

INTERNATIONAL
EDITION



Optoelectronics and Photonics

Principles and Practices

SECOND EDITION

S.O. Kasap



Selected definitions and basic equations

Photon energy

$$E_{\text{ph}} = h\nu = \hbar\omega; \omega = 2\pi\nu$$

Photon momentum

$$p_{\text{ph}} = \frac{h}{\lambda} = \hbar k$$

Photon flux Φ_{ph} and irradiance (intensity)

$$\Phi_{\text{ph}} = \frac{\text{Photons crossing area } A \text{ in time } \Delta t}{A \Delta t} = \frac{\Delta N_{\text{ph}}}{A \Delta t}$$

$$I = h\nu\Phi_{\text{ph}}$$

Propagation constant (wave vector)

$$k = \frac{2\pi}{\lambda}$$

Phase velocity

$$v = \lambda\nu = \frac{\omega}{k}; \quad v = \frac{c}{n} = \frac{c}{\sqrt{\epsilon_r}}$$

Changes in wavelength and frequency

$$\frac{\delta\lambda}{\lambda} = -\frac{\delta\nu}{\nu}; \quad \delta\lambda = -\frac{\lambda^2}{c}\delta\nu = -\frac{c}{\nu^2}\delta\nu$$

Group velocity

$$v_g = \frac{d\omega}{dk}$$

Group index

$$v_g(\text{medium}) = \frac{c}{N_g}; \quad N_g = n - \lambda_o \frac{dn}{d\lambda_o}$$

Electric and magnetic fields

$$E_x = \nu B_y = \frac{c}{n} B_y$$

Poynting vector and irradiance

$$\mathbf{S} = \nu^2 \epsilon_o \epsilon_r \mathbf{E} \times \mathbf{B}; \quad I = S_{\text{average}} = \frac{1}{2} \nu \epsilon_o \epsilon_r E_o^2$$

Snell's law and the Brewster angle

$$n_1 \sin \theta_i = n_2 \sin \theta_t; \sin \theta_c = \frac{n_2}{n_1}; \tan \theta_p = \frac{n_2}{n_1}$$

Phase change in total internal reflection (TIR)

$$\tan\left(\frac{1}{2}\phi_{\perp}\right) = \frac{[\sin^2 \theta_i - n^2]^{1/2}}{\cos \theta_i}; \quad n = \frac{n_2}{n_1}$$

$$\tan\left(\frac{1}{2}(\phi_{\parallel} + \pi)\right) = \frac{[\sin^2 \theta_i - n^2]^{1/2}}{n^2 \cos \theta_i}$$

Attenuation in second medium in TIR

$$\alpha_2 = \frac{2\pi n_2}{\lambda_o} \left[\left(\frac{n_1}{n_2} \right)^2 \sin^2 \theta_i - 1 \right]^{1/2}$$

Reflectance, transmittance (normal incidence)

$$R = R_{\perp} = R_{\parallel} = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2;$$

$$T = T_{\perp} = T_{\parallel} = \frac{4n_1 n_2}{(n_1 + n_2)^2}$$

Fabry–Perot cavity

$$v_m = m \left(\frac{c}{2L} \right) = mv_f, \quad m = 1, 2, 3, \dots$$

$$\delta v_m = \frac{v_f}{F}; \quad F = \frac{\pi R^{1/2}}{1 - R}$$

Single slit diffraction

$$I(\theta) = I(0) \text{sinc}^2(\beta); \quad \beta = \frac{1}{2} (ka \sin \theta)$$

Airy disk, angular radius, divergence

$$\sin \theta_o = 1.22 \frac{\lambda}{D}$$

$$\text{Divergence} = 2\theta_o \approx 2 \times 1.22 \frac{\lambda}{D}$$

Diffraction grating

$$d(\sin \theta_m - \sin \theta_i) = m\lambda; \quad m = 0, \pm 1, \pm 2, \dots$$

V-number, normalized frequency

$$V = \frac{2\pi a}{\lambda} (n_1^2 - n_2^2)^{1/2}; \quad V = \frac{2\pi a}{\lambda} \text{NA}$$

Normalized index difference

$$\Delta = (n_1 - n_2)/n_1$$

Acceptance angle and numerical aperture (NA)

$$2\alpha_{\text{max}}; \sin \alpha_{\text{max}} = \frac{(n_1^2 - n_2^2)^{1/2}}{n_0}; \quad \sin \alpha_{\text{max}} = \frac{\text{NA}}{n_0}$$

Normalized propagation constant

$$b = \frac{(\beta/k)^2 - n_2^2}{n_1^2 - n_2^2} \approx \frac{(\beta/k) - n_2}{n_1 - n_2}$$

$$b \approx \left(1.1428 - \frac{0.996}{V} \right)^2 \text{ for } 1.5 < V < 2.5$$

Single mode waveguides

Planar waveguide: $V < \pi/2$

Step-index fiber: $V < 2.405$

Mode field diameter

$$2w = 2a(0.65 + 1.619V^{-3/2} + 2.879V^{-6});$$

$$0.8 < V < 2.5$$

Dispersion in multimode step-index fiber

$$\frac{\Delta\tau}{L} \approx \frac{n_1 - n_2}{c} = \frac{n_1\Delta}{c}$$

Dispersion coefficient

$$\begin{aligned}\Delta\tau/L &= \text{Spread in group delay per unit length} \\ &= D\Delta\lambda\end{aligned}$$

$$D = \frac{\Delta\tau}{L\Delta\lambda}$$

Chromatic dispersion

$$\frac{\Delta\tau}{L} = |D_m + D_w + D_p|\Delta\lambda$$

Maximum RTZ bit rate

$$B \approx \frac{0.25}{\sigma}$$

Attenuation in optical fibers

$$\alpha_{\text{dB}} = \frac{1}{L} 10 \log \left(\frac{P_{\text{in}}}{P_{\text{out}}} \right) = 4.34\alpha$$

where α is the attenuation coefficient.



Optical gain coefficient

$$g(v) = \sigma_{\text{em}}(v)N_2 - \sigma_{\text{ab}}(v)N_1$$

Optical gain

$$G = \exp(gL)$$

Threshold gain in lasers

$$g_{\text{th}} = \alpha_s + \frac{1}{2L} \ln \left(\frac{1}{R_1 R_2} \right) = \alpha_t$$

Photon cavity lifetime

$$\tau_{\text{ph}} \approx n/c\alpha_t$$

Bandgap light and wavelength

$$\lambda_g(\mu\text{m}) = \frac{1.24}{E_g(\text{eV})}$$

Responsivity of a photodetector

$$R = \frac{\text{Photocurrent (A)}}{\text{Incident optical power (W)}} = \frac{I_{ph}}{P_o}$$

External quantum efficiency of a photodetector

$$\eta_e = \frac{I_{ph}/e}{P_o/h\nu}$$

Phase change between e - and o -waves

$$\phi = \frac{2\pi}{\lambda}(n_e - n_o)L$$

Arthur L. Schawlow is adjusting a ruby optical maser during an experiment at Bell Labs, while C.G.B. Garrett prepares to photograph the maser flash. In 1981, Arthur Schawlow shared the Nobel Prize in Physics for his “contribution to the development of laser spectroscopy.” (Reprinted with permission of Alcatel-Lucent USA Inc.)

2,929,922



2,929,922

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

Whereas Arthur L. Schawlow, of Madison, New Jersey, and Charles H. Townes, of New York, N. Y., assignors to Bell Telephone Laboratories, Incorporated, of New York, N. Y., a corporation of New York,

PRESENTED TO THE **Commissioner of Patents** A PETITION PRAYING FOR THE GRANT OF LETTERS PATENT FOR AN ALLEGED NEW AND USEFUL INVENTION THE TITLE AND A DESCRIPTION OF WHICH ARE CONTAINED IN THE SPECIFICATION OF WHICH A COPY IS HERETOUPON ANNEXED AND MADE A PART HEREOF, AND COMPLIED WITH THE VARIOUS REQUIREMENTS OF LAW IN SUCH CASES MADE AND PROVIDED, AND

Whereas UPON DUE EXAMINATION MADE THE SAID CLAIMANTS ARE ADJUDGED TO BE JUSTLY ENTITLED TO A PATENT UNDER THE LAW.

NOW THEREFORE THESE Letters Patent ARE TO GRANT UNTO THE SAID Bell Telephone Laboratories, Incorporated, its successors OR ASSIGNS FOR THE TERM OF SEVENTEEN YEARS FROM THE DATE OF THIS GRANT

THE RIGHT TO EXCLUDE OTHERS FROM MAKING, USING OR SELLING THE SAID INVENTION THROUHOUT THE UNITED STATES.

In testimony whereof I have hereunto set my hand and caused the seal of the Patent Office to be affixed at the City of Washington this twenty-second day of March, in the year of our Lord one thousand nine hundred and sixty, and of the Independence of the United States of America one hundred and eighty-fourth.

Attest:

Mark H. Miller
Attesting Officer.

Robert C. Watson
Commissioner of Patents

FORM PD-377A
1910-521

The patent for the invention of the laser by Charles H. Townes and Arthur L. Schawlow in 1960 (Reprinted with permission of Alcatel-Lucent USA Inc.). This laser patent was later bitterly disputed for almost three decades in the so-called "laser patent wars" by Gordon Gould, an American physicist, and his designated agents. Gordon Gould eventually received the U.S. patent for optical pumping of the laser in 1977 inasmuch as the original laser patent did not detail such a pumping procedure. In 1987 he also received a patent for the gas discharge laser, thereby winning his 30-year patent war. His original notebook even contained the word "laser." (See "Winning the laser-patent war", Jeff Hecht, *Laser Focus World*, December 1994, pp. 49-51).

Second Edition

Optoelectronics and Photonics: Principles and Practices

S.O. Kasap

University of Saskatchewan

Canada

International Edition Contributions by

Ravindra Kumar Sinha

Delhi Technological University

India

PEARSON

Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Vice President and Editorial Director, ECS: *Marcia J. Horton*
Executive Editor: *Andrew Gilfillan*
Editorial Assistant: *William Opaluch*
Marketing Manager: *Tim Galligan*
Marketing Assistant: *Jon Bryant*
Permissions Project Manager: *Karen Sanatar*
Senior Managing Editor: *Scott Disanno*
Production Project Manager: *Clare Romeo*
Publisher, International Editions: *Angshuman Chakraborty*
Publishing Administrator and Business Analyst, International Editions: *Shokhi Shah Khandelwal*
Senior Print and Media Editor, International Editions:
Ashwitha Jayakumar

Acquisitions Editor, International Editions: *Sandhya Ghoshal*
Publishing Administrator, International Editions:
Hema Mehta
Project Editor, International Editions:
Karthik Subramanian
Senior Manufacturing Controller, Production, International Edition: *Trudy Kimber*
Creative Director: *Jayne Conte*
Cover Design: *Suzanne Behnke*
Cover Illustration/Photo: *Courtesy of Teledyne-DALSA*
Image Permission Coordinator: *Karen Sanatar*
Full-Service Project Management/Composition:
Integra Software Services, Pvt. Ltd.

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the world

Visit us on the World Wide Web at:
www.pearsoninternationaleditions.com

© Pearson Education Limited 2013

The right of S.O. Kasap to be identified as author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

Authorized adaptation from the United States edition, entitled Optoelectronics and Photonics: Principles and Practices, 2nd edition, ISBN 978-0-13-215149-8, by S.O. Kasap, published by Pearson Education © 2013.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without either the prior written permission of the publisher or a license permitting restricted copying in the United Kingdom issued by the Copyright Licensing Agency Ltd, Saffron House, 6–10 Kirby Street, London EC1N 8TS.

All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

Microsoft and/or its respective suppliers make no representations about the suitability of the information contained in the documents and related graphics published as part of the services for any purpose. All such documents and related graphics are provided "as is" without warranty of any kind. Microsoft and/or its respective suppliers hereby disclaim all warranties and conditions with regard to this information, including all warranties and conditions of merchantability, whether express, implied or statutory, fitness for a particular purpose, title and non-infringement. In no event shall Microsoft and/or its respective suppliers be liable for any special, indirect or consequential damages or any damages whatsoever resulting from loss of use, data or profits, whether in an action of contract, negligence or other tortious action, arising out of or in connection with the use or performance of information available from the services.

The documents and related graphics contained herein could include technical inaccuracies or typographical errors. Changes are periodically added to the information herein. Microsoft and/or its respective suppliers may make improvements and/or changes in the product(s) and/or the program(s) described herein at any time. Partial screen shots may be viewed in full within the software version specified.

Microsoft® and Windows® are registered trademarks of the Microsoft Corporation in the U.S.A. and other countries. This book is not sponsored or endorsed by or affiliated with the Microsoft Corporation.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

10 9 8 7 6 5 4 3 2 1
14 13 12 11 10

Typeset in 10/12 Times LT Std by Integra Software Services, Pvt. Ltd.

Printed and bound by Courier Westford in The United States of America

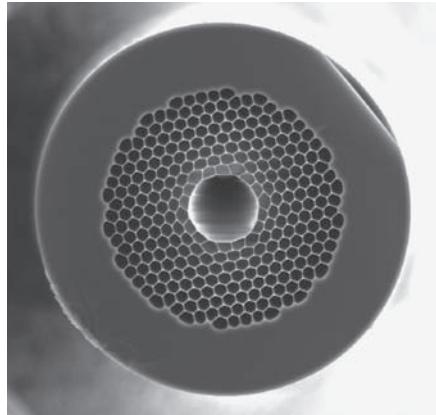
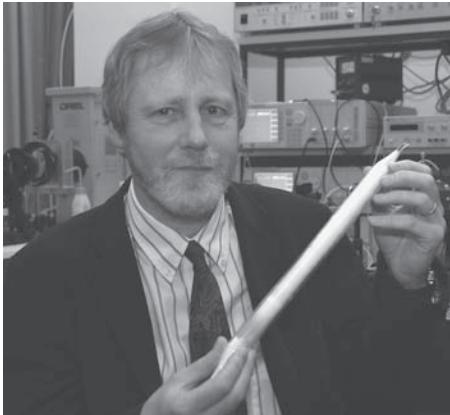
The publisher's policy is to use paper manufactured from sustainable forests.

PEARSON

ISBN 10: 0-273-77417-4
ISBN 13: 978-0-273-77417-4

We have a habit in writing articles published in scientific journals to make the work as finished as possible, to cover up all the tracks, to not worry about the blind alleys or describe how you had the wrong idea first, and so on. So there isn't any place to publish, in a dignified manner, what you actually did in order to get to do the work.

—RICHARD P. FEYNMAN
NOBEL LECTURE, 1966



Philip Russell led a team of researchers at the University Bath in the 1990s where photonic crystal fibers were drawn. Thin hollow capillary tubes were stacked together and then fused to make a preform as shown on the left. A photonic crystal fiber was then drawn at a high temperature from this preform. Photonic crystal fibers have the ability to guide light endlessly in a single mode, and have highly desirable nonlinear properties for various photonics applications in the manipulation of light, such as the generation of supercontinuum light. (Courtesy of Professor Philip Russell.)



Peter Schultz, Donald Keck, and Bob Maurer (left to right) at Corning were the first to produce low-loss optical fibers in the 1970s by using the outside vapor deposition method for the fabrication of preforms, which were then used to draw fibers with low losses.
(Courtesy of Corning.)

*To Nicolette, who brightens my every day
and makes me smile with joy every time I see her.*

PREFACE

The first edition of this book was written more than 12 years ago. At the time it was meant as an easy-to-read book for third-year engineering or applied physics undergraduate students; it emphasized qualitative explanations and relied heavily on intuitive derivations. As things turned out, the first edition ended up being used in fourth-year elective classes, and even in graduate courses on optoelectronics. Many of the instructors teaching at that level rightly needed better derivations, more rigor, better explanations, and, of course, many more topics and problems. We have all at one time or another suffered from how wrong some intuitive short-cut derivations can be. The second edition was therefore prepared by essentially rewriting the text almost from scratch with much better rigor and explanations, but without necessarily dwelling on mathematical details. Many new exciting practical examples have been introduced, and numerous new problems have been added. The book also had to be totally modernized given that much had happened in the intervening 12 years that deserved being covered in an undergraduate course.

FEATURES, CHANGES, AND REVISIONS IN THE SECOND EDITION

The second edition represents a total revision of the first edition, with numerous additional features and enhancements.

- All chapters have been totally revised and extended.
- Numerous modern topics in photonics have been added to all the chapters.
- There are Additional Topics that can be covered in more advanced courses, or in courses that run over two semesters.
- There are many more new examples and solved problems within chapters, and many more practical end-of-chapter problems that start from basic concepts and build up onto advanced applications.
- Nearly all the illustrations and artwork in the first edition have been revised and redrawn to better reflect the concepts.
- Numerous new illustrations have been added to convey the concepts as clearly as possible.
- Photographs have been added, where appropriate, to enhance the readability of the book and to illustrate typical modern photonic/optoelectronic devices.
- The previous edition's Chapter 7 on photovoltaics has been incorporated into this edition's Chapter 5 as an Additional Topic, thus allowing more photonics-related topics to be covered.
- Advanced or complicated mathematical derivations are avoided and, instead, the emphasis is placed on concepts and engineering applications.
- Useful and essential equations in photonics are given with explanations and are used in examples and problems to give the student a sense of what typical values are.
- Cross referencing in the second edition has been avoided as much as possible, without too much repetition, to allow various sections and chapters to be skipped as desired by the reader.
- There is greater emphasis on practical or engineering examples; care has been taken to consider various photonics/optoelectronics courses at the undergraduate level across major universities.

- The second edition is supported by an extensive PowerPoint presentation for instructors who have adopted the book for their course. The PowerPoint slides have all the illustrations in color, and include additional color photos. The basic concepts and equations are also highlighted in additional slides. There are also numerous slides with examples and solved problems. Instructors should visit www.pearsoninternationaleditions.com/kasap to access the PowerPoints.
- The second edition is also supported by an extensive Solutions Manual for instructors only. This is available from the publisher at www.pearsoninternationaleditions/kasap.

The second edition continues to represent a first course in optoelectronic materials and devices suitable for a half- or one-semester course at the undergraduate level either at the third- or fourth-year level in *electrical engineering*, *engineering physics*, and *materials science and engineering* departments. With its additional topics, it can also be used as an introductory textbook at the graduate level. Normally the students would not have covered Maxwell's equations. Although Maxwell's equations are mentioned in the text to alert the student, they are not used in developing the principles. It is assumed that the students would have taken a basic first- or second-year physics course, with modern physics, and would have seen rudimentary concepts in geometrical optics, interference, and diffraction, but not Fresnel's equations and concepts such as group velocity and group index. Typically an optoelectronics course would be given either after a semiconductor devices course or concurrently with it. Students would have been exposed to elementary quantum mechanics concepts, perhaps in conjunction with a basic semiconductor science course.

Most topics are initially introduced through qualitative explanations to allow the concept to be grasped first before any mathematical development. The mathematical level is assumed to include vectors, complex numbers, and partial differentiation but excludes reliance on Fourier transforms. On the one hand, we are required to cover as much as possible and, on the other hand, professional engineering accreditation requires students to solve numerical problems and carry out "design calculations." In preparing the text, I tried to satisfy engineering degree accreditation requirements in as much breadth as possible. Obviously one cannot solve numerical problems, carry out design calculations, and at the same time derive each equation without expanding the size of the text to an intolerable level. I have missed many topics but I have also covered many, though, undoubtedly, it is my own very biased selection.

I would like to thank two very special colleagues, whom I have known for a very long time, for their comments and help: Harry Ruda (University of Toronto) and Raman Kashyap (École Polytechnique de Montréal)—two perfect gentlemen who read some of the manuscript and made valuable criticisms toward this final version.

No textbook is perfect and I can only improve the text with your input. Please feel free to write to me with your comments. Although I may not be able to reply to each individual comment and suggestion, I do read all my email messages and take good note of suggestions and comments. Many instructors did, in fact, write to me on the first edition, pointed out how things could have been done better, and various mistakes one never seems to be able to eliminate totally. I hope that the second edition will at least go far in satisfying some of their criticisms. There is an important old adage that goes something like this (somewhat paraphrased), "a good diagram is worth a thousand words, but a bad diagram takes a thousand words to explain." I used a software package called Canvas to draw nearly all the line-art in the second edition as clearly as possible, and errors are all *mea culpa*; feel free to email me the errors you notice in the figures. All third-party artwork and photographs have been used with permission; and I'm grateful to

Pearson Education for meticulously obtaining permission from copyright holders. If you like the second edition, and cannot wait for the third, you can always write your comments and recommendations directly to the Sponsoring Editor for Electrical Engineering, Pearson Higher Education, One Lake Street, Upper Saddle River, NJ 07458, USA. This is the best way to have your input heard.

Resources for Instructors

- **Instructor's Solutions Manual.** An instructor's solutions manual was prepared by the author.
- **Presentation Resources.** All art from the text is available in PowerPoint slide and JPEG format.

These files are available for download from the instructor Resource Center at www.pearsonhighered.com/Kasap. If you are in need of a login and password for this site, please contact your local Pearson Prentice-Hall representative.

Safa Kasap

safa.kasap@yahoo.com

Saskatoon (March 2012)

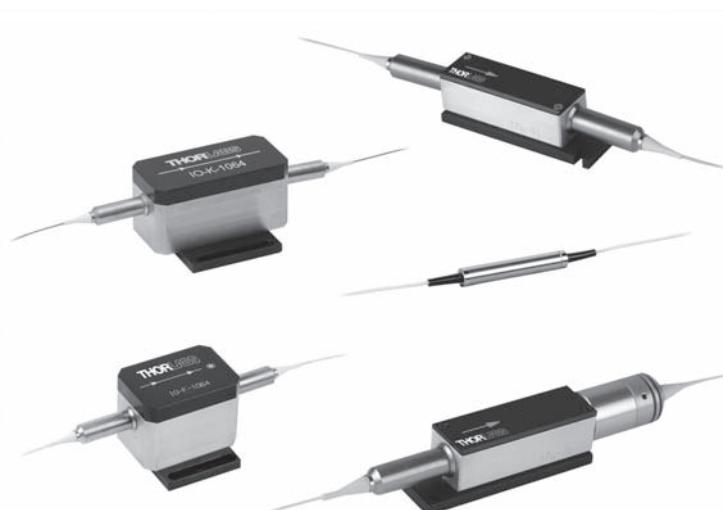
The publishers would like to thank Sheenu Thomas of Cochin University of Science and Technology for reviewing the content of the International Edition.



Gordon Gould (1920–2005) obtained his BSc in Physics (1941) from Union College in Schenectady and MSc from Yale University. Gould came up with the idea of an optically pumped laser during his PhD work at Columbia University around 1957—he is now recognized for the invention of *optical pumping* as a means of exciting masers and lasers. He has been also credited for collisional pumping as in gas lasers, and a variety of application-related laser patents. After nearly three decades of legal disputes, in 1987, he eventually won rights to the invention of the laser. Gould's laboratory logbook even had an entry with the heading "Some rough calculations on the feasibility of a LASER: Light Amplification by Stimulated Emission of Radiation," which is the first time that this acronym appears. Union College awarded Gould an honorary Doctor of Sciences in 1978 and the Eliphalet Nott Medal in 1995. (*Courtesy of Union College Alumni Office.*)



Fiber-coupled phase and amplitude modulators. The Pockels effect in lithium niobate is used to modulate the refractive index and hence the phase of an optical signal. In amplitude modulators, the Pockels effect is used to modulate the refractive indices of the two arms of a Mach-Zehnder interferometer, and hence the optical output. (© JENOPTIK Optical System GmbH.)



Fiber-coupled optical isolators: fiber isolators. The signal is allowed to propagate in one direction only, along the arrow shown on the device. The principle is based on Faraday rotation. (Courtesy of Thorlabs.)

CONTENTS

Chapter 1 Wave Nature of Light 19

1.1 Light Waves in a Homogeneous Medium 19

A. Plane Electromagnetic Wave 19

B. Maxwell's Wave Equation and Diverging Waves 22

Example 1.1.1 A diverging laser beam 26

1.2 Refractive Index and Dispersion 26

Example 1.2.1 Sellmeier equation and diamond 29

Example 1.2.2 Cauchy equation and diamond 30

1.3 Group Velocity and Group Index 30

Example 1.3.1 Group velocity 33

Example 1.3.2 Group velocity and index 33

Example 1.3.3 Group and phase velocities 34

1.4 Magnetic Field, Irradiance, and Poynting Vector 34

Example 1.4.1 Electric and magnetic fields in light 37

Example 1.4.2 Power and irradiance of a Gaussian beam 37

1.5 Snell's Law and Total Internal Reflection (TIR) 38

Example 1.5.1 Beam displacement 41

1.6 Fresnel's Equations 42

A. Amplitude Reflection and Transmission Coefficients (r and t) 42

B. Intensity, Reflectance, and Transmittance 48

C. Goos-Hänchen Shift and Optical Tunneling 49

Example 1.6.1 Reflection of light from a less dense medium (internal reflection) 51

Example 1.6.2 Reflection at normal incidence, and internal and external reflection 52

Example 1.6.3 Reflection and transmission at the Brewster angle 53

1.7 Antireflection Coatings and Dielectric Mirrors 54

A. Antireflection Coatings on Photodetectors and Solar Cells 54

Example 1.7.1 Antireflection coating on a photodetector 55

B. Dielectric Mirrors and Bragg Reflectors 56

Example 1.7.2 Dielectric mirror 58

1.8 Absorption of Light and Complex Refractive Index 59

Example 1.8.1 Complex refractive index of InP 62

Example 1.8.2 Reflectance of CdTe around resonance absorption 63

1.9 Temporal and Spatial Coherence 63

Example 1.9.1 Coherence length of LED light 66

1.10 Superposition and Interference of Waves 67

1.11 Multiple Interference and Optical Resonators 69

Example 1.11.1 Resonator modes and spectral width of a semiconductor Fabry–Perot cavity 73

1.12 Diffraction Principles 74

A. Fraunhofer Diffraction 74

Example 1.12.1 Resolving power of imaging systems 79

B. Diffraction Grating 80

Example 1.12.2 A reflection grating 83

Additional Topics 84**1.13 Interferometers 84****1.14 Thin Film Optics: Multiple Reflections in Thin Films 86**

Example 1.14.1 Thin film optics 88

1.15 Multiple Reflections in Plates and Incoherent Waves 89**1.16 Scattering of Light 90****1.17 Photonic Crystals 92**

Questions and Problems 98

Chapter 2 Dielectric Waveguides and Optical Fibers 111**2.1 Symmetric Planar Dielectric Slab Waveguide 111**

A. Waveguide Condition 111

B. Single and Multimode Waveguides 116

C. TE and TM Modes 116

Example 2.1.1 Waveguide modes 117

Example 2.1.2 *V*-number and the number of modes 118

Example 2.1.3 Mode field width, $2w_o$ 119

2.2 Modal and Waveguide Dispersion in Planar Waveguides 120

A. Waveguide Dispersion Diagram and Group Velocity 120

B. Intermodal Dispersion 121

C. Intramodal Dispersion 122

2.3 Step-Index Optical Fiber 123

A. Principles and Allowed Modes 123

Example 2.3.1 A multimode fiber 128

Example 2.3.2 A single-mode fiber 128

B. Mode Field Diameter 128

Example 2.3.3 Mode field diameter 129

C. Propagation Constant and Group Velocity 130

Example 2.3.4 Group velocity and delay 131

D. Modal Dispersion in Multimode Step-Index Fibers 132

Example 2.3.5 A multimode fiber and dispersion 132

2.4 Numerical Aperture	133
Example 2.4.1	A multimode fiber and total acceptance angle 134
Example 2.4.2	A single-mode fiber 134
2.5 Dispersion In Single-Mode Fibers	135
A. Material Dispersion	135
B. Waveguide Dispersion	136
C. Chromatic Dispersion	138
D. Profile and Polarization Dispersion Effects	138
Example 2.5.1	Material dispersion 140
Example 2.5.2	Material, waveguide, and chromatic dispersion 141
Example 2.5.3	Chromatic dispersion at different wavelengths 141
Example 2.5.4	Waveguide dispersion 142
2.6 Dispersion Modified Fibers and Compensation	142
A. Dispersion Modified Fibers	142
B. Dispersion Compensation	144
Example 2.6.1	Dispersion compensation 146
2.7 Bit Rate, Dispersion, and Electrical and Optical Bandwidth	146
A. Bit Rate and Dispersion	146
B. Optical and Electrical Bandwidth	149
Example 2.7.1	Bit rate and dispersion for a single-mode fiber 151
2.8 The Graded Index (GRIN) Optical Fiber	151
A. Basic Properties of GRIN Fibers	151
B. Telecommunications	155
Example 2.8.1	Dispersion in a graded index fiber and bit rate 156
Example 2.8.2	Dispersion in a graded index fiber and bit rate 157
2.9 Attenuation in Optical Fibers	158
A. Attenuation Coefficient and Optical Power Levels	158
Example 2.9.1	Attenuation along an optical fiber 160
B. Intrinsic Attenuation in Optical Fibers	160
C. Intrinsic Attenuation Equations	162
Example 2.9.2	Rayleigh scattering equations 163
D. Bending losses	164
Example 2.9.3	Bending loss for SMF 167
2.10 Fiber Manufacture	168
A. Fiber Drawing	168
B. Outside Vapor Deposition	169
Example 2.10.1	Fiber drawing 171
Additional Topics	171
2.11 Wavelength Division Multiplexing: WDM	171
2.12 Nonlinear Effects in Optical Fibers and DWDM	173

- 2.13 Bragg Fibers 175
- 2.14 Photonic Crystal Fibers—Holey Fibers 176
- 2.15 Fiber Bragg Gratings and Sensors 179
 - Example 2.15.1** Fiber Bragg grating at 1550 nm 183
 - Questions and Problems* 183

Chapter 3 Semiconductor Science and Light-Emitting Diodes 195

- 3.1 Review of Semiconductor Concepts and Energy Bands 195
 - A. Energy Band Diagrams, Density of States, Fermi–Dirac Function and Metals 195
 - B. Energy Band Diagrams of Semiconductors 198
- 3.2 Semiconductor Statistics 200
- 3.3 Extrinsic Semiconductors 203
 - A. *n*-Type and *p*-Type Semiconductors 203
 - B. Compensation Doping 206
 - C. Nondegenerate and Degenerate Semiconductors 207
 - D. Energy Band Diagrams in an Applied Field 208
 - Example 3.3.1** Fermi levels in semiconductors 209
 - Example 3.3.2** Conductivity of *n*-Si 209
- 3.4 Direct and Indirect Bandgap Semiconductors:
E–*k* Diagrams 210
- 3.5 *pn* Junction Principles 214
 - A. Open Circuit 214
 - B. Forward Bias and the Shockley Diode Equation 217
 - C. Minority Carrier Charge Stored in Forward Bias 222
 - D. Recombination Current and the Total Current 222
- 3.6 *pn* Junction Reverse Current 225
- 3.7 *pn* Junction Dynamic Resistance and Capacitances 227
 - A. Depletion Layer Capacitance 227
 - B. Dynamic Resistance and Diffusion Capacitance for Small Signals 229
- 3.8 Recombination Lifetime 230
 - A. Direct Recombination 230
 - B. Indirect Recombination 232
 - Example 3.8.1** A direct bandgap *pn* junction 232
- 3.9 *pn* Junction Band Diagram 234
 - A. Open Circuit 234
 - B. Forward and Reverse Bias 236
 - Example 3.9.1** The built-in voltage from the band diagram 237
- 3.10 Heterojunctions 238

3.11	Light-Emitting Diodes: Principles	240
A.	Homojunction LEDs	240
B.	Heterostructure High Intensity LEDs	242
C.	Output Spectrum	244
	Example 3.11.1 LED spectral linewidth	247
	Example 3.11.2 LED spectral width	248
	Example 3.11.3 Dependence of the emission peak and linewidth on temperature	249
3.12	Quantum Well High Intensity LEDs	249
	Example 3.12.1 Energy levels in the quantum well	252
3.13	LED Materials and Structures	253
A.	LED Materials	253
B.	LED Structures	254
	Example 3.13.1 Light extraction from a bare LED chip	257
3.14	LED Efficiencies and Luminous Flux	258
	Example 3.14.1 LED efficiencies	260
	Example 3.14.2 LED brightness	261
3.15	Basic LED Characteristics	261
3.16	LEDs for Optical Fiber Communications	262
3.17	Phosphors and White LEDs	265
Additional Topics 267		
3.18	LED Electronics	267
<i>Questions and Problems</i> 270		

Chapter 4 Stimulated Emission Devices: Optical Amplifiers and Lasers 281

4.1	Stimulated Emission, Photon Amplification, and Lasers	281
A.	Stimulated Emission and Population Inversion	281
B.	Photon Amplification and Laser Principles	282
C.	Four-Level Laser System	285
4.2	Stimulated Emission Rate and Emission Cross-Section	286
A.	Stimulated Emission and Einstein Coefficients	286
	Example 4.2.1 Minimum pumping power for three-level laser systems	288
B.	Emission and Absorption Cross-Sections	289
	Example 4.2.2 Gain coefficient in a Nd^{3+} -doped glass fiber	291
4.3	Erbium-Doped Fiber Amplifiers	292
A.	Principle of Operation and Amplifier Configurations	292
B.	EDFA Characteristics, Efficiency, and Gain Saturation	296
	Example 4.3.1 An erbium-doped fiber amplifier	299
C.	Gain-Flattened EDFA and Noise Figure	300

4.4	Gas Lasers: The He-Ne Laser	303
	Example 4.4.1	Efficiency of the He-Ne laser
4.5	The Output Spectrum of a Gas Laser	306
	Example 4.5.1	Doppler broadened linewidth
4.6	Laser Oscillations: Threshold Gain Coefficient and Gain Bandwidth	311
	A.	Optical Gain Coefficient g
	B.	Threshold Gain Coefficient g_{th} and Output Power
	Example 4.6.1	Threshold population inversion for the He-Ne laser
	C.	Output Power and Photon Lifetime in the Cavity
	Example 4.6.2	Output power and photon cavity lifetime τ_{ph}
	D.	Optical Cavity, Phase Condition, Laser Modes
4.7	Broadening of the Optical Gain Curve and Linewidth	319
4.8	Pulsed Lasers: Q-Switching and Mode Locking	323
	A.	Q-Switching
	B.	Mode Locking
4.9	Principle of the Laser Diode	327
4.10	Heterostructure Laser Diodes	331
	Example 4.10.1	Modes in a semiconductor laser and the optical cavity length
4.11	Quantum Well Devices	337
	Example 4.11.1	A GaAs quantum well
4.12	Elementary Laser Diode Characteristics	340
	Example 4.12.1	Laser output wavelength variation with temperature
	Example 4.12.2	Laser diode efficiencies for a sky-blue LD
	Example 4.12.3	Laser diode efficiencies
4.13	Steady State Semiconductor Rate Equations:	
	The Laser Diode Equation	348
	A.	Laser Diode Equation
	B.	Optical Gain Curve, Threshold, and Transparency Conditions
	Example 4.13.1	Threshold current and optical output power from a Fabry-Perot heterostructure laser diode
4.14	Single Frequency Semiconductor Lasers	354
	A.	Distributed Bragg Reflector LDs
	B.	Distributed Feedback LDs
	C.	External Cavity LDs
	Example 4.14.1	DFB LD wavelength
4.15	Vertical Cavity Surface Emitting Lasers	360
4.16	Semiconductor Optical Amplifiers	364

Additional Topics 366**4.17 Superluminescent and Resonant Cavity LEDs:
SLD and RCLED 366****4.18 Direct Modulation of Laser Diodes 367****4.19 Holography 370***Questions and Problems 373***Chapter 5 Photodetectors and Image Sensors 381****5.1 Principle of the *pn* Junction Photodiode 381****A. Basic Principles 381****B. Energy Band Diagrams and Photodetection Modes 383****C. Current-Voltage Convention and Modes of Operation 385****5.2 Shockley–Ramo Theorem and External Photocurrent 386****5.3 Absorption Coefficient and Photodetector Materials 388****5.4 Quantum Efficiency and Responsivity 391****Example 5.4.1 Quantum efficiency and responsivity 394****Example 5.4.2 Maximum quantum efficiency 395****5.5 The *pin* Photodiode 395****Example 5.5.1 Operation and speed of a *pin* photodiode 399****Example 5.5.2 Photocarrier diffusion in a *pin* photodiode 399****Example 5.5.3 Responsivity of a *pin* photodiode 400****Example 5.5.4 Steady state photocurrent in the *pin* photodiode 401****5.6 Avalanche Photodiode 402****A. Principles and Device Structures 402****Example 5.6.1 InGaAs APD responsivity 406****Example 5.6.2 Silicon APD 406****B. Impact Ionization and Avalanche Multiplication 406****Example 5.6.3 Avalanche multiplication in Si APDs 408****5.7 Heterojunction Photodiodes 409****A. Separate Absorption and Multiplication APD 409****B. Superlattice APDs 411****5.8 Schottky Junction Photodetector 413****5.9 Phototransistors 417****5.10 Photoconductive Detectors and Photoconductive
Gain 418****5.11 Basic Photodiode Circuits 421****5.12 Noise in Photodetectors 424****A. The *pn* Junction and *pin* Photodiodes 424****Example 5.12.1 NEP of a Si *pin* photodiode 428**

Example 5.12.2 Noise of an ideal photodetector 428

Example 5.12.3 SNR of a receiver 429

B. Avalanche Noise in the APD 430

Example 5.12.4 Noise in an APD 430

5.13 Image Sensors 431

A. Basic Principles 431

B. Active Matrix Array and CMOS Image Sensors 433

C. Charge-Coupled Devices 435

Additional Topics 437

5.14 Photovoltaic Devices: Solar Cells 437

A. Basic Principles 437

B. Operating Current and Voltage and Fill Factor 439

C. Equivalent Circuit of a Solar Cell 440

D. Solar Cell Structures and Efficiencies 442

Example 5.14.1 Solar cell driving a load 444

Example 5.14.2 Open circuit voltage and short circuit current 445

Questions and Problems 445

Chapter 6 Polarization and Modulation of Light 457

6.1 Polarization 457

A. State of Polarization 457

Example 6.1.1 Elliptical and circular polarization 460

B. Malus's Law 460

6.2 Light Propagation in an Anisotropic Medium:

Birefringence 461

A. Optical Anisotropy 461

B. Uniaxial Crystals and Fresnel's Optical Indicatrix 463

C. Birefringence of Calcite 466

D. Dichroism 467

6.3 Birefringent Optical Devices 468

A. Retarding Plates 468

Example 6.3.1 Quartz-half wave plate 469

Example 6.3.2 Circular polarization from linear polarization 470

B. Soleil-Babinet Compensator 470

C. Birefringent Prisms 471

6.4 Optical Activity and Circular Birefringence 472

6.5 Liquid Crystal Displays 474

6.6 Electro-Optic Effects 478

A. Definitions 478

B. Pockels Effect 479
Example 6.6.1 Pockels Cell Modulator 484
C. Kerr Effect 484
Example 6.6.2 Kerr Effect Modulator 486
6.7 Integrated Optical Modulators 486
A. Phase and Polarization Modulation 486
B. Mach-Zehnder Modulator 487
C. Coupled Waveguide Modulators 489
Example 6.7.1 Modulated Directional Coupler 492
6.8 Acousto-Optic Modulator 492
A. Photoelastic Effect and Principles 492
B. Acousto-Optic Modulators 494
Example 6.8.1 AO Modulator 499
6.9 Faraday Rotation and Optical Isolators 499
Example 6.9.1 Faraday rotation 500
6.10 Nonlinear Optics and Second Harmonic Generation 501

Additional Topics 505

6.11 Jones Vectors 505
<i>Questions and Problems</i> 506

Appendices

<i>Appendix A Gaussian Distribution</i> 514
<i>Appendix B Solid Angles</i> 516
<i>Appendix C Basic Radiometry and Photometry</i> 518
<i>Appendix D Useful Mathematical Formulae</i> 521
<i>Appendix E Notation and Abbreviations</i> 523

Index 535



CMOS image sensors with wide dynamic range. (*Courtesy of New Imaging Technologies (NIT), France*)

Physicists use the wave theory on Mondays, Wednesdays and Fridays and the particle theory on Tuesdays, Thursdays and Saturdays.

—Sir William Henry Bragg¹



Augustin Jean Fresnel (1788–1827) was a French physicist and a civil engineer for the French government who was one of the principal proponents of the wave theory of light. He made a number of distinct contributions to optics including the well-known Fresnel lens that was used in lighthouses in the nineteenth century. He fell out with Napoleon in 1815 and was subsequently put under house arrest until the end of Napoleon's reign. During his enforced leisure time he formulated his wave ideas of light into a mathematical theory. (© INTERFOTO/Alamy.)

If you cannot saw with a file or file with a saw, then you will be no good as an experimentalist.

—Attributed to Augustin Fresnel

¹ As quoted in Alan Mackay, *Dictionary of Scientific Quotes*, 2nd Edition (Institute of Physics Publishing, Bristol, 1991), p. 37.

1

WAVE NATURE OF LIGHT

1.1 LIGHT WAVES IN A HOMOGENEOUS MEDIUM

A. Plane Electromagnetic Wave

The wave nature of light, quite aside from its photonic behavior, is well recognized by such phenomena as interference and diffraction. We can treat light as an electromagnetic (EM) wave with time-varying electric and magnetic fields, E_x and B_y , respectively, which are propagating through space in such a way that they are always perpendicular to each other and the direction of propagation z as illustrated in Figure 1.1. The simplest traveling wave is a sinusoidal wave that, for propagation along z , has the general mathematical form

Traveling
wave
along z

$$E_x = E_o \cos(\omega t - kz + \phi_o) \quad (1.1.1)$$

in which E_x is the electric field at position z at time t , k is the **propagation constant**² given by $2\pi/\lambda$, where λ is the wavelength, ω is the angular frequency, E_o is the amplitude of the wave, and ϕ_o is a phase constant, which accounts for the fact that at $t = 0$ and $z = 0$; E_x may or may not necessarily be zero depending on the choice of origin. The argument $(\omega t - kz + \phi_o)$ is called the **phase** of the wave and denoted by ϕ . Equation (1.1.1) describes a **monochromatic plane wave** of infinite extent traveling in the positive z direction as depicted in Figure 1.2. In any plane perpendicular to the direction of propagation (along z), the phase of the wave, according to Eq. (1.1.1), is constant, which means that the field in this plane is also constant. A surface over which the phase of a wave is constant at a given instant is referred to as a **wavefront**. A wavefront of a plane wave is obviously an infinite plane perpendicular to the direction of propagation as shown in Figure 1.2.

We know from electromagnetism that time-varying magnetic fields result in time-varying electric fields (Faraday's law) and vice versa. A time-varying electric field would set up a time-varying magnetic field with the same

²Some authors also call k the *wave number*. However, in spectroscopy, the wave number implies $1/\lambda$, reciprocal wavelength. To avoid any confusion, propagation constant would be preferred for k .

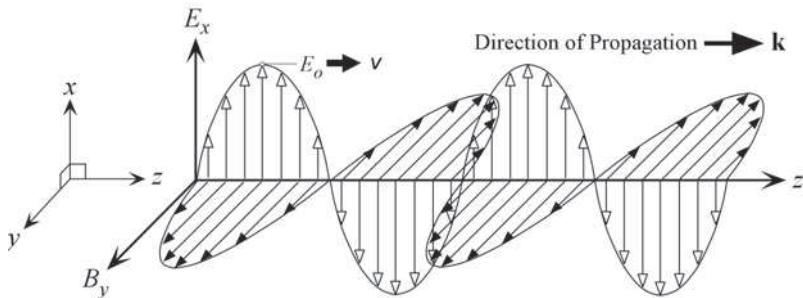


FIGURE 1.1 An electromagnetic wave in a homogenous and isotropic medium is a traveling wave that has time-varying electric and magnetic fields which are perpendicular to each other and the direction of propagation z . This is a snapshot at a given time of a particular harmonic or a sinusoidal EM wave. At a time δt later, a point on the wave, such as the maximum field, would have moved a distance $v\delta t$ in the z -direction.

frequency. According to electromagnetic principles,³ a traveling electric field E_x as represented by Eq. (1.1.1) would always be accompanied by a traveling magnetic field B_y with the same wave frequency and propagation constant (ω and k) but the directions of the two fields would be orthogonal as in Figure 1.1. Thus, there is a similar traveling wave equation for the magnetic field component B_y . We generally describe the interaction of a light wave with a non-conducting matter (conductivity, $\sigma = 0$) through the electric field component E_x rather than B_y because it is the electric field that displaces the electrons in molecules or ions in the crystal and thereby gives rise to the polarization of matter. However, the two fields are linked, as in Figure 1.1, and there is an intimate relationship between them. The **optical field** refers to the electric field E_x .

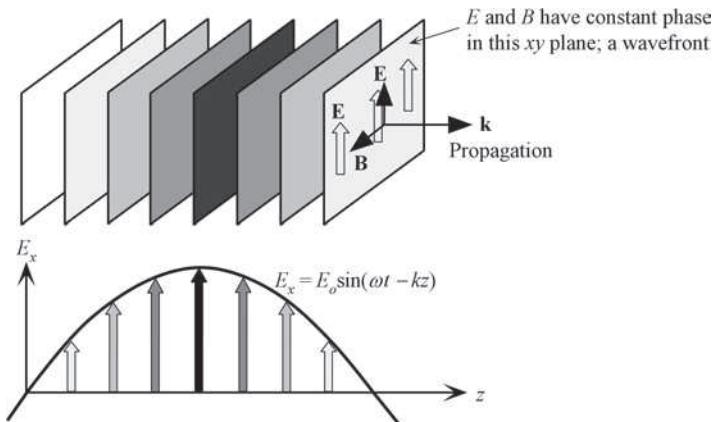


FIGURE 1.2 A plane EM wave traveling along z has the same E_x (or B_y) at any point in a given xy plane. All electric field vectors in a given xy plane are therefore in phase. The xy planes are of infinite extent in the x and y directions.

³Maxwell's equations formulate electromagnetic phenomena and provide relationships between the electric and magnetic fields and their space and time derivatives. We need only to use a few selected results from Maxwell's equation without delving into their derivations. The magnetic field B is also called the magnetic induction or magnetic flux density. The magnetic field intensity H and magnetic field B in a non-magnetic material are related by $B = \mu_0 H$ in which μ_0 is the absolute permeability of the medium.

We can also represent a traveling wave using the exponential notation since $\cos \phi = \text{Re}[\exp(j\phi)]$ in which **Re** refers to the real part. We then need to take the real part of any complex result at the end of calculations. Thus, we can write Eq. (1.1.1) as

$$E_x(z, t) = \text{Re}[E_o \exp(j\phi_o) \exp j(\omega t - kz)]$$

or

$$E_x(z, t) = \text{Re}[E_c \exp j(\omega t - kz)] \quad (1.1.2)$$

Traveling
wave
along z

in which $E_c = E_o \exp(j\phi_o)$ is a complex number that represents the amplitude of the wave and includes the constant phase information ϕ_o . Note that in Eq. (1.1.2), $\exp j(\omega t - kz)$ represents $e^{j(\omega t - kz)}$.

We indicate the direction of propagation with a vector **k**, called the **wave vector** (or **propagation vector**), whose magnitude is the **propagation constant**, $k = 2\pi/\lambda$. It is clear that **k** is perpendicular to constant phase planes as indicated in Figure 1.2. Consider an electromagnetic wave that is propagating along some arbitrary direction **k**, as indicated in Figure 1.3. The electric field $E(\mathbf{r}, t)$ at an arbitrary point \mathbf{r} is given by

$$E(\mathbf{r}, t) = E_o \cos(\omega t - \mathbf{k} \cdot \mathbf{r} + \phi_o) \quad (1.1.3)$$

Traveling
wave in
3D with
wave
vector k

because the dot product $\mathbf{k} \cdot \mathbf{r}$ is along the direction of propagation similar to kz as indicated in Figure 1.3. The latter can be shown by drawing a plane that has the point \mathbf{r} and is perpendicular to **k** as illustrated in Figure 1.3. The dot product is the product of **k** and the projection of \mathbf{r} onto **k**, which is \mathbf{r}' in Figure 1.3, so that $\mathbf{k} \cdot \mathbf{r} = kr'$. Indeed, if propagation is along z , $\mathbf{k} \cdot \mathbf{r}$ becomes kz . In general, if **k** has components k_x , k_y , and k_z along x , y , and z , then from the definition of the dot product, $\mathbf{k} \cdot \mathbf{r} = k_x x + k_y y + k_z z$.

The relationship between time and space for a given phase, ϕ for example, that corresponds to a maximum field, according to Eq. (1.1.1), is described by

$$\phi = \omega t - kz + \phi_o = \text{constant}$$

During a time interval δt , this constant phase (and hence the maximum field) moves a distance δz . The phase velocity of this wave is therefore $\delta z/\delta t$. Thus the **phase velocity** v is

$$v = \frac{dz}{dt} = \frac{\omega}{k} = v\lambda \quad (1.1.4)$$

Phase
velocity

in which v is the frequency ($\omega = 2\pi\nu$) of the EM wave. For an EM wave propagating in free space v is the speed of light in vacuum or c .

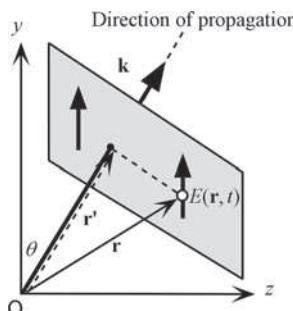


FIGURE 1.3 A traveling plane EM wave along a direction **k**.

We are often interested in the phase difference $\Delta\phi$ at a given time between two points on a wave (Figure 1.1) that are separated by a certain distance. If the wave is traveling along z with a wave vector k , as in Eq. (1.1.1), then the phase difference between two points separated by Δz is simply $k\Delta z$ since ωt is the same for each point. If this phase difference is 0 or multiples of 2π then the two points are in phase. Thus phase difference $\Delta\phi$ can be expressed as $k\Delta z$ or $2\pi\Delta z/\lambda$.

B. Maxwell's Wave Equation and Diverging Waves

Consider the plane EM wave in Figure 1.2. All constant phase surfaces are xy planes that are perpendicular to the z -direction. A cut of a plane wave parallel to the z -axis is shown in Figure 1.4 (a) in which the parallel dashed lines at right angles to the z -direction are wavefronts. We normally show wavefronts that are separated by a phase of 2π or a whole wavelength λ as in the figure. The vector that is normal to a wavefront surface at a point such as P represents the direction of wave propagation (\mathbf{k}) at that point P . Clearly, the propagation vectors everywhere are all parallel and the plane wave propagates without the wave diverging; *the plane wave has no divergence*. The amplitude of the planar wave E_o does not depend on the distance from a reference point, and it is the same at all points on a given plane perpendicular to \mathbf{k} (i.e., independent of x and y). Moreover, as these planes extend to infinity there is infinite energy in the plane wave. A plane wave such as the one in Figure 1.4 (a) is an idealization that is useful in analyzing many wave phenomena. In reality, however, the electric field in a plane at right angles to \mathbf{k} does not extend to infinity since the light beam would have a finite cross-sectional area and finite power. We would need an infinitely large EM source with infinite power to generate a perfect plane wave!

In practice there are many types of possible EM waves. These waves must obey a special wave equation that describes the time and space dependence of the electric field. In an isotropic and linear dielectric medium, the relative permittivity (ϵ_r) is the same in all directions and is independent of the electric field. The field E in such a medium obeys **Maxwell's EM wave equation**

Maxwell's
wave
equation

$$\frac{\partial^2 E}{\partial x^2} + \frac{\partial^2 E}{\partial y^2} + \frac{\partial^2 E}{\partial z^2} - \epsilon_0 \epsilon_r \mu_0 \frac{\partial^2 E}{\partial t^2} = 0 \quad (1.1.5)$$

in which μ_0 is the absolute permeability, ϵ_0 is the absolute permittivity, and ϵ_r is the relative permittivity of the medium. Equation (1.1.5) assumes an isotropic medium (as discussed later) and

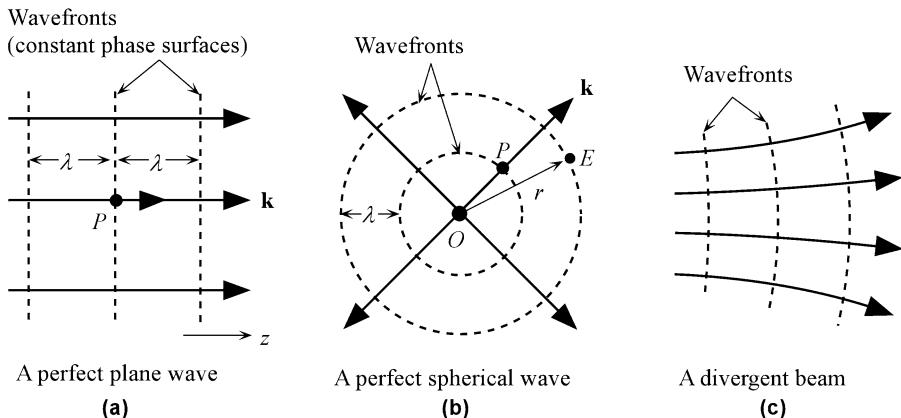


FIGURE 1.4 Examples of possible EM waves. (a) A perfect plane wave. (b) A perfect spherical wave. (c) A divergent beam.

that the conductivity of the medium is zero. To find the time and space dependence of the field, we must solve Eq. (1.1.5) in conjunction with the initial and boundary conditions. We can easily show that the plane wave in Eq. (1.1.1) satisfies Eq. (1.1.5). There are many possible waves that satisfy Eq. (1.1.5) that can therefore exist in nature.

A **spherical wave** is described by a traveling field that emerges from a point EM source and whose amplitude decays with distance r from the source. At any point r from the source, the field is given by

$$E = \frac{A}{r} \cos(\omega t - kr) \quad (1.1.6)$$

Spherical wave

in which A is a constant. We can substitute Eq. (1.1.6) into Eq. (1.1.5) to show that Eq. (1.1.6) is indeed a solution of Maxwell's equation (transformation from Cartesian to spherical coordinates would help). A cut of a spherical wave is illustrated in Figure 1.4 (b) where it can be seen that *wavefronts are spheres* centered at the point source O . The direction of propagation \mathbf{k} at any point such as P is determined by the normal to the wavefront at that point. Clearly \mathbf{k} vectors diverge out and, as the wave propagates, the constant phase surfaces become larger. **Optical divergence** refers to the angular separation of wave vectors on a given wavefront. The spherical wave has 360° of divergence in all planes through the point source. It is apparent that plane and spherical waves represent two extremes of wave propagation behavior from perfectly parallel to fully diverging wave vectors. They are produced by two extreme sizes of EM wave source: an infinitely large source for the plane wave and a point source for the spherical wave. In reality, an EM source is neither of infinite extent nor in point form, but would have a finite size and finite power. Figure 1.4 (c) shows a more practical example in which a light beam exhibits some inevitable divergence while propagating; the wavefronts are slowly bent away thereby spreading the wave. Light rays of geometric optics are drawn to be normal to constant phase surfaces (wavefronts). Light rays therefore follow the wave vector directions. Rays in Figure 1.4 (c) slowly diverge away from each other. The reason for favoring plane waves in many optical explanations is that, at a distance far away from a source, over a *small spatial region*, the wavefronts will appear to be plane even if they are actually spherical. Figure 1.4 (a) may be a small part of a huge spherical wave.

Many light beams, such as the output from a laser, can be described by assuming that they are **Gaussian beams**. Figure 1.5 illustrates a Gaussian beam traveling along the z -axis. The beam still has an $\exp(j(\omega t - kz))$ dependence to describe propagation characteristics but the amplitude varies spatially away from the beam axis and also *along* the beam axis. Such a beam has similarities to that in Figure 1.4 (c); it slowly diverges⁴ and is the result of radiation from a source of finite extent. The light intensity (*i.e.*, the radiation energy flow per unit area per unit time) distribution across the beam cross-section anywhere along z is Gaussian as shown in Figures 1.5 (b) and (c). The **beam diameter** $2w$ at any point z is defined in such a way that the cross-sectional area πw^2 at that point contains 86% of the beam power. Thus, the beam diameter $2w$ increases as the beam travels along z . The Gaussian beam shown in Figure 1.5 (a) starts from O with a finite width $2w_0$ where the wavefronts are parallel and then the beam slowly diverges as the wavefronts curve out during propagation along z . The finite width $2w_0$ where the wavefronts are parallel is called the **waist** of the beam; w_0 is the **waist radius**.

⁴The divergence is due to the self-diffraction of the beam—the beam is diffraction limited. Diffraction is covered later in this chapter. Further, the intensity of light will be defined quantitatively in Section 1.4. For the present discussion it represents the radiation energy flow per unit time per unit area.

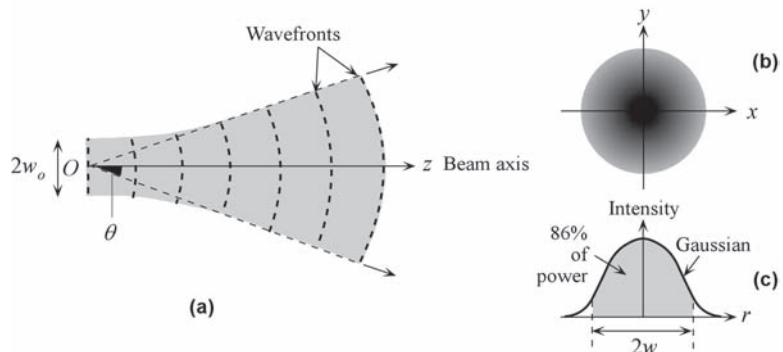


FIGURE 1.5 (a) Wavefronts of a Gaussian light beam. (b) Light intensity across a beam cross-section. (c) Light intensity vs. radial distance r from beam axis (z).

and $2w_o$ is the **spot size**. Far away from the source, the beam diameter $2w$ increases linearly with distance z . The increase in beam diameter $2w$ with z makes an angle 2θ at O , as shown in Figure 1.5, which is called the **beam divergence**. The greater the waist, the narrower the divergence. The two are related by

Far field
Gaussian
beam
divergence

$$2\theta = \frac{4\lambda}{\pi(2w_o)} \quad (1.1.7)$$

It can be shown that the Gaussian beam is a solution of Maxwell's equations when the beam divergence is small,⁵ and the intensity decays slowly with distance z . Suppose that we reflect the Gaussian beam back on itself (by using a spherical mirror that has the right curvature to match the incident wavefront) so that the beam is traveling in the $-z$ direction and converging toward O ; simply reverse the direction of travel in Figure 1.5 (a). The wavefronts would be "straightening out," and at O they would be parallel again. The beam would still have the same finite diameter $2w_o$ (waist) at O . From then on, the beam again diverges out just as it did traveling in $+z$ direction as illustrated in Figure 1.6 (a). When we try to focus a Gaussian beam using a lens or a spherical mirror, as in Figure 1.6 (a), the beam cannot be brought to a point but to a finite spot size $2w_o$. The relationship in Eq. (1.1.7) for the beam divergence can be used to find this minimum spot size $2w_o$ to which a Gaussian beam can be focused.

At a certain distance z_o from O , the beam diameter becomes $2^{1/2}(2w_o)$ as illustrated in Figure 1.6 (a). The distance z_o is called the **Rayleigh range**, and is given by

Gaussian
beam
Rayleigh
range

$$z_o = \frac{\pi w_o^2}{\lambda} \quad (1.1.8)$$

The Rayleigh range is also known as the **depth of focus**. $2z_o$ is called the **confocal parameter**. The region along z far away from the Rayleigh region, $z \gg z_o$, is called the **far-field region**.

Gaussian
beam
width at
distance

The width $2w$ of a Gaussian beam at a position along z increases with z and is given by

$$2w = 2w_o \left[1 + \left(\frac{z}{z_o} \right)^2 \right]^{1/2} = 2w_o \left[1 + \left(\frac{z\lambda}{\pi w_o^2} \right)^2 \right]^{1/2} \quad (1.1.9a)$$

⁵This is called the *paraxial approximation* in which the normals to the wavefronts, the wave vectors, make small angles with the z -axis.

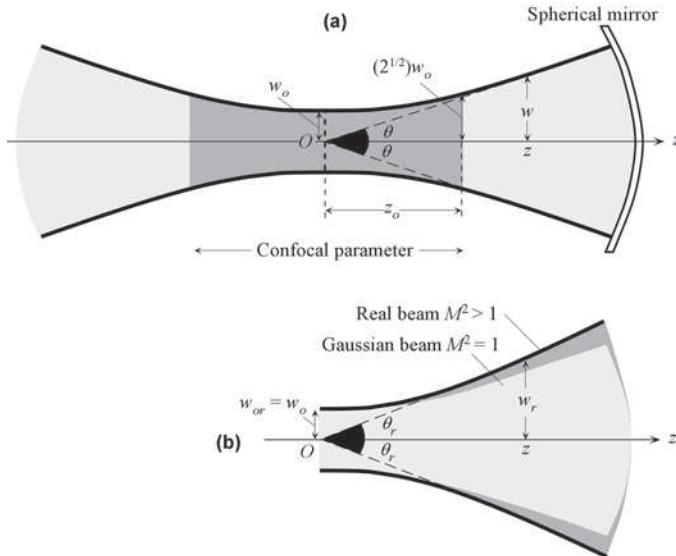


FIGURE 1.6 (a) Gaussian beam definitions. The region z_o is called the Rayleigh range and also the depth of focus. (b) Comparison of a real beam with $M^2 > 1$ with a Gaussian beam $M^2 = 1$ and the same waist $2w_o$.

Far away from the Rayleigh range, for $z \gg z_o$, in the far-field region, the beam width increases *linearly* with z , that is,

$$2w \approx (2w_o) \frac{z}{z_o} \quad (1.1.9b)$$

Gaussian beam width at distance z

Notice that the product of the beam radius w_o (half the beam waist) and half the divergence angle θ from Eq. (1.1.7) is given by $w_o\theta = \lambda/\pi$, that is, it depends only on the wavelength and is a well-defined constant for a given wavelength. The product $w_o\theta$ is called the **beam parameter product**.

The Gaussian beam concept is so useful in photonics that a special quantity, called the M^2 -factor, has been introduced to compare a given laser beam to an ideal Gaussian beam. The **M^2 factor** measures the deviation of the real laser beam from the Gaussian characteristics, in which $M^2 = 1$ for an ideal (theoretical) Gaussian beam shape. Suppose that $2\theta_r$ and $2w_{or}$ are the divergence and waist, respectively, of the *real* laser beam, and 2θ and $2w_o$ are those for the ideal Gaussian. The M^2 factor is defined by⁶

$$M^2 = \frac{w_{or}\theta_r}{w_o\theta} = \frac{w_{or}\theta_r}{(\lambda/\pi)} \quad (1.1.10) \quad M^2 \text{ factor definition}$$

where we have used $w_o\theta = \lambda/\pi$ for an ideal Gaussian beam.

According to Eq. (1.1.10), M^2 is the ratio of the beam parameter product of the real beam to that of a Gaussian beam, and hence M^2 gauges the **beam quality** of the laser beam. For many lasers, M^2 is greater than unity, and can be as high as 10–30 in multimode lasers.

⁶Some authors define M , instead of M^2 , as $M = \theta_r/\theta = w_{or}/w_o$. In addition, the reader should not be too concerned with the terms “multimode” and “single mode” at this point, except that they represent the types of radiation that is emitted from lasers.

For stable He-Ne single mode lasers, it is very close to unity. We can apply the above Gaussian equations to a real laser beam by replacing w_o with w_{or}/M^2 . For example, the width at a distance z becomes

Beam width at distance z

$$2w = 2w_{or} \left[1 + \left(\frac{z\lambda M^2}{\pi w_{or}} \right)^2 \right]^{1/2} \quad (1.1.11)$$

Far away from the Rayleigh range, $2w_r = M^2(2w)$, where $2w$ is the ideal Gaussian beam width at the same location. Suppose that we put a Gaussian beam with a waist w_o onto the real beam and adjust w_o to be the same as w_{or} , i.e., $w_{or} = w_o$. Then, from Eq. (1.1.10) the divergence of the real beam is greater inasmuch as $\theta_r = M^2\theta$, which is shown in Figure 1.6 (b).

EXAMPLE 1.1.1 A diverging laser beam

Consider a He-Ne laser beam at 633 nm with a spot size of 1 mm. Assuming a Gaussian beam, what is the divergence of the beam? What are the Rayleigh range and the beam width at 25 m?

Solution

Using Eq. (1.1.7), we find

$$2\theta = \frac{4\lambda}{\pi(2w_o)} = \frac{4(633 \times 10^{-9} \text{ m})}{\pi(1 \times 10^{-3} \text{ m})} = 8.06 \times 10^{-4} \text{ rad} = 0.046^\circ$$

The Rayleigh range is

$$z_o = \frac{\pi w_o^2}{\lambda} = \frac{\pi [(1 \times 10^{-3} \text{ m})/2]^2}{(633 \times 10^{-9} \text{ m})} = 1.24 \text{ m}$$

The beam width at a distance of 25 m is

$$\begin{aligned} 2w &= 2w_o [1 + (z/z_o)^2]^{1/2} = (1 \times 10^{-3} \text{ m}) \{1 + [(25 \text{ m})/(1.24 \text{ m})]^2\}^{1/2} \\ &= 0.0202 \text{ m} \quad \text{or} \quad 20 \text{ mm.} \end{aligned}$$

1.2 REFRACTIVE INDEX AND DISPERSION

When an EM wave is traveling in a dielectric medium, the oscillating electric field polarizes the molecules of the medium at the frequency of the wave. Indeed, the EM wave propagation can be considered to be the propagation of this polarization in the medium. The field and the induced molecular dipoles become coupled. The relative permittivity ϵ_r measures the ease with which the medium becomes polarized and hence it indicates the extent of interaction between the field and the induced dipoles. To find the nature of propagation of an EM wave in a dielectric medium, and hence the phase velocity, we have to solve Maxwell's equations in a dielectric medium. If we assume the medium is insulating, nonmagnetic, and also **isotropic**, that is the relative permittivity is independent of the direction of propagation of the EM wave and the optical field, then the solution becomes quite simple and leads to Maxwell's wave equation stated in Eq. (1.1.5). We can continue to represent the EM wave in a similar fashion to its propagation in vacuum but

we need to assign to it a new phase velocity v , and a new wavelength, both of which depend on ϵ_r . In a dielectric medium of relative permittivity ϵ_r , the phase velocity v is given by

$$v = \frac{1}{\sqrt{\epsilon_r \epsilon_0 \mu_0}} \quad (1.2.1)$$

Phase
velocity in
a medium

It is important to use the relative permittivity at the frequency of operation in Eq. (1.2.1) since ϵ_r depends on the frequency. Typical frequencies that are involved in optoelectronic devices are in the infrared (including far infrared), visible, and UV, and we generically refer to these frequencies as *optical frequencies*; they cover a somewhat arbitrary range from roughly 10^{12} Hz to 10^{16} Hz.

For an EM wave traveling in free space, $\epsilon_r = 1$ and $v_{\text{vacuum}} = 1/(\epsilon_0 \mu_0)^{1/2} = c = 3 \times 10^8 \text{ m s}^{-1}$, the velocity of light in vacuum. The ratio of the speed of light in free space to its speed in a medium is called the **refractive index** n of the medium, that is,

$$n = \frac{c}{v} = \sqrt{\epsilon_r} \quad (1.2.2)$$

Definition
of refrac-
tive index

If k is the propagation constant ($k = 2\pi/\lambda$) and λ is the wavelength, both in free space, then in the medium⁷ $k_{\text{medium}} = nk$ and $\lambda_{\text{medium}} = \lambda/n$. Equation (1.2.2) is in agreement with our intuition that light propagates more slowly in a denser medium that has a higher refractive index. We should note that the frequency v (or ω) remains the same.⁸

The refractive index of a medium is not necessarily the same in all directions. In noncrystalline materials such as glasses and liquids, the material structure is the same in all directions and n does not depend on the direction. The refractive index is then **isotropic**. In crystals, however, the atomic arrangements and interatomic bonding are different along different directions. Crystals, in general, have nonisotropic, or **anisotropic**, properties. Depending on the crystal structure, the relative permittivity ϵ_r is different along different crystal directions. This means that, in general, the refractive index n seen by a propagating electromagnetic wave in a crystal will depend on the value of ϵ_r along the direction of the oscillating electric field (*i.e.*, along the direction of polarization). For example, suppose that the wave in Figure 1.1 is traveling along the z -direction in a particular crystal with its electric field oscillating along the x -direction. If the relative permittivity along this x -direction is ϵ_{rx} , then $n_x = (\epsilon_{rx})^{1/2}$. The wave therefore propagates with a phase velocity that is c/n_x . The variation of n with direction of propagation and the direction of the electric field depends on the particular crystal structure. With the exception of cubic crystals (such as diamond), all crystals exhibit a degree of optical anisotropy that leads to a number of important applications as discussed in Chapter 6. Typically, noncrystalline solids, such as glasses and liquids, and cubic crystals are **optically isotropic**; they possess only one refractive index for all directions.

Relative permittivity ϵ_r or the dielectric constant of materials, in general, depends on the frequency of the electromagnetic wave. The relationship $n = (\epsilon_r)^{1/2}$ between the refractive index n and ϵ_r must be applied at the same frequency for both n and ϵ_r . The relative permittivity for many materials can be vastly different at high and low frequencies because different polarization

⁷On occasions, we will need to use k_o and λ_o for the free-space propagation constant and wavelength (as in the next section) and use k and λ for those values inside the medium. In each case, these quantities will be clearly defined to avoid confusion.

⁸We are accustomed to describing light in terms of its wavelength and often quote wavelengths in nm (or Å) in the visible and IR regions. However, there would be certain advantages to using the frequency instead of the wavelength, for example, terahertz (THz) instead of nm, one of which is that the frequency v does not change in the medium. (See Roger A. Lewis, *Am. J. Phys.*, 79, 341, 2011.)

TABLE 1.1 Low-frequency (LF) relative permittivity $\epsilon_r(\text{LF})$ and refractive index n

Material	$\epsilon_r(\text{LF})$	$[\epsilon_r(\text{LF})]^{1/2}$	n (at λ)	Comment
Si	11.9	3.44	3.45 (at 2.15 μm)	Electronic bond polarization up to optical frequencies
Diamond	5.7	2.39	2.41 (at 590 nm)	Electronic bond polarization up to UV light
GaAs	13.1	3.62	3.30 (at 5 μm)	Ionic polarization contributes to $\epsilon_r(\text{LF})$
SiO_2	3.84	2.00	1.46 (at 600 nm)	Ionic polarization contributes to $\epsilon_r(\text{LF})$
Water	80	8.9	1.33 (at 600 nm)	Dipolar polarization contributes to $\epsilon_r(\text{LF})$, which is large

mechanisms operate at these frequencies.⁹ At low frequencies all polarization mechanisms present can contribute to ϵ_r , whereas at optical frequencies only the electronic polarization can respond to the oscillating field. Table 1.1 lists the relative permittivity $\epsilon_r(\text{LF})$ at low frequencies (*e.g.*, 60 Hz or 1 kHz as would be measured, for example, using a capacitance bridge in the laboratory) for various materials. It then compares $[\epsilon_r(\text{LF})]^{1/2}$ with n .

For silicon and diamond there is an excellent agreement between $[\epsilon_r(\text{LF})]^{1/2}$ and n . Both are covalent solids in which electronic polarization (electronic bond polarization) is the only polarization mechanism at low and high frequencies. Electronic polarization involves the displacement of light electrons with respect to positive ions of the crystal. This process can readily respond to the field oscillations up to optical or even ultraviolet frequencies.

For GaAs and SiO_2 $[\epsilon_r(\text{LF})]^{1/2}$ is larger than n because at low frequencies both of these solids possess a degree of ionic polarization. The bonding is not totally covalent and there is a degree of ionic bonding that contributes to polarization at frequencies below far-infrared wavelengths.

In the case of water, the $\epsilon_r(\text{LF})$ is dominated by orientational or dipolar polarization, which is far too sluggish to respond to high-frequency oscillations of the field at optical frequencies.

It is instructive to consider what factors affect n . The relative permittivity depends on the polarizability α per molecule (or atom) in the solid. (α is defined as the induced electric dipole moment per unit applied field.) The simplest and approximate expression for the relative permittivity is

$$\epsilon_r \approx 1 + \frac{N\alpha}{\epsilon_0}$$

in which N is the number of molecules per unit volume. Both the atomic concentration, or density, and polarizability therefore increase n . For example, glasses of given type but with greater density tend to have higher n .

The frequency or wavelength dependence of ϵ_r and hence n is called the **dispersion relation**, or simply dispersion. There are various theoretical and empirical models that describe the n vs. λ behavior. The Cauchy dispersion equation in its simplest form is given by¹⁰

$$n = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} \quad (1.2.3)$$

Cauchy
short form
dispersion
equation

⁹Chapters 7 and 9 in *Principles of Electronic Materials and Devices*, 3rd Edition, S. O. Kasap (McGraw-Hill, 2006) provides a semiquantitative description of the frequency dependence of ϵ_r and hence the wavelength dependence of n .

¹⁰Dispersion relations like the one in Eq. (1.2.3) are always in terms of the free-space wavelength λ . (It does not make sense to give them in terms of the actual wavelength in the medium.)

TABLE 1.2 Sellmeier and Cauchy coefficients

Sellmeier	A_1	A_2	A_3	λ_1 (μm)	λ_2 (μm)	λ_3 (μm)
SiO ₂ (fused silica)	0.696749	0.408218	0.890815	0.0690660	0.115662	9.900559
86.5%SiO ₂ -13.5%GeO ₂	0.711040	0.451885	0.704048	0.0642700	0.129408	9.425478
GeO ₂	0.80686642	0.71815848	0.85416831	0.068972606	0.15396605	11.841931
Sapphire	1.023798	1.058264	5.280792	0.0614482	0.110700	17.92656
Diamond	0.3306	4.3356	—	0.1750	0.1060	—

Cauchy	Range of $h\nu$ (eV)	n_{-2} (eV ²)	n_0	n_2 (eV ⁻²)	n_4 (eV ⁻⁴)
Diamond	0.05–5.47	-1.07×10^{-5}	2.378	8.01×10^{-3}	1.04×10^{-4}
Silicon	0.002–1.08	-2.04×10^{-8}	3.4189	8.15×10^{-2}	1.25×10^{-2}
Germanium	0.002–0.75	-1.0×10^{-8}	4.003	2.2×10^{-1}	1.4×10^{-1}

Source: Sellmeier coefficients combined from various sources. Cauchy coefficients from D. Y. Smith *et al.*, *J. Phys. CM*, 13, 3883, 2001.

where A , B , and C are material-specific constants. A more general Cauchy dispersion relation is of the form

$$n = n_{-2}(h\nu)^{-2} + n_0 + n_2(h\nu)^2 + n_4(h\nu)^4 \quad (1.2.4) \quad \text{Cauchy dispersion equation in photon energy}$$

where $h\nu$ is the photon energy, and n_0 , n_{-2} , n_2 , and n_4 are constants; values for diamond, Si, and Ge are listed in Table 1.2. The general Cauchy equation is usually applicable over a wide photon energy range.

Another useful dispersion relation that has been widely used, especially in optical fibers, is the Sellmeier equation given by

$$n^2 = 1 + \frac{A_1\lambda^2}{\lambda^2 - \lambda_1^2} + \frac{A_2\lambda^2}{\lambda^2 - \lambda_2^2} + \frac{A_3\lambda^2}{\lambda^2 - \lambda_3^2} \quad (1.2.5) \quad \text{Sellmeier equation}$$

where A_1 , A_2 , A_3 and λ_1 , λ_2 , λ_3 are constants, called **Sellmeier coefficients**.¹¹ Equation (1.2.5) turns out to be quite a useful semi-empirical expression for calculating n at various wavelengths if the Sellmeier coefficients are known. Higher terms involving A_4 and higher A coefficients can generally be neglected in representing n vs. λ behavior over typical wavelengths of interest. For example, for diamond, we only need the A_1 and A_2 terms. The Sellmeier coefficients are listed in various optical data handbooks.

EXAMPLE 1.2.1 Sellmeier equation and diamond

Using the Sellmeier coefficients for diamond in Table 1.2, calculate its refractive index at 610 nm (red light) and compare with the experimental quoted value of 2.415 to three decimal places.

¹¹This is also known as the Sellmeier-Herzberger formula.

Solution

The Sellmeier dispersion relation for diamond is

$$n^2 = 1 + \frac{0.3306\lambda^2}{\lambda^2 - 175 \text{ nm}^2} + \frac{4.3356\lambda^2}{\lambda^2 - 106 \text{ nm}^2}$$

$$n^2 = 1 + \frac{0.3306(610 \text{ nm})^2}{(610 \text{ nm})^2 - (175 \text{ nm})^2} + \frac{4.3356(610 \text{ nm})^2}{(610 \text{ nm})^2 - (106 \text{ nm})^2} = 5.8308$$

So that

$$n = 2.4147$$

which is 2.415 to three decimal places and matches the experimental value.

EXAMPLE 1.2.2 Cauchy equation and diamond

Using the Cauchy coefficients for diamond in Table 1.2, calculate the refractive index at 610 nm.

Solution

At $\lambda = 610 \text{ nm}$, the photon energy is

$$h\nu = \frac{hc}{\lambda} = \frac{(6.626 \times 10^{-34} \text{ Js})(2.998 \times 10^8 \text{ m s}^{-1})}{(610 \times 10^{-9} \text{ m})} \times \frac{1}{1.602 \times 10^{-19} \text{ eV}^{-1}} = 2.0325 \text{ eV}$$

Using the Cauchy dispersion relation for diamond with coefficients from Table 1.2,

$$\begin{aligned} n &= n_{-2}(h\nu)^{-2} + n_0 + n_2(h\nu)^2 + n_4(h\nu)^4 \\ &= (-1.07 \times 10^{-5})(2.0325)^{-2} + 2.378 + (8.01 \times 10^{-3})(2.0325)^2 \\ &\quad + (1.04 \times 10^{-4})(2.0325)^4 \\ &= 2.4140 \end{aligned}$$

which is slightly different than the value calculated in Example 1.2.1; one reason for the discrepancy is due to the Cauchy coefficients quoted in Table 1.2 being applicable over a wider wavelength range at the expense of some accuracy. Although both dispersion relations have four parameters, A_1 , A_2 , λ_1 , λ_2 for Sellmeier and n_{-2} , n_0 , n_2 , n_4 for Cauchy, the functional forms are different.

1.3 GROUP VELOCITY AND GROUP INDEX

Since there are no perfect monochromatic waves in practice, we have to consider the way in which a group of waves differing slightly in wavelength will travel along the z -direction. Figure 1.7 shows how two perfectly harmonic waves of slight different frequencies $\omega - \delta\omega$ and $\omega + \delta\omega$ interfere to generate a periodic **wave packet** that contains an oscillating field at the mean frequency ω that is amplitude modulated by a slowly varying field of frequency $\delta\omega$. We are interested in the velocity of this wave packet. The two sinusoidal waves of frequencies $\omega - \delta\omega$ and $\omega + \delta\omega$ will propagate with propagation constants $k - \delta k$ and $k + \delta k$ respectively inside the material so that their sum will be

$$E_x(z, t) = E_o \cos[(\omega - \delta\omega)t - (k - \delta k)z] + E_o \cos[(\omega + \delta\omega)t - (k + \delta k)z]$$

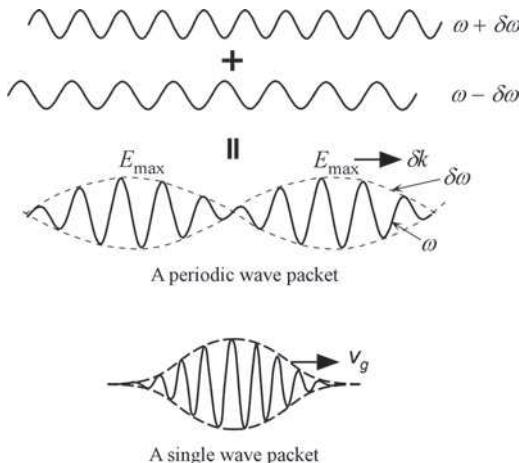


FIGURE 1.7 Two waves of slightly different wavelengths traveling in the same direction result in a wave packet that has an amplitude variation which travels at the group velocity.

By using the trigonometric identity $\cos A + \cos B = 2 \cos\left[\frac{1}{2}(A - B)\right] \cos\left[\frac{1}{2}(A + B)\right]$ we arrive at

$$E_x(z, t) = 2E_o \cos[(\delta\omega)t - (\delta k)z] \cos(\omega t - kz)$$

As illustrated in Figure 1.7, this represents a sinusoidal wave of frequency ω , which is amplitude modulated by a very slowly varying sinusoid of frequency $\delta\omega$. The system of waves, that is, the modulation, travels along z at a speed determined by the modulating term, $\cos[(\delta\omega)t - (\delta k)z]$. The maximum in the field occurs when $[(\delta\omega)t - (\delta k)z] = 2m\pi = \text{constant}$ (m is an integer), which travels with a velocity

$$\frac{\delta z}{\delta t} = \frac{\delta\omega}{\delta k}$$

or¹²

$$v_g = \frac{d\omega}{dk} \quad (1.3.1) \quad \begin{matrix} \text{Group} \\ \text{velocity} \end{matrix}$$

where the velocity v_g is the **group velocity** of the waves, since it determines the speed of propagation of the maximum electric field along z . The group velocity represents the speed with which energy or information is propagated since it defines the speed of the envelope of the amplitude variation. The maximum electric field in Figure 1.7 advances with a velocity v_g whereas the phase variations in the electric field propagate at the phase velocity v .

The wave packet we generated by adding two slightly different harmonic waves is periodic with a period $2\pi/\delta\omega$. By adding many such harmonic waves with slightly different frequencies but with the right amplitudes, we can generate a *single wave packet* that is nonperiodic as illustrated in Figure 1.7. This wave packet travels with a group velocity v_g .

¹²It is quite common to replace small changes in expressions like $\delta z/\delta t$ by differentials, and write it as dz/dt , which we will often do in the future.

We know that in vacuum, $\omega = ck$ and the group velocity is

*Group
velocity in
vacuum*

$$v_g(\text{vacuum}) = \frac{d\omega}{dk} = c = \text{phase velocity} \quad (1.3.2)$$

In vacuum or air, the group velocity is the same as the phase velocity.

For an EM wave in a medium, k in Eq. (1.3.2) is the propagation constant inside the medium, which can be written $k = 2\pi n/\lambda_o$ where λ_o is the free space wavelength. The group velocity then is not necessarily the same as the phase velocity v , which depends on ω/k and is given by c/n . The group velocity v_g , on the other hand, is $d\omega/dk$, which depends on how the propagation changes in the medium, δk , with the change in frequency $\delta\omega$, and $d\omega/dk$, is not necessarily the same as ω/k when the refractive index has a wavelength dependence. Suppose that the refractive index $n = n(\lambda_o)$ is a function of (free space) wavelength λ_o , perhaps being described by one of the expressions in Section 1.2. Its gradient would be $dn/d\lambda_o$. We can easily find the group velocity, as shown in Example 1.3.1, by first finding $d\omega$ and dk in terms of dn and $d\lambda_o$, and then using Eq. (1.3.2),

*Group
velocity in
a medium*

$$v_g(\text{medium}) = \frac{d\omega}{dk} = \frac{c}{n - \lambda_o \left(\frac{dn}{d\lambda_o} \right)} \quad (1.3.3)$$

This can be written as

*Group
velocity in
a medium*

$$v_g(\text{medium}) = \frac{c}{N_g} \quad (1.3.4)$$

in which

*Group
index*

$$N_g = n - \lambda_o \frac{dn}{d\lambda_o} \quad (1.3.5)$$

is defined as the **group index of the medium**. Equation (1.3.5) defines the group refractive index N_g of a medium and determines the effect of the medium on the group velocity via Eq. (1.3.4). What is important in Eqs. (1.3.4) and (1.3.5) is the gradient of the refractive index, $dn/d\lambda_o$. If the refractive index is constant and independent of the wavelength, at least over the wavelength range of interest, then $N_g = n$; and the group and phase velocities are the same.

In general, for many materials the refractive index n and hence the group index N_g depend on the wavelength of light by virtue of ϵ_r being frequency dependent. Then, both the phase velocity v and the group velocity v_g depend on the wavelength and the medium is called a **dispersive medium**. The refractive index n and the group index N_g of pure SiO₂ (silica) glass are important parameters in optical fiber design in optical communications. Both of these parameters depend on the wavelength of light as shown in Figure 1.8. Around 1300 nm, N_g is minimum, which means that for wavelengths close to 1300 nm, N_g is wavelength independent. Thus, light waves with wavelengths around 1300 nm travel with the same group velocity and do not experience dispersion. This phenomenon is significant in the propagation of light in optical fibers as discussed in Chapter 2.

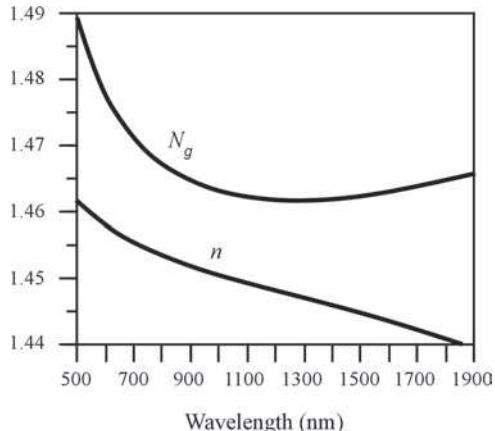


FIGURE 1.8 Refractive index n and the group index N_g of pure SiO_2 (silica) glass as a function of wavelength.

EXAMPLE 1.3.1 Group velocity

Consider two sinusoidal waves that are close in frequency, that is, waves of frequencies $\omega - \delta\omega$ and $\omega + \delta\omega$ as in Figure 1.7. Their propagation constant will be $k - \delta k$ and $k + \delta k$. The resultant wave will be

$$E_x(z, t) = E_o \cos[(\omega - \delta\omega)t - (k - \delta k)z] + E_o \cos[(\omega + \delta\omega)t - (k + \delta k)z]$$

By using the trigonometric identity $\cos A + \cos B = 2\cos[\frac{1}{2}(A - B)]\cos[\frac{1}{2}(A + B)]$ we arrive at

$$E_x(z, t) = 2E_o \cos[(\delta\omega)t - (\delta k)z] \cos(\omega t - kz)$$

As illustrated in Figure 1.7, this represents a sinusoidal wave of frequency ω , which is amplitude modulated by a very slowly varying sinusoid of frequency $\delta\omega$. The system of waves, that is, the modulation, travels along z at a speed determined by the modulating term, $\cos[(\delta\omega)t - (\delta k)z]$. The maximum in the field occurs when $[(\delta\omega)t - (\delta k)z] = 2m\pi = \text{constant}$ (m is an integer), which travels with a velocity

$$\frac{\delta z}{\delta t} = \frac{\delta\omega}{\delta k} \quad \text{or} \quad v_g = \frac{d\omega}{dk}$$

This is the group velocity of the waves, as stated in Eq. (1.3.1), since it determines the speed of propagation of the maximum electric field along z .¹³

EXAMPLE 1.3.2 Group velocity and index

Consider $\omega = 2\pi c/\lambda_o$ and $k = 2\pi n/\lambda_o$, where λ_o is the free-space wavelength. By finding expressions for $d\omega$ and dk in terms of dn and $d\lambda_o$ derive Eq. (1.3.4) for the group velocity v_g .

Solution

Differentiate $\omega = 2\pi c/\lambda_o$ to get $d\omega = -(2\pi c/\lambda_o^2)d\lambda_o$, and then differentiate $k = 2\pi n/\lambda_o$ to find

$$dk = 2\pi n(-1/\lambda_o^2)d\lambda_o + (2\pi/\lambda_o)\left(\frac{dn}{d\lambda_o}\right)d\lambda_o = -(2\pi/\lambda_o^2)\left(n - \lambda_o \frac{dn}{d\lambda_o}\right)d\lambda_o$$

¹³It is left as an exercise to show that the same result can be obtained by using a sine instead of cosine function for the waves.

We can now substitute for $d\omega$ and dk in Eq. (1.3.1),

$$v_g = \frac{d\omega}{dk} = \frac{-(2\pi c/\lambda_o^2)d\lambda_o}{-(2\pi/\lambda_o^2)\left(n - \lambda_o \frac{dn}{d\lambda_o}\right)d\lambda_o} = \frac{c}{n - \lambda_o \frac{dn}{d\lambda_o}}$$

EXAMPLE 1.3.3 Group and phase velocities

Consider a light wave traveling in a pure SiO₂ (silica) glass medium. If the wavelength of light is 1 μm and the refractive index at this wavelength is 1.450, what is the phase velocity, group index (N_g), and group velocity (v_g)?

Solution

The phase velocity is given by

$$v = c/n = (3 \times 10^8 \text{ m s}^{-1})/(1.450) = 2.069 \times 10^8 \text{ m s}^{-1}$$

From Figure 1.8, at $\lambda = 1 \mu\text{m}$, $N_g = 1.463$, so that

$$v_g = c/N_g = (3 \times 10^8 \text{ m s}^{-1})/(1.463) = 2.051 \times 10^8 \text{ m s}^{-1}$$

The group velocity is about ~0.9% smaller than the phase velocity.

1.4 MAGNETIC FIELD, IRRADIANCE, AND POYNTING VECTOR

Although we have considered the electric field component E_x of the electromagnetic wave, we should recall that the magnetic field (magnetic induction) component B_y always accompanies E_x in an EM wave propagation. In fact, if v is the phase velocity of an EM wave in an isotropic dielectric medium and n is the refractive index, then according to electromagnetism, at all times and anywhere in an EM wave,¹⁴

*Fields in
an EM
wave*

$$E_x = vB_y = \frac{c}{n}B_y \quad (1.4.1)$$

in which $v = (\epsilon_0\epsilon_r\mu_0)^{-1/2}$ and $n = \epsilon_r^{1/2}$. Thus, the two fields are simply and intimately related for an EM wave propagating in an isotropic medium. Any process that alters E_x also intimately changes B_y in accordance with Eq. (1.4.1).

As the EM wave propagates in the direction of the wave vector \mathbf{k} as shown in Figure 1.9, there is an energy flow in this direction. The wave brings with it electromagnetic energy. A small region of space in which the electric field is E_x has an energy density, that is, energy per unit volume, given by $(1/2)\epsilon_0\epsilon_rE_x^2$. Similarly, a region of space where the magnetic field is B_y has an energy density $(1/2\mu_0)B_y^2$. Since the two fields are related by Eq. (1.4.1), the energy densities in the E_x and B_y fields are the same.

*Energy
densities
in an EM
wave*

$$\frac{1}{2}\epsilon_0\epsilon_rE_x^2 = \frac{1}{2\mu_0}B_y^2 \quad (1.4.2)$$

The total energy density in the wave is therefore $\epsilon_0\epsilon_rE_x^2$. Suppose that an ideal “energy meter” is placed in the path of the EM wave so that the receiving area A of this meter is

¹⁴This is actually a statement of Faraday’s law for EM waves. In vector notation it is often expressed as $\omega\mathbf{B} = \mathbf{k} \times \mathbf{E}$.

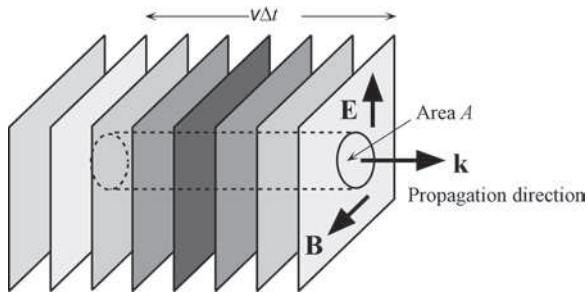


FIGURE 1.9 A plane EM wave traveling along \mathbf{k} crosses an area A at right angles to the direction of propagation. In time Δt , the energy in the cylindrical volume $Av\Delta t$ (shown dashed) flows through A .

perpendicular to the direction of propagation. In a time interval Δt , a portion of the wave of spatial length $v\Delta t$ crosses A . Thus, a volume $Av\Delta t$ of the EM wave crosses A in time Δt . The energy in this volume consequently becomes received. If S is the EM power flow per unit area, then

$$S = \text{Energy flow per unit time per unit area}$$

giving

$$S = \frac{(Av\Delta t)(\epsilon_0\epsilon_r E_x^2)}{A\Delta t} = v\epsilon_0\epsilon_r E_x^2 = v^2\epsilon_0\epsilon_r E_x B_y \quad (1.4.3)$$

In an isotropic medium, the energy flow is in the direction of wave propagation. If we use the vectors \mathbf{E} and \mathbf{B} to represent the electric and magnetic fields in the EM wave, then the wave propagates in a direction $\mathbf{E} \times \mathbf{B}$ because this direction is perpendicular to both \mathbf{E} and \mathbf{B} . The EM power flow per unit area in Eq. (1.4.3) can be written as,

$$\mathbf{S} = v^2\epsilon_0\epsilon_r \mathbf{E} \times \mathbf{B} \quad (1.4.4) \quad \begin{matrix} \text{Poynting} \\ \text{vector} \end{matrix}$$

in which \mathbf{S} , called the **Poynting vector**, represents the energy flow per unit time per unit area in a direction determined by $\mathbf{E} \times \mathbf{B}$ (direction of propagation). Its magnitude, power flow per unit area, is called the **irradiance**.¹⁵

The field E_x at the receiver location (say, $z = z_1$) varies sinusoidally, which means that the energy flow also varies sinusoidally. The irradiance in Eq. (1.4.3) is the **instantaneous irradiance**. If we write the field as $E_x = E_o \sin(\omega t)$ and then calculate the average irradiance by averaging S over one period we would find the **average irradiance**,

$$I = S_{\text{average}} = \frac{1}{2} v\epsilon_0\epsilon_r E_o^2 \quad (1.4.5) \quad \begin{matrix} \text{Average} \\ \text{irradiance} \\ (\text{intensity}) \end{matrix}$$

Since $v = c/n$ and $\epsilon_r = n^2$ we can write Eq. (1.4.5) as

$$I = S_{\text{average}} = \frac{1}{2} c\epsilon_0 n E_o^2 = (1.33 \times 10^{-3})n E_o^2 \quad (1.4.6) \quad \begin{matrix} \text{Average} \\ \text{irradiance} \\ (\text{intensity}) \end{matrix}$$

The instantaneous irradiance can be measured only if the power meter can respond more quickly than the oscillations of the electric field, and since this is in the optical frequencies range,

¹⁵The term *intensity* is widely used and interpreted by many engineers as power flow per unit area even though the strictly correct term is *irradiance*. Many optoelectronic data books simply use intensity to mean irradiance. By the way, the Poynting vector was named after John H. Poynting (1851–1914), an English physicist, who was a physics professor at the University of Birmingham (known as Mason Science College at the time).

all practical measurements invariably yield the average irradiance. This is because all detectors have a response rate much slower than the frequency of the wave.

The irradiance of an EM wave depends on the distance from the EM source. For an *ideal* plane wave, the irradiance at a distance r from the “infinite extended plane EM source” is independent of r . On the other hand, for a spherical wave, the irradiance drops as $1/r^2$ because the electric field drops as $1/r$. Suppose that the total power emitted by a point source placed at some point O is P_o as shown in Figure 1.10. Then the irradiance I at a distance r from O is

Irradiance
of a
spherical
wave

$$I = \frac{P_o}{4\pi r^2} \quad (1.4.7)$$

For a Gaussian beam, the irradiance has a Gaussian distribution across the beam cross-section, as illustrated in Figure 1.5 (b) and (c), and also decreases as the beam propagates along z ; the power in the beam becomes spread over larger and larger wavefront surfaces as the wave propagates along z . The irradiance I at a point z from O , and at a radial distance r from the beam axis (Figures 1.5 and 1.6) is given by

Irradiance
of a
Gaussian
beam

$$I(z, r) = I_o \left(\frac{w_o}{w} \right)^2 \exp \left(-\frac{2r^2}{w^2} \right) \quad (1.4.8)$$

where w_o is the beam waist, w is the beam width at a distance z from O , and I_o is the maximum beam irradiance, which occurs at $z = 0$ when $w = w_o$. Since w depends on z through $2w = 2w_o [1 + (z/z_o)^2]^{1/2}$, the irradiance also depends on z and I decreases with z . At far away from the Rayleigh range, $z \gg z_o$, the irradiance on the beam axis is

Irradiance
of a
Gaussian
beam on
axis

$$I_{\text{axis}}(z) = I_o \frac{z_o^2}{z^2} \quad (1.4.9)$$

which shows that the decay of light intensity with distance is similar to that for a spherical wave. The radial dependence, of course, remains Gaussian.

Total
power and
irradiance
of a
Gaussian
beam

The total optical power P_o is the EM power carried by a wave, and can be found by integrating the irradiance. For a Gaussian beam P_o and I_o are related by

$$P_o = \frac{1}{2} [I_o(\pi w_o^2)] \quad (1.4.10)$$

where it can be seen that the apparent “cross-sectional area” of the beam, πw_o^2 , is used to multiply the maximum irradiance with a factor of half to yield the total optical power.

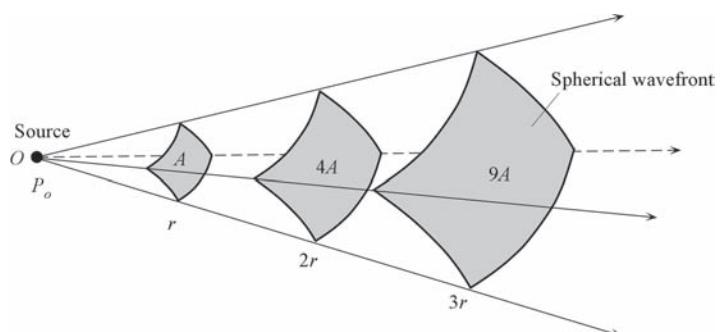


FIGURE 1.10 The irradiance of a spherical wave decreases with distance from the source because the area through which the power flows increases as r^2 .

EXAMPLE 1.4.1 Electric and magnetic fields in light

The intensity (irradiance) of the red laser beam from a He-Ne laser at a certain location has been measured to be 1 mW cm^{-2} . What are the magnitudes of the electric and magnetic fields? What are the magnitudes if this beam were in a glass medium with a refractive index $n = 1.45$ and the irradiance was still 1 mW cm^{-2} ?

Solution

Using Eq. (1.4.6) for the average irradiance, the field in air is

$$E_o = \sqrt{\frac{2I}{ce_o n}} = \sqrt{\frac{2(1 \times 10^{-3} \times 10^4 \text{ W m}^{-2})}{(3 \times 10^8 \text{ m s}^{-1})(8.85 \times 10^{-12} \text{ F m}^{-1})}} \quad (1)$$

so that

$$E_o = 87 \text{ V m}^{-1} \quad \text{or} \quad 0.87 \text{ V cm}^{-1}$$

The corresponding magnetic field is

$$B_o = E_o/c = (87 \text{ V m}^{-1})/(3 \times 10^8 \text{ m s}^{-1}) = 0.29 \mu\text{T}$$

If the same intensity were in a glass medium with $n = 1.45$, then

$$E_o(\text{medium}) = \sqrt{\frac{2I}{ce_o n}} = \sqrt{\frac{2(1 \times 10^{-3} \times 10^4 \text{ W m}^{-2})}{(3 \times 10^8 \text{ m s}^{-1})(8.85 \times 10^{-12} \text{ F m}^{-1})(1.45)}} \quad (1)$$

or

$$E_o(\text{medium}) = 72 \text{ V m}^{-1}$$

and

$$B_o(\text{medium}) = nE_o(\text{medium})/c = (1.45)(72 \text{ V m}^{-1})/(3 \times 10^8 \text{ m s}^{-1}) = 0.35 \mu\text{T}$$

EXAMPLE 1.4.2 Power and irradiance of a Gaussian beam

Consider a 5 mW He-Ne laser that is operating at 633 nm, and has a spot size of 1 mm. Find the maximum irradiance of the beam and the axial (maximum) irradiance at 25 m from the laser.

Solution

The 5 mW rating refers to the total optical power P_o available, and 633 nm is the free-space output wavelength λ . Apply Eq. (1.4.10), $P_o = (1/2)[I_o(\pi w_o^2)]$,

$$5 \times 10^{-3} \text{ W} = \frac{1}{2} I_o(\pi) \left(\frac{1}{2} \times 1 \times 10^{-3} \text{ m} \right)^2$$

which gives

$$I_o = 1.273 \times 10^4 \text{ W m}^{-2} = 1.273 \text{ W cm}^{-2}$$

The Rayleigh range z_o was calculated previously as $z_o = \pi w_o^2 / \lambda = 1.24 \text{ m}$ in Example 1.1.1. At $z = 25 \text{ m}$, the axial irradiance is

$$I_{\text{axis}} = (1.273 \times 10^4 \text{ W m}^{-2}) \frac{(1.24 \text{ m})^2}{(25 \text{ m})^2} = 31.3 \text{ W m}^{-2} = 3.13 \text{ mW cm}^2$$

1.5 SNELL'S LAW AND TOTAL INTERNAL REFLECTION (TIR)



WILLEBROD SNELL.



Willebrord Snellius (Willebrord Snel van Royen, 1580–1626) was a Dutch astronomer and a mathematician, who was a professor at the University of Leiden. He discovered his law of refraction in 1621 which was published by René Descartes in France 1637; it is not known whether Descartes knew of Snell's law or formulated it independently. (*Courtesy of AIP Emilio Segre Visual Archives, Brittle Books Collection.*)

René Descartes (1596–1650) was a French philosopher who was also involved with mathematics and sciences. He has been called the “Father of Modern Philosophy.” Descartes was responsible for the development of Cartesian coordinates and analytical geometry. He also made significant contributions to optics, including reflection and refraction. (*Courtesy of Georgios Kollidas/Shutterstock.com.*)

We consider a traveling plane EM wave in a medium (1) of refractive index n_1 propagating toward a medium (2) with a refractive index n_2 . Constant phase fronts are joined with broken lines and the wave vector \mathbf{k}_i is perpendicular to the wavefronts as shown in Figure 1.11. When the wave reaches the plane boundary between the two media, a transmitted wave in medium 2 and a reflected wave in medium 1 appear. The transmitted wave is called the **refracted light**. The angles θ_i , θ_t , θ_r define the directions of the incident, transmitted, and reflected waves, respectively, with respect to the normal to the boundary plane as shown in Figure 1.11. The wave vectors of the reflected and transmitted waves are denoted as \mathbf{k}_r and \mathbf{k}_t . Since both the incident and reflected waves are in the same medium, the magnitudes of \mathbf{k}_r and \mathbf{k}_i are the same, $k_r = k_i$.

Simple arguments based on constructive interference can be used to show that there can be only one reflected wave that occurs at an angle equal to the incidence angle. The two waves along A_i and B_i are in phase. When these waves are reflected to become waves A_r and B_r then they must still be in phase, otherwise they will interfere destructively and destroy each other. The only way the two waves can stay in phase is if $\theta_r = \theta_i$. All other angles lead to the waves A_r and B_r being out of phase and interfering destructively.

The refracted waves A_t and B_t are propagating in a medium of refracted index $n_2 (< n_1)$ that is different than n_1 . Hence the waves A_t and B_t have different velocities than A_i and B_i . We consider what happens to a wavefront such as AB , corresponding perhaps to the maximum field, as it propagates from medium 1 to 2. We recall that the points A and B on this front are always in phase. During the time it takes for the phase B on wave B_i to reach B' , phase A on wave A_i has

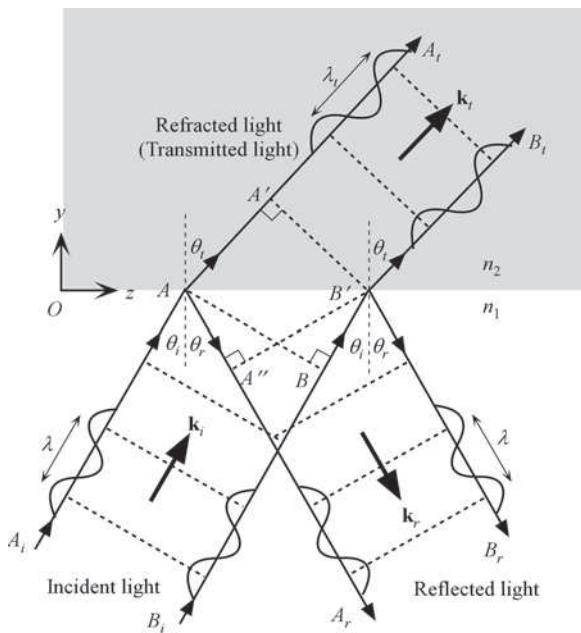


FIGURE 1.11 A light wave traveling in a medium with a greater refractive index ($n_1 > n_2$) suffers reflection and refraction at the boundary. (Notice that λ_t is slightly longer than λ .)

progressed to A' . The wavefront AB thus becomes the front $A'B'$ in medium 2. Unless the two waves at A' and B' still have the same phase, there will be no transmitted wave. A' and B' points on the front are in phase only for one particular transmitted angle, θ_t .

If it takes time t for the phase at B on wave B_i to reach B' , then $BB' = v_1 t = ct/n_1$. During this time t , the phase A has progressed to A' where $AA' = v_2 t = ct/n_2$. A' and B' belong to the same front just like A and B so that AB is perpendicular to \mathbf{k}_i in medium 1 and $A'B'$ is perpendicular to \mathbf{k}_t in medium 2. From geometrical considerations, $AB' = BB'/\sin \theta_t$ and $AB' = AA'/\sin \theta_t$ so that

$$AB' = \frac{v_1 t}{\sin \theta_t} = \frac{v_2 t}{\sin \theta_t}$$

or

$$\frac{\sin \theta_i}{\sin \theta_t} = \frac{v_1}{v_2} = \frac{n_2}{n_1} \quad (1.5.1) \quad \text{Snell's Law}$$

This is **Snell's law**,¹⁶ which relates the angles of incidence and refraction to the refractive indices of the media.

If we consider the reflected wave, the wavefront AB becomes $A''B'$ in the reflected wave. In time t , phase B moves to B' and A moves to A'' . Since they must still be in phase to constitute the reflected wave, BB' must be equal to AA'' . Suppose it takes time t for the wavefront B to move to B' (or A to A''). Then, since $BB' = AA'' = v_1 t$, from geometrical considerations,

$$AB' = \frac{v_1 t}{\sin \theta_i} = \frac{v_1 t}{\sin \theta_r}$$

so that $\theta_i = \theta_r$. Angles of incidence and reflection are the same.

¹⁶Snell's law is known as Descartes's law in France as he was the first to publish it in his "Discourse on Method" in 1637.

When $n_1 > n_2$ then obviously the transmitted angle is greater than the incidence angle as apparent in Figure 1.11. When the refraction angle θ_t reaches 90° , the incidence angle is called the **critical angle** θ_c , which is given by

Total
internal
reflection
(TIR)

$$\sin \theta_c = \frac{n_2}{n_1} \quad (1.5.2)$$

When the incidence angle θ_i exceeds θ_c then there is no transmitted wave but only a reflected wave. The latter phenomenon is called **total internal reflection (TIR)**. The effect of increasing the incidence angle is shown in Figure 1.12. It is the TIR phenomenon that leads to the propagation of waves in a dielectric medium surrounded by a medium of smaller refractive index as shown in Chapter 2. Although Snell's law for $\theta_i > \theta_c$ shows that $\sin \theta_t > 1$ and hence θ_t is an "imaginary" angle of refraction, there is however a wave called the evanescent wave, whose amplitude decays exponentially with distance into the second medium as discussed below. The wave exists only in the interface region from which the reflected wave emerges (not outside).

Snell's law can also be viewed as the \mathbf{k} -vector of light parallel to the interface being continuous through the interface, that is, having the same value on both sides of the interface. In medium n_1 , \mathbf{k}_i parallel to the interface is $k_i \sin \theta_i$ or $kn_1 \sin \theta_i$, where $k_i = kn_1$, and k is the magnitude of the wave vector in free space. In medium n_2 , \mathbf{k}_t parallel to the interface is $k_t \sin \theta_t$ or $kn_2 \sin \theta_t$. If \mathbf{k} 's component tangential to the interface remains constant, $kn_1 \sin \theta_i = kn_2 \sin \theta_t$, then we obtain Snell's law in Eq. (1.5.1). Put differently, Snell's law is equivalent to

Snell's
Law

$$n \sin \theta = \text{constant through an interface between different media} \quad (1.5.3)$$

Snell's law of refraction and TIR play a very important role in many optoelectronic and photonic devices. A prism is a transparent optical component that can deflect a light beam as illustrated in Figure 1.13. There are two basic types of prism. In a **refracting prism**, the light deflection is caused by refractions whereas in a **reflecting prism** it is caused by one or more TIRs. (Some prisms such as composite prisms need both refraction and TIR to achieve their desired deflection.) The deflection δ depends not only on the incidence angle of the light beam on the prism, the prism material (n), and geometry, but also on the wavelength and the polarization state of the incident light. The reason is that the refractive index n of the prism material normally depends on the wavelength, and further, for certain materials (e.g., quartz, calcite), it depends on the polarization state (direction of the electric field) of light as well.

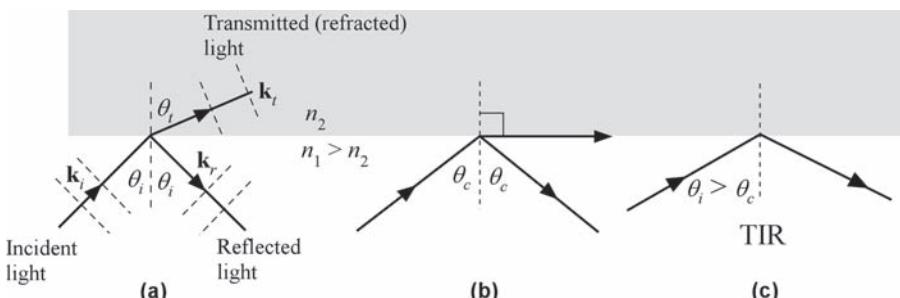


FIGURE 1.12 Light wave travelling in a more dense medium strikes a less dense medium. Depending on the incidence angle with respect to θ_c , which is determined by the ratio of the refractive indices, the wave may be transmitted (refracted) or reflected. (a) $\theta_i < \theta_c$ (b) $\theta_i = \theta_c$ (c) $\theta_i > \theta_c$ and total internal reflection. (Wavefronts are only indicated in (a).)

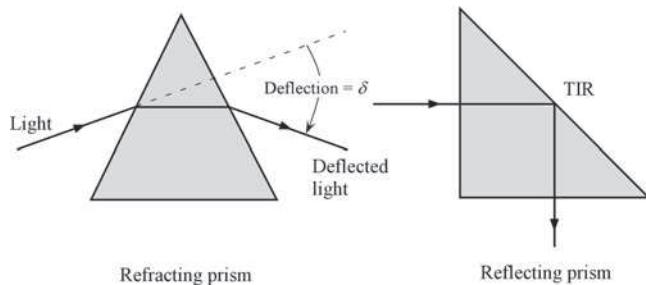


FIGURE 1.13 Basic types of prism: refracting and reflecting prisms.

EXAMPLE 1.5.1 Beam displacement

Lateral displacement of light, or **beam displacement**, occurs when a beam of light passes obliquely through a plate of transparent material, such as a glass plate. When a light beam is incident on a plate of transparent material of refractive index n , it emerges from the other side traveling parallel to the incident light but displaced from it by a distance d , called *lateral displacement*, as illustrated in Figure 1.14. Find the displacement d in terms of the incidence angle the plate thickness. What is d for a glass of $n = 1.600$, $d = 10 \text{ mm}$ if the incidence angle is 45° ?

Solution

The displacement $d = BC = AB \sin(\theta_i - \theta_t)$. Further, $L/AB = \cos \theta_t$ so that combining these two equations we find

$$d = L \left[\frac{\sin(\theta_i - \theta_t)}{\cos \theta_t} \right]$$

We can expand $\sin(\theta_i - \theta_t) = \sin \theta_i \cos \theta_t - \cos \theta_i \sin \theta_t$, use $\cos \theta_t = \sqrt{1 - \sin^2 \theta_t}$ and then apply Snell's law $n \sin \theta_t = n_o \sin \theta_i$ at the top surface to find

$$\frac{d}{L} = \sin \theta_i \left[1 - \frac{\cos \theta_i}{\sqrt{(n/n_o)^2 - \sin^2 \theta_i}} \right]$$

which is maximum with $d = L$ when $\theta_i \approx 90^\circ$, glazing incidence. Substituting $n = 1.600$, $n_o = 1$, $\theta_i = 45^\circ$, and $L = 10 \text{ mm}$, we find, $d = 3.587 \text{ mm}$. If the refractive index increases by 1%, $n = 1.616$, then $d = 3.630$ and the change in d is 0.043 mm or $43 \mu\text{m}$, which can be measured electronically by using, for example, CCD or CMOS photodiode arrays.

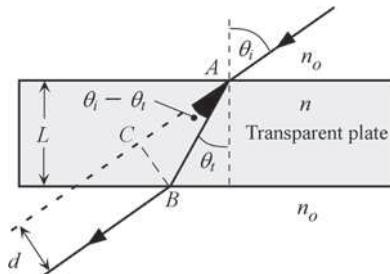


FIGURE 1.14 Lateral displacement of light passing obliquely through a transparent plate.

1.6 FRESNEL'S EQUATIONS

A. Amplitude Reflection and Transmission Coefficients (r and t)

Although the ray picture with constant phase wavefronts is useful in understanding refraction and reflection, to obtain the magnitude of the reflected and refracted waves and their relative phases, we need to consider the electric field in the light wave. The electric field in the wave must be perpendicular to the direction of propagation as shown in Figure 1.15. We can resolve the field E_i of the incident wave into two components, one in the plane of incidence, $E_{i,\parallel}$, and the other perpendicular to the plane of incidence, $E_{i,\perp}$. The plane of incidence is defined as the plane containing the incident and the reflected rays, which in Figure 1.15 corresponds to the plane of the paper.¹⁷ Similarly for both the reflected and transmitted waves, we will have field components parallel and perpendicular to the plane of incidence, that is, $E_{r,\parallel}$, $E_{r,\perp}$, and $E_{t,\parallel}$, $E_{t,\perp}$.

As apparent from Figure 1.15, the incident, transmitted, and reflected wave all have a wave vector component along the z -direction; that is, they have an effective velocity along z . The fields $E_{i,\perp}$, $E_{r,\perp}$, and $E_{t,\perp}$ are all perpendicular to the z -direction. These waves are called **transverse electric field** (TE) waves. On the other hand, waves with $E_{i,\parallel}$, $E_{r,\parallel}$, and $E_{t,\parallel}$ have only their magnetic field components perpendicular to the z -direction, and these are called **transverse magnetic field** (TM) waves.

We will describe the incident, reflected, and refracted waves each by the exponential representation of a traveling wave:

Incident
wave

$$E_i = E_{io} \exp(j(\omega t - \mathbf{k}_i \cdot \mathbf{r})) \quad (1.6.1)$$

Reflected
wave

$$E_r = E_{ro} \exp(j(\omega t - \mathbf{k}_r \cdot \mathbf{r})) \quad (1.6.2)$$

Transmitted
wave

$$E_t = E_{to} \exp(j(\omega t - \mathbf{k}_t \cdot \mathbf{r})) \quad (1.6.3)$$

in which \mathbf{r} is the position vector, the wave vectors \mathbf{k}_i , \mathbf{k}_r , and \mathbf{k}_t describe the directions of the incident, reflected, and transmitted waves, and E_{io} , E_{ro} , and E_{to} are the respective amplitudes. Any phase changes such as ϕ_r and ϕ_t in the reflected and transmitted waves with respect to the phase of the incident wave are incorporated into the complex amplitudes, E_{ro} and E_{to} . Our objective is to find E_{ro} and E_{to} with respect to E_{io} .

We should note that similar equations can be stated for the magnetic field components in the incident, reflected, and transmitted waves but these will be perpendicular to the corresponding electric fields. The electric and magnetic fields anywhere on the wave must be perpendicular to each other as a requirement of electromagnetic wave theory. This means that with E_{\parallel} in the EM wave we have a magnetic field B_{\perp} associated with it such that $B_{\perp} = (n/c)E_{\parallel}$. Similarly, E_{\perp} will have a magnetic field B_{\parallel} associated with it such that $B_{\parallel} = (n/c)E_{\perp}$.

There are two useful fundamental rules in electromagnetism that govern the behavior of the electric and magnetic fields at a boundary between two dielectric media, which we can arbitrarily label as 1 and 2. These rules are called **boundary conditions**. The first states that the

¹⁷The definitions of the field components follow those of S. G. Lipson *et al.*, *Optical Physics*, 3rd Edition (Cambridge University Press, 1995) and Grant Fowles, *Introduction to Modern Optics*, 2nd Edition (Dover Publications, Inc., 1975), whose clear treatments of this subject are highly recommended. The majority of authors use a different convention, which leads to different signs later in the equations (“they [Fresnel’s equations] must be related to the specific electric field directions from which they are derived.” Eugene Hecht, *Optics*, 4th Edition (Addison Wesley, Pearson Education, 2002), p. 115.

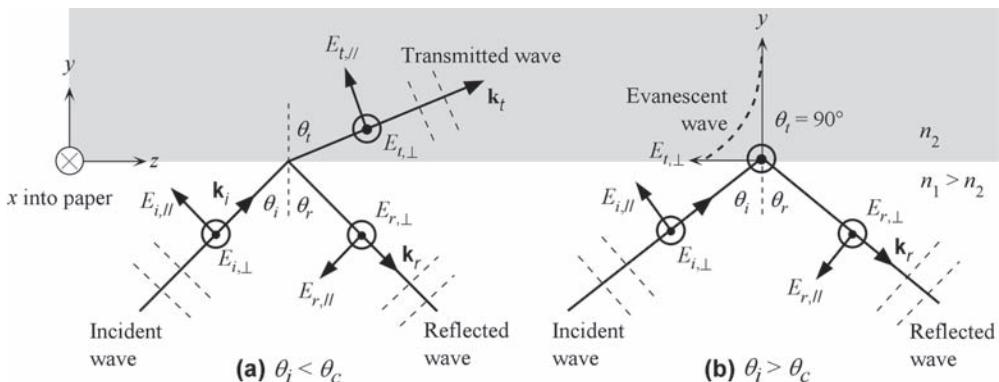


FIGURE 1.15 Light wave traveling in a more dense medium strikes a less dense medium. The plane of incidence is the plane of the paper and is perpendicular to the flat interface between the two media. The electric field is normal to the direction of propagation. It can be resolved into perpendicular (\perp) and parallel ($/\!\!/$) components. (a) $\theta_i < \theta_c$ then some of the wave is transmitted into the less dense medium. Some of the wave is reflected. (b) $\theta_i > \theta_c$ then the incident wave suffers total internal reflection. There is a decaying evanescent wave into the n_2 -medium.

electric field that is tangential to the boundary surface, $E_{\text{tangential}}$, must be continuous across the boundary from medium 1 to 2, that is, at the boundary, $y = 0$ in Figure 1.15,

$$E_{\text{tangential}}(1) = E_{\text{tangential}}(2) \quad (1.6.4) \quad \text{Boundary condition}$$

The second rule is that the tangential component of the magnetic field, $B_{\text{tangential}}$, to the boundary must likewise be continuous from medium 1 to 2, provided that the two media are non-magnetic (relative permeability, $\mu_r = 1$),

$$B_{\text{tangential}}(1) = B_{\text{tangential}}(2) \quad (1.6.5) \quad \text{Boundary condition}$$

Using the boundary conditions above for the fields at $y = 0$ and the relationship between the electric and magnetic fields, we can find the reflected and transmitted waves in terms of the incident wave. The boundary conditions can be satisfied only if the reflection and incidence angles are equal, $\theta_r = \theta_i$, and the angles for the transmitted and incident wave obey Snell's law, $n_1 \sin \theta_1 = n_2 \sin \theta_2$.

Applying the boundary conditions above to the EM wave going from medium 1 to 2, the amplitudes of the reflected and transmitted waves can be readily obtained in terms of n_1 , n_2 , and the incidence angle θ_i alone.¹⁸ These relationships are called **Fresnel's equations**. If we define $n = n_2/n_1$ as the relative refractive index of medium 2 to that of 1, then the **reflection and transmission coefficients** for E_{\perp} are

$$r_{\perp} = \frac{E_{ro,\perp}}{E_{io,\perp}} = \frac{\cos \theta_i - [n^2 - \sin^2 \theta_i]^{1/2}}{\cos \theta_i + [n^2 - \sin^2 \theta_i]^{1/2}} \quad (1.6.6a) \quad \text{Reflection coefficient}$$

¹⁸These equations are readily available in any electromagnetism textbook. Their derivation from the two boundary conditions involves extensive algebraic manipulation, which we will not carry out here. The electric and magnetic field components on both sides of the boundary are resolved tangentially to the boundary surface and the boundary conditions are then applied. We then use such relations as $\cos \theta_t = [1 - \sin^2 \theta_t]^{1/2}$ and $\sin \theta_t$ is determined by Snell's law, etc.

and

Transmission coefficient

$$t_{\perp} = \frac{E_{to,\perp}}{E_{io,\perp}} = \frac{2\cos\theta_i}{\cos\theta_i + [n^2 - \sin^2\theta_i]^{1/2}} \quad (1.6.6b)$$

There are corresponding coefficients for the E_{\parallel} fields with corresponding **reflection and transmission coefficients**, r_{\parallel} and t_{\parallel} .

Reflection coefficient

$$r_{\parallel} = \frac{E_{ro,\parallel}}{E_{io,\parallel}} = \frac{[n^2 - \sin^2\theta_i]^{1/2} - n^2\cos\theta_i}{[n^2 - \sin^2\theta_i]^{1/2} + n^2\cos\theta_i} \quad (1.6.7a)$$

Transmission coefficient

$$t_{\parallel} = \frac{E_{to,\parallel}}{E_{io,\parallel}} = \frac{2n\cos\theta_i}{n^2\cos\theta_i + [n^2 - \sin^2\theta_i]^{1/2}} \quad (1.6.7b)$$

Further, the coefficients above are related by

Transmission coefficient

$$r_{\parallel} + nt_{\parallel} = 1 \quad \text{and} \quad r_{\perp} + 1 = t_{\perp} \quad (1.6.8)$$

The significance of these equations is that they allow the amplitudes and phases of the reflected and transmitted waves to be determined from the coefficients r_{\perp} , r_{\parallel} , t_{\parallel} , and t_{\perp} . For convenience we take E_{io} to be a real number so that phase angles of r_{\perp} and t_{\perp} correspond to the **phase changes** measured with respect to the incident wave. For example, if r_{\perp} is a complex quantity then we can write this as $r_{\perp} = |r_{\perp}| \exp(-j\phi_{\perp})$ in which $|r_{\perp}|$ and ϕ_{\perp} represent the relative amplitude and phase of the reflected wave with respect to the incident wave for the field perpendicular to the plane of incidence. Of course, when r_{\perp} is a real quantity, then a positive number represents no phase shift and a negative number is a phase shift of 180° (or π). As with all waves, a negative sign corresponds to a 180° phase shift. Complex coefficients can be obtained only from Fresnel's equations if the terms under the square roots become negative and this can happen only when $n < 1$ (or $n_1 > n_2$), and also when $\theta_i > \theta_c$, the critical angle. Thus, phase changes other than 0 or 180° occur only when there is total internal reflection. Fresnel's equations for normal incidence are greatly simplified. Putting $\theta_i = 0$ into Eqs. (1.6.6) and (1.6.7) we find

Normal incidence

$$r_{\parallel} = r_{\perp} = \frac{n_1 - n_2}{n_1 + n_2} \quad \text{and} \quad t_{\parallel} = t_{\perp} = \frac{2n_1}{n_1 + n_2} \quad (1.6.9)$$

Figure 1.16 (a) shows how the magnitudes of the reflection coefficients, $|r_{\perp}|$ and $|r_{\parallel}|$, vary with the incidence angle θ_i for a light wave traveling from a more dense medium, $n_1 = 1.44$, to a less dense medium, $n_2 = 1.00$, as predicted by Fresnel's equations. Figure 1.16 (b) shows the changes in the phase of the reflected wave, ϕ_{\perp} and ϕ_{\parallel} , with θ_i . The critical angle θ_c as determined from $\sin\theta_c = n_2/n_1$ in this case is 44° . It is clear that for incidence close to normal (small θ_i), there is no phase change in the reflected wave. The reflection coefficient in Eq. (1.6.9) is a positive quantity for $n_1 > n_2$, which means that the reflected wave suffers no phase change. This is confirmed by ϕ_{\perp} and ϕ_{\parallel} in Figure 1.16 (b). As the incidence angle increases, eventually r_{\parallel} becomes zero at an angle of about 35° . We can find this special incidence angle, labeled as θ_p , by solving the Fresnel equation (1.6.7a) for $r_{\parallel} = 0$. The field in the reflected wave is then always perpendicular to the plane of incidence and hence well-defined as illustrated in

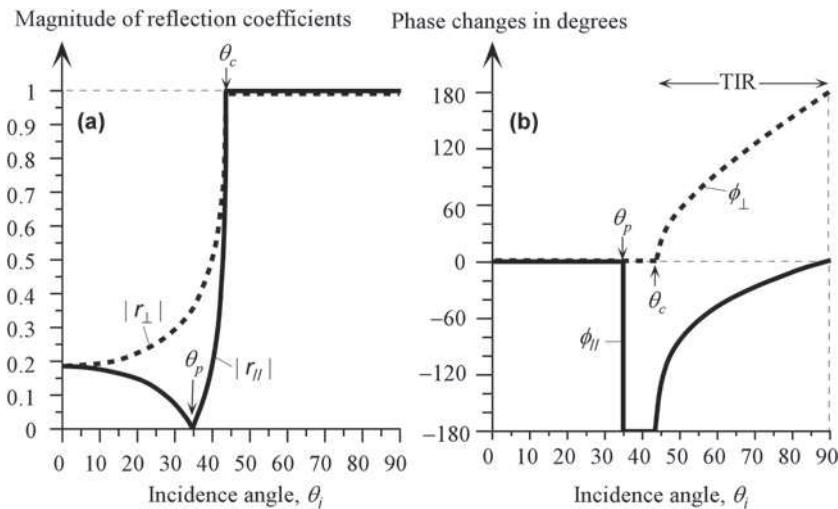


FIGURE 1.16 Internal reflection: (a) Magnitude of the reflection coefficients r_{\parallel} and r_{\perp} vs. angle of incidence θ_i for $n_1 = 1.44$ and $n_2 = 1.00$. The critical angle is 44° . (b) The corresponding phase changes ϕ_{\parallel} and ϕ_{\perp} vs. incidence angle θ_i .

Figure 1.17. This special angle is called the **Brewster's¹⁹ angle** or **Brewster's angle** and from Eq. (1.6.7a) is given by

$$\tan \theta_p = \frac{n_2}{n_1} \quad (1.6.10)$$

Brewster's polarization angle

The reflected wave is then said to be **linearly polarized** because it contains *electric field oscillations that are contained within a well-defined plane*, which is perpendicular to the plane of incidence and also to the direction of propagation. Electric field oscillations in **unpolarized light**, on the other hand, can be in any one of infinite number of directions that are perpendicular

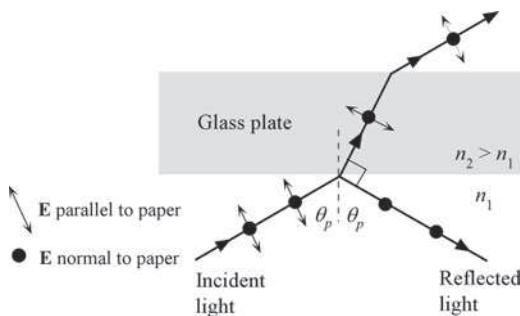


FIGURE 1.17 At the Brewster angle of incidence $\theta_i = \theta_p$, the reflected light contains only field oscillations normal to the plane of incidence (paper).

¹⁹ After Sir David Brewster (1781–1868), a Scottish physicist who was educated in theology at the University of Edinburgh in Scotland. He became interested in the polarization properties of light from 1799 onwards, and reported some of his experiments in scientific journals, including the *Philosophical Transactions of London*.

to the direction of propagation. In linearly polarized light, however, the field oscillations are contained within a well-defined plane. Light emitted from many light sources, such as a tungsten light bulb or an LED diode, is unpolarized.²⁰ Unpolarized light can be viewed as a stream or collection of EM waves whose fields are randomly oriented in a direction that is perpendicular to the direction of light propagation.

For incidence angles greater than θ_p but smaller than θ_c , Fresnel's equation (1.6.7a) gives a negative number for r_{\parallel} , which indicates a phase shift of 180° as shown in ϕ_{\parallel} in Figure 1.16 (b). The magnitude of both r_{\parallel} and r_{\perp} increases with θ_i as illustrated in Figure 1.16 (a). At the critical angle and beyond (past 44° in Figure 1.16), that is, when $\theta_i \geq \theta_c$, the magnitudes of both r_{\parallel} and r_{\perp} go to unity so that the reflected wave has the same amplitude as the incident wave. The incident wave has suffered **total internal reflection**, TIR. When $\theta_i > \theta_c$, in the presence of TIR, both Eqs. (1.6.6) and (1.6.7) are complex quantities because then $\sin \theta_i > n$ and the terms under the square roots become negative. The reflection coefficients become complex quantities of the type $r_{\perp} = 1 \cdot \exp(-j\phi_{\perp})$ and $r_{\parallel} = 1 \cdot \exp(-j\phi_{\parallel})$ with the phase angles ϕ_{\perp} and ϕ_{\parallel} being other than zero or 180° . The reflected wave therefore suffers phase changes, ϕ_{\perp} and ϕ_{\parallel} , in the components E_{\perp} and E_{\parallel} . These phase changes depend on the incidence angle, as illustrated in Figure 1.16 (b), and on n_1 and n_2 .

Examination of Eq. (1.6.6) for r_{\perp} shows that for $\theta_i > \theta_c$, we have $|r_{\perp}| = 1$, but the phase change ϕ_{\perp} is given by

Phase
change
in TIR

$$\tan\left(\frac{1}{2}\phi_{\perp}\right) = \frac{[\sin^2\theta_i - n^2]^{1/2}}{\cos\theta_i} \quad (1.6.11)$$

For the E_{\parallel} component, the phase change ϕ_{\parallel} is given by

Phase
change
in TIR

$$\tan\left(\frac{1}{2}\phi_{\parallel} + \frac{1}{2}\pi\right) = \frac{[\sin^2\theta_i - n^2]^{1/2}}{n^2 \cos\theta_i} \quad (1.6.12)$$

We can summarize that in internal reflection ($n_1 > n_2$), the amplitude of the reflected wave from TIR is equal to the amplitude of the incident wave but its phase has shifted by an amount determined by Eqs. (1.6.11) and (1.6.12). The fact that ϕ_{\parallel} has an additional π shift that makes ϕ_{\parallel} negative for $\theta_i > \theta_c$ is due to the choice for the direction of the reflected optical field $E_{r,\parallel}$ in Figure 1.15. This π shift can be ignored if we by simply invert $E_{r,\parallel}$. (In many books Eq. (1.6.12) is written without the π -shift.)

The reflection coefficients in Figure 1.16 considered the case in which $n_1 > n_2$. When light approaches the boundary from the higher index side, that is $n_1 > n_2$, the reflection is said to be **internal reflection** and *at normal incidence there is no phase change*. On the other hand, if light approaches the boundary from the lower index side, that is $n_1 < n_2$, then it is called **external reflection**. Thus in external reflection light becomes reflected by the surface of an optically denser (higher refractive index) medium. There is an important difference between the two. Figure 1.18 shows how the reflection coefficients r_{\perp} and r_{\parallel} depend on the incidence angle θ_i for external reflection ($n_1 = 1$ and $n_2 = 1.44$). At normal incidence, both coefficients are negative, which means that *in external reflection at normal incidence there is a phase shift of 180°* . Further, r_{\parallel} goes through zero at the **Brewster angle** θ_p given by Eq. (1.6.10). At this

²⁰Even light from a tungsten light bulb or an LED has some polarization but this is usually negligibly small.

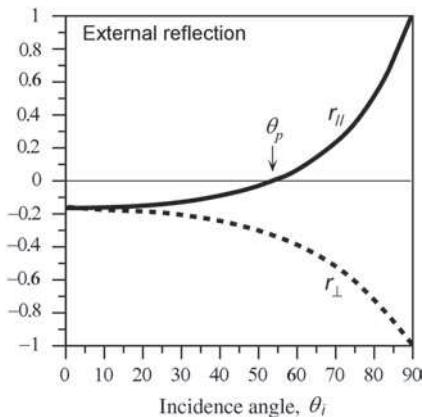


FIGURE 1.18 External reflection coefficients r_{\parallel} and r_{\perp} vs. angle of incidence θ_i for $n_1 = 1.00$ and $n_2 = 1.44$.

angle of incidence, the reflected wave is polarized in the E_{\perp} component only. Transmitted light in both internal reflection (when $\theta_i < \theta_c$) and external reflection does not experience a phase shift.

What happens to the transmitted wave when $\theta_i > \theta_c$? According to the boundary conditions, there must still be an electric field in medium 2; otherwise, the boundary conditions cannot be satisfied. When $\theta_i > \theta_c$, the field in medium 2 is a wave that travels near the surface of the boundary along the z direction as shown in Figure 1.19. The wave is called an **evanescent wave** and advances along z with its field decreasing as we move into medium 2; that is,

$$E_{t,\perp}(y, z, t) \propto e^{-\alpha_2 y} \exp(j(\omega t - k_i z)) \quad (1.6.13)$$

Evanescent wave

in which $k_i z = k_i \sin \theta_i$ is the wave vector of the incident wave along the z -axis, and α_2 is an **attenuation coefficient** for the electric field penetrating into medium 2,²¹

$$\alpha_2 = \frac{2\pi n_2}{\lambda_o} \left[\left(\frac{n_1}{n_2} \right)^2 \sin^2 \theta_i - 1 \right]^{1/2} \quad (1.6.14)$$

Attenuation of evanescent wave

in which λ is the free-space wavelength. According to Eq. (1.6.13), the evanescent wave travels along z and has an amplitude that decays exponentially as we move from the boundary into medium 2 (along y). The field of the evanescent wave is e^{-1} in medium 2 when $y = 1/\alpha_2 = \delta$, which is called the **penetration depth**. It is not difficult to show that the evanescent wave is correctly predicted by Snell's law when $\theta_i > \theta_c$. The evanescent wave propagates along the boundary (along z) with the same speed as the z -component velocity of the incident and reflected waves. In Eqs. (1.6.1) and (1.6.2) we had assumed that the incident and reflected waves were *plane waves*, that is, of *infinite extent*. If we were to extend the plane wavefronts on the reflected wave, these would cut the boundary as shown in Figure 1.19. The evanescent wave traveling along z can be thought of arising from these plane wavefronts at the boundary as in Figure 1.19. (Evanescent wave is important in light propagation in optical waveguides such as optical fibers.) If the incident wave is a narrow

²¹Normally the term *attenuation coefficient* refers to the attenuation of the irradiance but in this case it refers to the electric field.

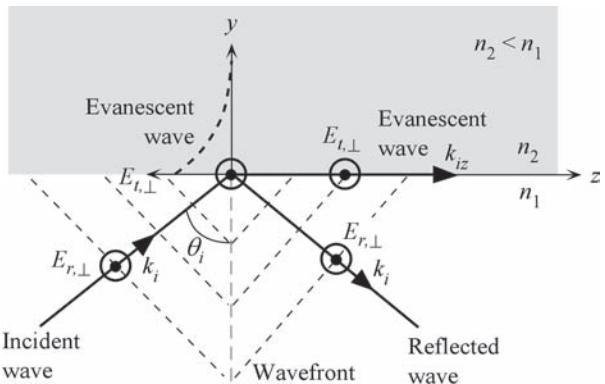


FIGURE 1.19 When $\theta_i > \theta_c$ for a *plane wave* that is reflected, there is an evanescent wave at the boundary whose magnitude decays into the n_2 -medium.

beam of light (e.g., from a laser pointer) then the reflected beam would have the same cross-section. There would still be an evanescent wave at the boundary, but it would exist only within the cross-sectional area of the reflected beam at the boundary.

B. Intensity, Reflectance, and Transmittance

It is frequently necessary to calculate the intensity or irradiance²² of the reflected and transmitted waves when light traveling in a medium of index n_1 is incident at a boundary where the refractive index changes to n_2 . In some cases we are simply interested in normal incidence where $\theta_i = 0^\circ$. For example, in laser diodes light is reflected from the ends of an optical cavity where there is a change in the refractive index.

For a light wave traveling with a velocity v in a medium with relative permittivity ϵ_r , the light intensity I is defined in terms of the electric field amplitude E_o as

Light intensity or irradiance

$$I = \frac{1}{2} v \epsilon_r \epsilon_o E_o^2 \quad (1.6.15)$$

Here $\frac{1}{2} \epsilon_r \epsilon_o E_o^2$ represents the energy in the field per unit volume. When multiplied by the velocity v it gives the rate at which energy is transferred through a unit area. Since $v = c/n$ and $\epsilon_r = n^2$ the intensity is proportional to $n E_o^2$.

Reflectance R measures the intensity of the reflected light with respect to that of the incident light and can be defined separately for electric field components parallel and perpendicular to the plane of incidence. The reflectances R_\perp and R_{\parallel} are defined by

Reflectances

$$R_\perp = \frac{|E_{ro,\perp}|^2}{|E_{io,\perp}|^2} = |r_\perp|^2 \quad \text{and} \quad R_{\parallel} = \frac{|E_{ro,\parallel}|^2}{|E_{io,\parallel}|^2} = |r_\parallel|^2 \quad (1.6.16)$$

Although the reflection coefficients can be complex numbers that can represent phase changes, reflectances are necessarily real numbers representing intensity changes. Magnitude of a complex number is defined in terms of its product with its complex conjugate. For example, when $E_{ro,\parallel}$ is a complex number then

$$|E_{ro,\parallel}|^2 = (E_{ro,\parallel})(E_{ro,\parallel})^*$$

in which $(E_{ro,\parallel})^*$ is the complex conjugate of $(E_{ro,\parallel})$.

²²Strictly the terms intensity and irradiance are not the same as mentioned in footnote 15.

From Eqs. (1.6.6a) and (1.6.7a) with normal incidence, these are simply given by

$$R = R_{\perp} = R_{\parallel} = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2 \quad (1.6.17)$$

Reflectance
at normal
incidence

Since a glass medium has a refractive index of around 1.5 this means that typically 4% of the incident radiation on an air–glass surface will be reflected back.

Transmittance T relates the intensity of the transmitted wave to that of the incident wave in a similar fashion to the reflectance. We must, however, consider that the transmitted wave is in a different medium and also that its direction with respect to the boundary is different from that of the incident wave by virtue of refraction. For normal incidence, the incident and transmitted beams are normal and the transmittances are defined and given by

$$T_{\perp} = \frac{n_2 |E_{to,\perp}|^2}{n_1 |E_{io,\perp}|^2} = \left(\frac{n_2}{n_1} \right) |t_{\perp}|^2 \quad \text{and} \quad T_{\parallel} = \frac{n_2 |E_{to,\parallel}|^2}{n_1 |E_{io,\parallel}|^2} = \left(\frac{n_2}{n_1} \right) |t_{\parallel}|^2 \quad (1.6.18)$$

or

$$T = T_{\perp} = T_{\parallel} = \frac{4n_1 n_2}{(n_1 + n_2)^2} \quad (1.6.19)$$

Trans-
mittance
at normal
incidence

Further, the fraction of light reflected and fraction transmitted must add to unity. Thus $R + T = 1$.

When the light is incident at an angle θ_i , as in Figure 1.15, then the transmitted light has an angle θ_t with respect to normal. The corresponding transmittances are given by

$$T_{\perp} = \left(\frac{n_2 \cos \theta_t}{n_1 \cos \theta_i} \right) |t_{\perp}|^2 \quad \text{and} \quad T_{\parallel} = \left(\frac{n_2 \cos \theta_t}{n_1 \cos \theta_i} \right) |t_{\parallel}|^2 \quad (1.6.20)$$

Trans-
mittance
at oblique
incidence

Each transmittance in Eq. (1.6.20) is with respect to the incident intensity in the corresponding polarization.

C. Goos-Hänchen Shift and Optical Tunneling

A light traveling in an optically more dense medium suffers total internal reflection when it is incident on a less dense medium at an angle of incidence greater than the critical angle ($\theta_i > \theta_c$) as shown in Figure 1.12 (c). Simple ray trajectory analysis gives the impression that the reflected ray emerges from the point of contact of the incident ray with the interface as in Figure 1.12 (c). However, careful optical experiments examining the incident and reflected beams have shown that the reflected wave appears to be laterally shifted from the point of incidence at the interface as illustrated in Figure 1.20. Although the angles of incidence and reflection are the same (as one expects from Fresnel's equation), the reflected beam, nonetheless, is laterally shifted and appears to be reflected from a *virtual plane inside the optically less dense medium*. The lateral shift is known as the **Goos-Hänchen shift**.

The lateral shift of the reflected beam can be understood by considering that the reflected beam experiences a phase change ϕ , as shown in Figure 1.16 (b), and that the electric field extends into the second medium roughly by a penetration depth $\delta = 1/\alpha_2$. We know that phase changes other than 0 or 180° occur only when there is total internal reflection. We can equivalently represent this phase change ϕ and the penetration into the second medium by shifting the reflected wave along the direction of propagation of the evanescent wave, that is, along z ,

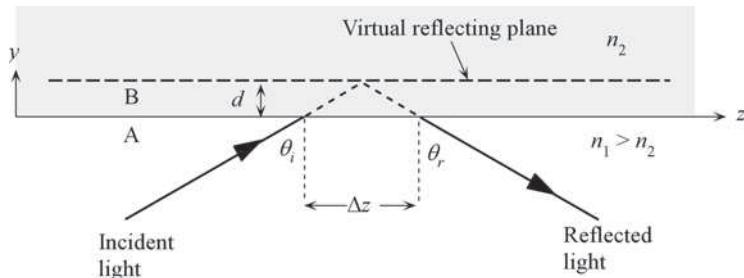


FIGURE 1.20 The reflected light beam in total internal reflection appears to have been laterally shifted by an amount Δz at the interface. It appears as though it is reflected from a virtual plane at a depth d in the second medium from the interface.

by an amount²³ Δz as shown in Figure 1.20. The lateral shift depends on the angle of incidence and the penetration depth. We can represent the reflection as if it were occurring from a virtual plane placed at some distance d from the interface (d is not the same as δ), then from simple geometric considerations, $\Delta z = 2d \tan \theta_i$. Actual d , however, is quite complicated to calculate, but its *order of magnitude* is the penetration depth δ , so that for light with $\lambda = 1 \mu\text{m}$ incident at 85° at a glass–glass ($n_1 = 1.450$ and $n_2 = 1.430$) interface and suffering TIR, $\delta = 0.78 \mu\text{m}$, which means $\Delta z \sim 18 \mu\text{m}$.

Total internal reflection occurs whenever a wave propagating in an optically denser medium, such as in A in Figure 1.20, is incident at an angle greater than the critical angle at the interface AB with a medium B of lower refractive index. If we were to shrink the thickness d of medium B, as in Figure 1.21, we would observe that when B is sufficiently thin, an attenuated light beam emerges on the other side of B in C. This phenomenon in which an incident wave is partially transmitted through a medium where it is forbidden in terms of simple geometrical optics is called **optical tunneling** and is a consequence of the electromagnetic wave nature of light. It is due to the fact that the field of the evanescent wave penetrates into medium B and reaches the interface BC before it vanishes. The optical tunneling phenomenon is illustrated in Figure 1.21 and is referred to as **frustrated total internal reflection** (FTIR): the proximity of medium C

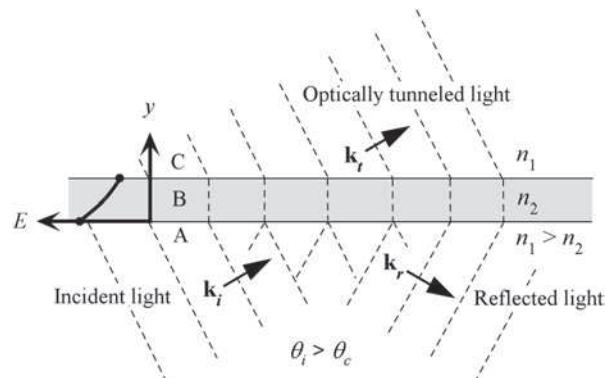


FIGURE 1.21 When medium B is thin, the field penetrates from the AB interface into medium B and reaches BC interface, and gives rise to a transmitted wave in medium C. The effect is the tunneling of the incident beam in A through B to C. The maximum field E_{\max} of the evanescent wave in B decays in B along y but is finite at the BC boundary and excites the transmitted wave.

²³The actual analysis of the Goos-Haenchen shift is more complicated; see for example J. E. Midwinter, *Optical Fibers for Transmission* (John Wiley and Sons, 1979), Ch. 3. d in Figure 1.20 is not simply the penetration depth δ . In fact, d is such that Δz represents the right phase change observed upon TIR.

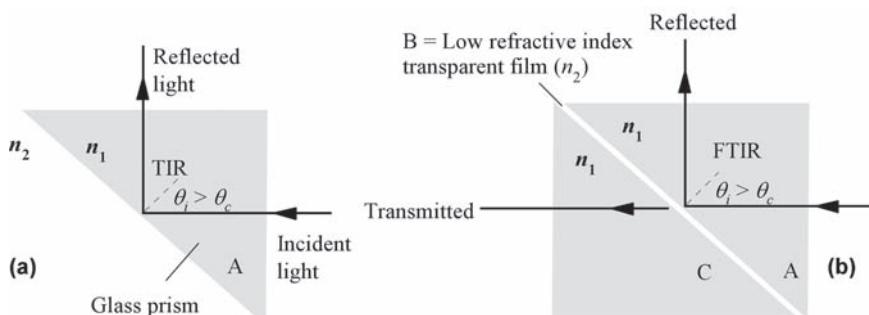
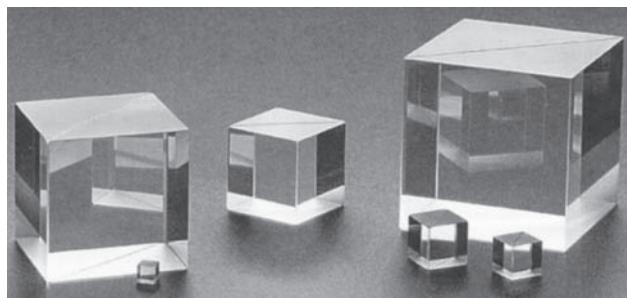


FIGURE 1.22 (a) A light incident at the long face of a glass prism suffers TIR; the prism deflects the light. (b) Two prisms separated by a thin low refractive index film forming a beam-splitter cube. The incident beam is split into two beams by FTIR.



Beam splitter cubes. (Courtesy of CVI Melles Griot.)

frustrates TIR. The transmitted wave in C carries some of the light intensity and thus the intensity of the reflected wave is reduced. Notice that the field of the evanescent wave in B decays and has a finite value at the BC interface, where the field oscillations are able to excite a transmitted wave.

Frustrated total internal reflection is utilized in beam splitters as shown in Figure 1.22. A light beam entering the glass prism A suffers TIR at the hypotenuse face ($\theta_i > \theta_c$ at the glass-air interface) and becomes reflected; the prism deflects the light as in Figure 1.22 (a). In the *beam splitter cube* in Figure 1.22 (b), two prisms, A and C, are separated by a thin film, B, of low refractive index. Some of the light energy is now tunneled through this thin layer and transmitted into C and out from the cube. FTIR at the hypotenuse face of A leads to a transmitted beam and hence to the splitting of the incident beam into two beams. The extent of energy division between the two beams depends on the thickness of the thin layer B and its refractive index.

EXAMPLE 1.6.1 Reflection of light from a less dense medium (internal reflection)

A ray of light that is traveling in a glass medium of refractive index $n_1 = 1.450$ becomes incident on a less dense glass medium of refractive index $n_2 = 1.430$. Suppose that the free-space wavelength (λ) of the light ray is $1 \mu\text{m}$.

- What should the minimum incidence angle for TIR be?
- What is the phase change in the reflected wave when $\theta_i = 85^\circ$ and when $\theta_i = 90^\circ$?
- What is the penetration depth of the evanescent wave into medium 2 when $\theta_i = 85^\circ$ and when $\theta_i = 90^\circ$?

Solution

- (a) The critical angle θ_c for TIR is given by $\sin \theta_c = n_2/n_1 = 1.430/1.450$ so that $\theta_c = 80.47^\circ$.
 (b) Since the incidence angle $\theta_i > \theta_c$, there is a phase shift in the reflected wave. The phase change in $E_{r,\perp}$ is given by ϕ_\perp . With $n_1 = 1.450$, $n_2 = 1.430$, and $\theta_i = 85^\circ$

$$\begin{aligned}\tan\left(\frac{1}{2}\phi_\perp\right) &= \frac{[\sin^2\theta_i - n^2]^{1/2}}{\cos\theta_i} = \frac{\left[\sin^2(85^\circ) - \left(\frac{1.430}{1.450}\right)^2\right]^{1/2}}{\cos(85^\circ)} \\ &= 1.61449 = \tan\left[\frac{1}{2}(116.45^\circ)\right]\end{aligned}$$

so that the phase change is 116.45° . For the $E_{r,\parallel}$ component, the phase change is

$$\tan\left(\frac{1}{2}\phi_\parallel + \frac{1}{2}\pi\right) = \frac{[\sin^2\theta_i - n^2]^{1/2}}{n^2 \cos\theta_i} = \frac{1}{n^2} \tan\left(\frac{1}{2}\phi_\perp\right)$$

so that

$$\tan\left(\frac{1}{2}\phi_\parallel + \frac{1}{2}\pi\right) = (n_1/n_2)^2 \tan(\phi_\perp/2) = (1.450/1.430)^2 \tan\left(\frac{1}{2}116.45^\circ\right)$$

which gives $\phi_\parallel = -62.1^\circ$. (Note: If we were to invert the reflected field, this phase change would be $+117.86^\circ$)

We can repeat the calculation with $\theta_i = 90^\circ$ to find $\phi_\perp = 180^\circ$ and $\phi_\parallel = 0^\circ$.

Note that as long as $\theta_i > \theta_c$, the magnitude of the reflection coefficients are unity. Only the phase changes.

- (c) The amplitude of the evanescent wave as it penetrates into medium 2 is

$$E_{t,\perp}(y) = E_{t_0,\perp} \exp(-\alpha_2 y)$$

We ignore the z -dependence, $\exp(j\omega t - k_z z)$, as it only gives a propagating property along z . The field strength drops to e^{-1} when $y = 1/\alpha_2 = \delta$, which is the *penetration depth*. The attenuation constant α_2 is

$$\alpha_2 = \frac{2\pi n_2}{\lambda_o} \left[\left(\frac{n_1}{n_2} \right)^2 \sin^2 \theta_i - 1 \right]^{1/2}$$

that is,

$$\alpha_2 = \frac{2\pi(1.430)}{(1.0 \times 10^{-6} \text{ m})} \left[\left(\frac{1.450}{1.430} \right)^2 \sin^2(85^\circ) - 1 \right]^{1/2} = 1.28 \times 10^6 \text{ m}^{-1}$$

so the penetration depth is $\delta = 1/\alpha_2 = 1/(1.28 \times 10^6 \text{ m}) = 7.8 \times 10^{-7} \text{ m}$, or $0.78 \mu\text{m}$. For 90° , repeating the calculation we find $\alpha_2 = 1.5 \times 10^6 \text{ m}^{-1}$, so that $\delta = 1/\alpha_2 = 0.66 \mu\text{m}$. We see that the penetration is greater for smaller incidence angles. This will be an important consideration later in analyzing light propagation in optical fibers.

EXAMPLE 1.6.2 Reflection at normal incidence, and internal and external reflection

Consider the reflection of light at normal incidence on a boundary between a glass medium of refractive index 1.5 and air of refractive index 1.

- (a) If light is traveling from air to glass, what is the reflection coefficient and the intensity of the reflected light?

- (b) If light is traveling from glass to air, what is the reflection coefficient and the intensity of the reflected light?

Solution

- (a) The light travels in air and becomes partially reflected at the surface of the glass that corresponds to external reflection. Thus $n_1 = 1$ and $n_2 = 1.5$. Then

$$r_{\parallel} = r_{\perp} = \frac{n_1 - n_2}{n_1 + n_2} = \frac{1 - 1.5}{1 + 1.5} = -0.2$$

This is negative, which means that there is a 180° phase shift. The reflectance (R), which gives the fractional reflected power, is

$$R = r_{\parallel}^2 = 0.04 \quad \text{or} \quad 4\%.$$

- (b) The light travels in glass and becomes partially reflected at the glass-air interface that corresponds to internal reflection. Thus $n_1 = 1.5$ and $n_2 = 1$. Then

$$r_{\parallel} = r_{\perp} = \frac{n_1 - n_2}{n_1 + n_2} = \frac{1.5 - 1}{1.5 + 1} = 0.2$$

There is no phase shift. The reflectance is again 0.04 or 4%. In both cases (a) and (b), the amount of reflected light is the same.

EXAMPLE 1.6.3 Reflection and transmission at the Brewster angle

A light beam traveling in air is incident on a glass plate of refractive index 1.50. What is the Brewster or polarization angle? What are the relative intensities of the reflected and transmitted light for the polarization perpendicular and parallel to the plane of incidence at the Brewster angle of incidence?

Solution

Light is traveling in air and is incident on the glass surface at the polarization angle θ_p . Here $n_1 = 1$, $n_2 = 1.5$, $n = n_2/n_1 = 1.5$, and $\tan \theta_p = (n_2/n_1) = 1.5$ so that $\theta_p = 56.31^\circ$. We now use Fresnel's equations to find the reflected and transmitted amplitudes. For the perpendicular polarization, from Eq. (1.6.6a),

$$r_{\perp} = \frac{\cos(56.31^\circ) - [1.5^2 - \sin^2(56.31^\circ)]^{1/2}}{\cos(56.31^\circ) + [1.5^2 - \sin^2(56.31^\circ)]^{1/2}} = -0.385$$

On the other hand, $r_{\parallel} = 0$. The reflectances $R_{\perp} = |r_{\perp}|^2 = 0.148$ and $R_{\parallel} = |r_{\parallel}|^2 = 0$ so that the reflected light has no parallel polarization in the plane of incidence. Notice the negative sign in r_{\perp} , which indicates a phase change of π .

From Eqs. (1.6.6b) and (1.7.7b), the transmission coefficients are

$$t_{\perp} = \frac{2\cos(56.31^\circ)}{\cos(56.31^\circ) + [1.5^2 - \sin^2(56.31^\circ)]^{1/2}} = 0.615$$

and

$$t_{\parallel} = \frac{2(1.5)\cos(56.31^\circ)}{(1.5)^2\cos(56.31^\circ) + [1.5^2 - \sin^2(56.31^\circ)]^{1/2}} = 0.667$$

Notice that $r_{\parallel} + nt_{\parallel} = 1$ and $r_{\perp} + t_{\perp} = 1$, as we expect. To find the transmittance for each polarization, we need the refraction angle θ_r . From Snell's law, $n_1 \sin \theta_i = n_2 \sin \theta_r$, that is $(1) \sin(56.31^\circ) = (1.5) \sin \theta_r$, we find $\theta_r = 33.69^\circ$. Then, from Eq. (1.6.20),

$$T_{\perp} = \left[\frac{(1.5) \cos(33.69^\circ)}{(1) \cos(56.31^\circ)} \right] (0.615)^2 = 0.852 \quad \text{and} \quad T_{\parallel} = \left[\frac{(1.5) \cos(33.69^\circ)}{(1) \cos(56.31^\circ)} \right] (0.667)^2 = 1$$

Clearly, light with polarization parallel to the plane of incidence has greater intensity. Note that $R + T = 1$ for both polarizations.

If we were to reflect light from a glass plate, keeping the angle of incidence at 56.3° , then the reflected light will be polarized with an electric field component perpendicular to the plane of incidence. The transmitted light will have the field greater in the plane of incidence; that is, it will be partially polarized. By using a stack of glass plates one can increase the polarization of the transmitted light. (This type of *pile-of-plates polarizer* was invented by Dominique F. J. Arago in 1812.)

1.7 ANTIREFLECTION COATINGS AND DIELECTRIC MIRRORS

Fresnel equations are routinely used in a number of applications in optoelectronics to design and fabricate optical coatings, that is, thin films, to reduce reflections and glare, and also in various components such as dielectric mirrors and filters. Section 1.14 on thin film optics provides a good example of their application for thin film coatings; but in this section we consider two practical applications in optoelectronics: antireflection (AR) coatings and dielectric mirrors.

A. Antireflection Coatings on Photodetectors and Solar Cells

When light is incident on the surface of a semiconductor, it becomes partially reflected. Partial reflection is an important consideration in solar cells where transmitted light energy into the semiconductor device is converted to electrical energy. The refractive index of Si is about 3.5 at wavelengths around 600–800 nm. Thus, the reflectance with $n_1(\text{air}) = 1$ and $n_2(\text{Si}) \approx 3.5$ is

$$R = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2 = \left(\frac{1 - 3.5}{1 + 3.5} \right)^2 = 0.309$$

This means that 30% of the light is reflected and is not available for conversion to electrical energy; a considerable reduction in the efficiency of the solar cell.

However, we can coat the surface of the semiconductor device with a thin layer of a dielectric material, such as an $a\text{-Si}_{1-x}\text{N}_x\text{:H}$ (amorphous hydrogenated silicon nitride based on silicon nitride, Si_3N_4 , and x is typically 0.4–0.6), that has an intermediate refractive index. Figure 1.23

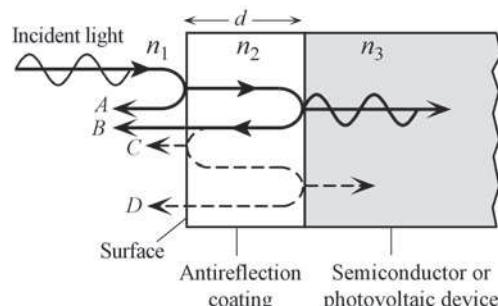


FIGURE 1.23 Illustration of how an antireflection (AR) coating reduces the reflected light intensity. The thickness d and the refractive index n_2 of the antireflection coefficient are such that the waves A and B have a phase difference of π and hence interfere destructively. There is no reflection.

illustrates how the thin dielectric coating, called an **antireflection (AR) coating**, reduces the reflected light intensity. In this case $n_1(\text{air}) = 1$, $n_2(\text{coating}) \approx 1.9$, and $n_3(\text{Si}) = 3.5$. Light is first incident on the air/coating surface and some of it becomes reflected and this reflected wave is shown as *A* in Figure 1.23. Wave *A* has experienced a 180° phase change on reflection as this is an external reflection. The wave that enters and travels in the coating then becomes reflected at the coating/semiconductor surface. This wave, which is shown as *B* in Figure 1.23, also suffers a 180° phase change since $n_3 > n_2$. When wave *B* reaches *A*, it has suffered a total delay corresponding to traversing the thickness d of the coating twice. The phase difference is equivalent to $k_c(2d)$ in which $k_c = 2\pi/\lambda_c$ is the propagation constant in the coating and is given by $2\pi/\lambda_c$, where λ_c is the wavelength in the coating. Since $\lambda_c = \lambda/n_2$, where λ is the free-space wavelength, the phase difference $\Delta\phi$ between *A* and *B* is $(2\pi n_2/\lambda)(2d)$. To reduce the reflected light, *A* and *B* must interfere destructively and this requires the phase difference to be π or odd-multiples of π , $m\pi$ in which $m = 1, 3, 5, \dots$ is an odd-integer. Thus

$$\left(\frac{2\pi n_2}{\lambda}\right)2d = m\pi \quad \text{or} \quad d = m\left(\frac{\lambda}{4n_2}\right) \quad (1.7.1) \quad \begin{matrix} \text{Antireflec-} \\ \text{tion coating} \end{matrix}$$

Thus, the thickness of the coating must be odd multiples of the quarter wavelength in the coating and depends on the wavelength.

To obtain a good degree of destructive interference between waves *A* and *B*, the two amplitudes must be comparable. (In fact, we have to consider multiple reflections as in Section 1.14, a subject of thin film optics) It turns out that we need $n_2 = (n_1 n_3)^{1/2}$. When $n_2 = (n_1 n_3)^{1/2}$, then the reflection coefficient between the air and coating is equal to that between the coating and the semiconductor. In this case we would need $(3.5)^{1/2}$ or 1.87. Thus, a-Si_{1-x}N_x:H is a good choice as an AR coating material on Si solar cells. Its refractive index can also be modified in thin film deposition by changing its composition x . Taking the wavelength to be 600 nm, $d = (600 \text{ nm})/[4(1.9)] = 79.0 \text{ nm}$ or odd-multiples of d .

Once Eq. (1.7.1) is satisfied for an AR coating, it is not difficult to calculate the minimum reflectance obtainable from such a coating. However, we need to consider all reflected waves arising from multiple reflections in the thin film, such as *C* and *D* and so on, and we need to properly account for the magnitude of the field by multiplying the field with the right reflection coefficient at each reflection, and with the right transmission coefficient at each transmission. We add all the reflected fields, $A + B + C + D + \dots$, and then square the total reflected field for the intensity. Once we have done all that, the minimum reflectance R_{\min} is given by

$$R_{\min} = \left(\frac{n_2^2 - n_1 n_3}{n_2^2 + n_1 n_3}\right)^2 \quad (1.7.2) \quad \begin{matrix} \text{Minimum} \\ \text{and} \\ \text{maximum} \\ \text{reflectance} \end{matrix}$$

Clearly, the best choice is for $n_2 = (n_1 n_3)^{1/2}$; recall that R_{\min} is only minimum at one wavelength that satisfies Eq. (1.7.1).

EXAMPLE 1.7.1 Antireflection coating on a photodetector

Consider an InGaAs photodetector for use at a wavelength of 1310 nm in optical communications. The refractive index of doped InGaAs that is used in the detector at 1310 nm is 3.61. What is the reflectance from its surface without an AR coating? What would be the ideal refractive index for an AR coating on InGaAs? What should be thickness of an AR coating that has a refractive index of 1.85. What is its reflectance?

Solution

We use $n_1 = 1$ for air, n_2 for the antireflection coating and $n_3 = 3.61$ for InGaAs,

Without an AR coating, the reflectance is

$$R = [(n_1 - n_3)/(n_1 + n_3)]^2 = [(1 - 3.61)/(1 + 3.61)]^2 = 0.32 \text{ or } 32\%$$

The ideal AR coating would satisfy Eq. (1.7.1) and would have $n_2 = (n_1 n_3)^{1/2} = 1.90$.

From Eq. (1.7.1), the thickness of the AR layer should be

$$d = \lambda/(4n_2) = (1310 \text{ nm})/[4(1.85)] = 177 \text{ nm}$$

From Eq. (1.7.2), the minimum reflectance with the AR coating is

$$R_{\min} = \left(\frac{1.85^2 - (1)(3.61)}{1.85^2 + (1)(3.61)} \right)^2 = 7.1 \times 10^{-4}, \text{ or } 0.07\%, \text{ very small}$$

B. Dielectric Mirrors and Bragg Reflectors

A **dielectric mirror** consists of a stack of dielectric layers of alternating refractive indices as schematically illustrated in Figure 1.24 (a), in which n_1 is greater than n_2 . It is customary to write $n_1 = n_H$ (high) and $n_2 = n_L$ (low) to represent the $n_H n_L$ stack shown in the figure. The thickness of each layer is a quarter of wavelength or $\lambda_{\text{layer}}/4$ in which λ_{layer} is the wavelength of light in that layer, or λ_o/n where λ_o is the free-space wavelength at which the mirror is required to reflect the incident light and n is the refractive index of the layer. Reflected waves from the interfaces interfere constructively and give rise to a substantial reflected light over a band of wavelengths centered around λ_o as shown in Figure 1.24 (b). If there are sufficient numbers of layers, the reflectance can approach unity at the wavelength λ_o . Since n_1 (high) and n_2 (low) layers are used in pairs, the total number of such pairs of layers, or double layers, is denoted as N ; as N increases, the reflectance also increases. The layers are coated, by vacuum deposition techniques, on a suitable substrate. The dielectric mirror illustrated in Figure 1.24 (a) is also known as a **quarter-wave dielectric stack**. Figure 1.24 (b) shows the typical reflectance vs. wavelength behavior of three dielectric mirrors with different n_1 to n_2 ratios and N values. The mirror has been designed to reflect at $1.55 \mu\text{m}$.

The reflection coefficient r_{12} for light in layer 1 being reflected at the 1-2 boundary, from Eq. (1.6.9), is $r_{12} = (n_1 - n_2)/(n_1 + n_2)$ and is a *positive* number indicating no phase change. The reflection coefficient for light in layer 2 being reflected at the 2-1 boundary is $r_{21} = (n_2 - n_1)/(n_2 + n_1)$, which is $-r_{12}$ or *negative*, indicating a π phase change. Thus the reflection coefficient alternates in sign through the mirror. Consider two arbitrary waves, B and C , which are reflected at two consecutive interfaces. The two waves are therefore already out of phase by π due to reflections at the different boundaries. Further, wave B travels an additional distance that is twice $(\lambda_2/4)$ (the thickness of layer d_2) before reaching wave B and therefore experiences a phase change equivalent to $2(\lambda_2/4)$ or $\lambda_2/2$, that is π . The phase difference between B and C is then $\pi + \pi$ or 2π . Thus, waves B and C are in phase and *interfere constructively*. We can similarly show that waves C and D also interfere constructively and so on, so that all reflected waves from the consecutive boundaries interfere constructively. After several layers (depending on the n_1/n_2 ratio) the transmitted intensity will be very small and the reflected light intensity will be close to unity as indicated in Figure 1.24 (b). These dielectric

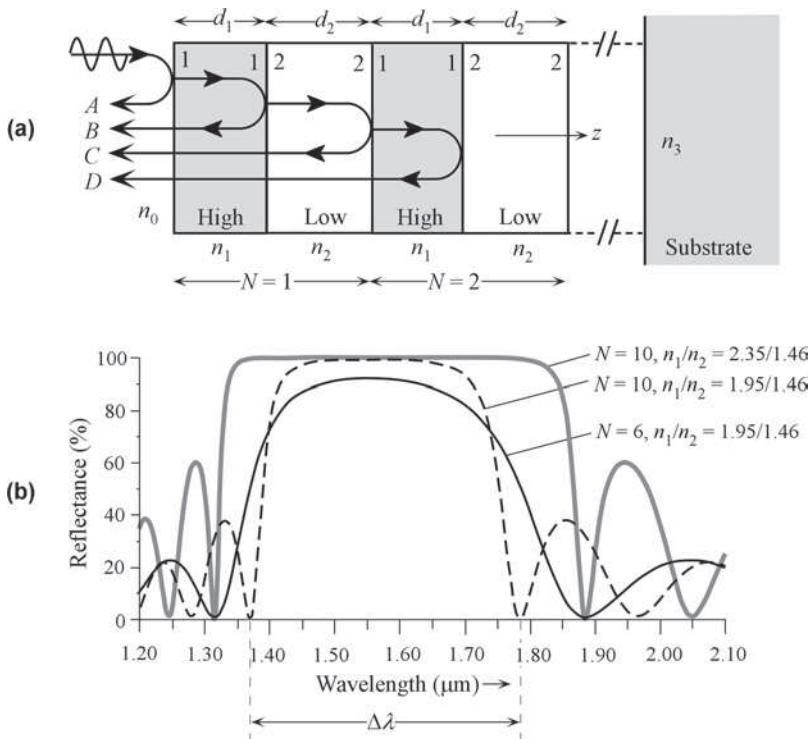


FIGURE 1.24 (a) Schematic illustration of the principle of the dielectric mirror with *many* high and low refractive index layers. Reflected waves A, B, C, D , and so on all interfere constructively if the layer thicknesses d_1 and d_2 are a quarter of a wavelength within the layer, that is $d_1 = \lambda/4n_1$ and $d_2 = \lambda/4n_2$, where λ is the free-space wavelength. The dielectric mirror is assumed to be coated on a substrate with an index n_3 . (b) The reflectance of three different dielectric mirrors that have $N = 10$, $n_1/n_2 = 2.35/1.46$; $N = 10$, $n_1/n_2 = 1.95/1.46$; $N = 6$, $n_1/n_2 = 1.95/1.46$, $n_3 = 1.52$. (Note: $n(\text{TiO}_2) = 2.35$, $n(\text{Si}_3\text{N}_4) = 1.95$, $n(\text{SiO}_2) = 1.46$.)

mirrors are widely used in photonics, for example, in solid state lasers such as the vertical cavity surface emitting laser diode. Since the dielectric mirror has a periodic variation in the refractive index (the period being $d_1 + d_2$), similar to a diffraction grating, it is sometimes referred to as a **Bragg reflector**.²⁴ It is left as an exercise to show that if we interchange the high and low layers, $n_1 = n_L$ and $n_2 = n_H$, we obtain the same result.

As shown in Figure 1.24 (b), with sufficient number of double layers, the reflectance is almost unity over a band of wavelengths, $\Delta\lambda$. Conversely, the transmittance vanishes over the same wavelength range $\Delta\lambda$. This wavelength range in which the transmittance vanishes is called **reflectance bandwidth**; or the **stop band** for the transmitted light. As discussed in Section 1.17, the dielectric mirror in Figure 1.24 (a) is a one-dimensional photonic crystal in which there is a certain stop band within which there can be no propagation of waves along the z axis within the multilayer dielectric structure in Figure 1.24 (a).

²⁴ As we will see later, a periodic variation in the refractive index is actually a diffraction grating and is able to diffract or reflect a wave in a certain direction if the periodicity is right. The dielectric mirror is also called a one-dimensional Bragg grating structure. It is also a one-dimensional photonic crystal for large N .



Various dielectric mirrors, which are quarter wave dielectric stacks on Pyrex or Zerodur substrates. (Courtesy of Newport.)

There are two general observations from the reflectance spectra in Figure 1.24 (b). The reflectance R increases with N , the number of double layers used. R also increases with the refractive index ratio n_1/n_2 ; or put differently, with the index contrast. For a large number of layers, the bandwidth $\Delta\lambda$ of the dielectric mirror increases with the index contrast. The *maximum reflectance* R_N for N pairs of layers is given by

Maximum reflectance, dielectric mirror

$$R_N = \left[\frac{n_1^{2N} - (n_0/n_3)n_2^{2N}}{n_1^{2N} + (n_0/n_3)n_2^{2N}} \right]^2 \quad (1.7.3)$$

The bandwidth $\Delta\lambda$ when $2N$ is large (for near-unity reflectance) is given by

Reflectance bandwidth

$$\frac{\Delta\lambda}{\lambda_o} \approx (4/\pi) \arcsin \left(\frac{n_1 - n_2}{n_1 + n_2} \right) \quad (1.7.4)$$

which increases with the refractive index contrast as apparent from Figure 1.24 (b). If we interchange the low-high layers so that the high index layer is on the air side, there is very little change in the reflectance or the bandwidth in a highly reflecting dielectric mirror.

EXAMPLE 1.7.2 Dielectric mirror

Consider a dielectric mirror that has quarter wave layers consisting of Ta_2O_5 with $n_H = 1.78$ and SiO_2 with $n_L = 1.55$ both at 850 nm, the central wavelength at which the mirror reflects light. Suppose the substrate is Pyrex glass with an index $n_s = 1.47$ and the outside medium is air with $n_0 = 1$. Calculate the maximum reflectance of the mirror when the number N of double layers is 4 and 12. What would happen if you use TiO_2 with $n_H = 2.49$, instead of Ta_2O_5 ? What is the bandwidth for these two dielectric stacks when they are highly reflecting (with many pairs of layers)? Suppose we use a Si wafer as the substrate, what happens to the maximum reflectance?

Solution

We use $n_0 = 1$ for air, $n_1 = n_H = 1.78$, $n_2 = n_L = 1.55$, $n_3 = n_s = 1.47$, $N = 4$ in Eq. (1.7.3). Thus, for four pairs of layers, the maximum reflectance R_4 is

$$R_4 = \left[\frac{(1.78)^{2(4)} - (1/1.47)(1.55)^{2(4)}}{(1.78)^{2(4)} + (1/1.47)(1.55)^{2(4)}} \right]^2 = 0.40 \quad \text{or} \quad 40\%$$

If we repeat the calculation for 12 pairs of layers, we will find $R_{12} = 90.6\%$.

If we use TiO_2 with $n_1 = n_H = 2.49$, we would find $R_4 = 94.0\%$ and $R_{12} = 100\%$ (to two decimal places). Obviously the refractive index contrast is important; with the $\text{TiO}_2\text{-SiO}_2$ stack we only need four double layers to get roughly the same reflectance as from 12 pairs of layers of $\text{Ta}_2\text{O}_5\text{-SiO}_2$. If we interchange n_H and n_L in the 12-pair stack, that is, $n_1 = n_L$ and $n_2 = n_H$, the $\text{Ta}_2\text{O}_5\text{-SiO}_2$ reflectance falls to 80.8% but the $\text{TiO}_2\text{-SiO}_2$ stack is unaffected since it is already reflecting nearly all the light.

We can only compare bandwidths $\Delta\lambda$ for “infinite” stacks (those with $R \approx 100\%$). For the $\text{TiO}_2\text{-SiO}_2$ stack, Eq. (1.7.4) gives

$$\Delta\lambda \approx \lambda_o(4/\pi) \arcsin\left(\frac{n_2 - n_1}{n_2 + n_1}\right) = (850 \text{ nm})(4/\pi) \arcsin\left(\frac{2.49 - 1.55}{2.49 + 1.55}\right) = 254 \text{ nm}$$

On the other hand, for the $\text{Ta}_2\text{O}_5\text{-SiO}_2$ infinite stack, we get $\Delta\lambda = 74.8 \text{ nm}$. As expected, $\Delta\lambda$ is narrower for the smaller contrast stack.

If we change the substrate to a silicon wafer with $n_3 = n_s = 3.50$, we would find that the $\text{Ta}_2\text{O}_5\text{-SiO}_2$ 4-pair stack gives a reflectance of 68.5%, higher than before because the large index changes from n_L to n_s at the substrate interface provides further reflections.

1.8 ABSORPTION OF LIGHT AND COMPLEX REFRACTIVE INDEX

Generally when light propagates through a material it becomes *attenuated* in the direction of propagation as illustrated in Figure 1.25. We distinguish between *absorption* and *scattering* both of which gives rise to a loss of intensity in the regular direction of propagation. In **absorption**, the loss in the power in the propagating electromagnetic wave is due to the conversion of light energy to other forms of energy; for example, lattice vibrations (heat) during the polarization of the molecules of the medium or during the local vibrations of impurity ions driven by the optical field. The excitation of electrons from the valence band to the conduction band, or from impurities to the conduction band, in insulators and semiconductors would also absorb energy from the propagating radiation. Further, free electrons inside a medium with a finite conductivity can be drifted by the optical field in the radiation. As these electrons become scattered by lattice vibrations (or impurities) they will pass the energy they have acquired from the EM wave to lattice vibrations. There are other examples as well. In all cases, some of energy from the propagating light wave is absorbed and converted to other forms of energy.

On the other hand, **scattering** is a process by which the energy from a propagating EM wave is redirected as secondary EM waves in various directions away from the original direction of propagation. (This is discussed later in this chapter) The **attenuation coefficient** α is defined as the fractional decrease in the irradiance I of a wave per unit distance along the direction of propagation z

$$\alpha = -\frac{dI}{Idz} \quad (1.8.1)$$

*Definition
of attenu-
ation
coefficient*

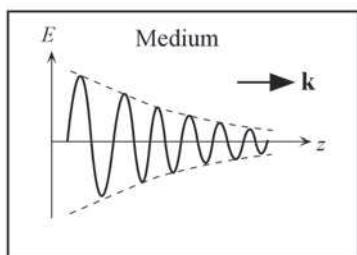


FIGURE 1.25 Attenuation of a traveling wave in a medium results in the decay of its amplitude.

When the irradiance decreases, dI/dz is negative, and the attenuation coefficient is a positive number. If the attenuation of the wave is due to absorption only, then α is the **absorption coefficient**.

It is instructive to consider what happens when a monochromatic light wave such as

Lossless propagation

$$E = E_o \exp j(\omega t - kz) \quad (1.8.2)$$

is propagating in a dielectric medium. The electric field E in Eq. (1.8.2) is either parallel to x or y since propagation is along z . As the wave travels through the medium, the molecules become polarized. This polarization effect is represented by the *relative permittivity* ϵ_r of the medium. If there were no losses in the polarization process, then the relative permittivity ϵ_r would be a real number and the corresponding refractive index $n = \epsilon_r^{1/2}$ would also be a real number. However, we know that there are always some losses in all polarization processes. For example, when the ions of an ionic crystal are displaced from their equilibrium positions by an alternating electric field and made to oscillate, some of the energy from the electric field is coupled and converted to lattice vibrations (called phonons). These losses are generally accounted by describing the whole medium in terms of a **complex relative permittivity (or dielectric constant)** ϵ_r , that is,

Complex dielectric constant

$$\epsilon_r = \epsilon'_r - j\epsilon''_r \quad (1.8.3)$$

where the real part ϵ'_r determines the polarization of the medium with losses ignored and the imaginary part ϵ''_r describes the losses in the medium.²⁵ For a lossless medium, obviously $\epsilon_r = \epsilon'_r$. The loss ϵ''_r depends on the frequency of the wave and usually peaks at certain natural (resonant) frequencies involved in the absorption process.

An EM wave that is traveling in a medium and experiencing attenuation due to absorption can be generally described by a **complex propagation constant** k , that is,

Complex propagation constant

$$k = k' - jk'' \quad (1.8.4)$$

where k' and k'' are the real and imaginary parts. If we put Eq. (1.8.4) into Eq. (1.8.2) we will find the following:

Attenuated propagation

$$E = E_o \exp(-k''z) \exp j(\omega t - k'z) \quad (1.8.5)$$

The amplitude decays exponentially while the wave propagates along z . The **real** k' part of the complex propagation constant (wave vector) describes the propagation characteristics (e.g., phase velocity $v = \omega/k'$). The **imaginary** k'' part describes the rate of attenuation along z . The irradiance I at any point along z is

$$I \propto |E|^2 \propto \exp(-2k''z)$$

so that the rate of change in the irradiance with distance is

Imaginary part k''

$$dI/dz = -2k''I \quad (1.8.6)$$

where the negative sign represents attenuation. Clearly $\alpha = 2k''$

²⁵See, for example, S. O. Kasap, *Principles of Electronic Materials and Devices*, 3rd Edition (McGraw-Hill, 2006), Ch. 7. Further, some books use $\epsilon_r = \epsilon'_r + j\epsilon''_r$ instead of Eq. (1.8.3), with a positive imaginary part, but the latter normally refers to an applied field that has a time dependence of the form $\exp(-j\omega t)$.

Suppose that k_o is the propagation constant in vacuum. This is a real quantity as a plane wave suffers no loss in free space. The **complex refractive index** N with a real part n and imaginary part K is defined as the ratio of the complex propagation constant in a medium to propagation constant in free space.

$$N = n - jK = k/k_o = (1/k_o)[k' - jk''] \quad (1.8.7)$$

 Complex
refractive
index

that is,

$$n = k'/k_o \quad \text{and} \quad K = k''/k_o$$

The real part n is simply and generally called the **refractive index** and K is called the **extinction coefficient**. In the absence of attenuation

$$k'' = 0, k = k' \quad \text{and} \quad N = n = k/k_o = k'/k_o$$

We know that in the absence of loss, the relationship between the refractive index n and the relative permittivity ϵ_r is $n = \epsilon_r^{1/2}$. This relationship is also valid in the presence of loss except that we must use a complex refractive index and complex relative permittivity, that is

$$N = n - jK = \sqrt{\epsilon_r} = \sqrt{\epsilon'_r - j\epsilon''_r} \quad (1.8.8)$$

 Complex
refractive
index

By squaring both sides we can relate n and K directly to ϵ'_r and ϵ''_r . The final result is

$$n^2 - K^2 = \epsilon'_r \quad \text{and} \quad 2nK = \epsilon''_r \quad (1.8.9)$$

 Complex
refractive
index

Optical properties of materials are typically reported either by showing the frequency dependences of n and K or ϵ'_r and ϵ''_r . Clearly we can use Eq. (1.8.9) to obtain one set of properties from the other.

Figure 1.26 shows the real (n) and imaginary (K) parts of the complex refractive index of CdTe as a function of wavelength in the infrared region to highlight their behavior around a resonance absorption phenomenon, when the energy transfer is maximum from the EM wave to the material. Both ionic and electronic polarizations contribute to n (≈ 3.3) at low frequencies whereas only electronic polarization contributes to n (≈ 2.6) at high frequencies. The extinction coefficient peaks at about 72 μm when the EM wave oscillations are efficiently coupled to the *lattice vibrations*,

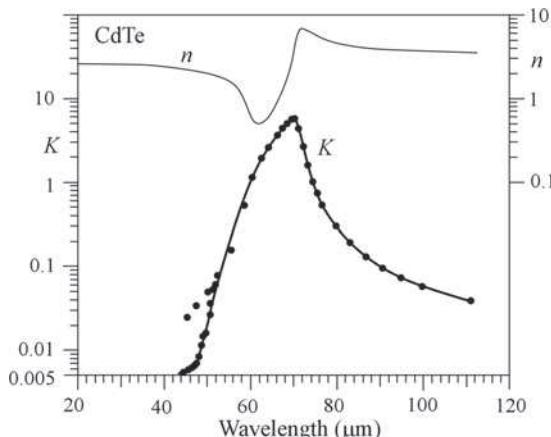


FIGURE 1.26 Optical properties of CdTe as a function of wavelength in the infrared region.

that is, the vibrations of chains of Cd^{2+} and Te^{2-} ions in the crystal,²⁶ so that energy is passed from the EM wave to these lattice vibrations. This type of absorption in which energy is passed to lattice vibrations is called **lattice or Reststrahlen absorption**. Notice that K has a clear peak whereas n shows a response that has a maximum, a minimum, and an inflection point (an S-like shape).

If we know the frequency dependence of the real part ε'_r of the relative permittivity of a material, we can also determine the frequency dependence of the imaginary part ε''_r , and vice versa. This may seem remarkable but it is true provided that we know the frequency dependence of either the real or imaginary part over as wide a range of frequencies as possible (ideally from dc to infinity) and the material is *linear*, that is, it has a relative permittivity that is independent of the applied field; the polarization response must be linearly proportional to the applied field.²⁷ The relationships that relate the real and imaginary parts of the relative permittivity are called **Kramers-Kronig relations**, which involve integral transformations. If $\varepsilon'_r(\omega)$ and $\varepsilon''_r(\omega)$ represent the frequency dependences of the real and imaginary parts, then one can be determined from the other. Similarly, if we know the wavelength dependence of n (or K), over as a wide wavelength range as possible, we can determine the wavelength dependence of K (or n) using Kramers-Kronig relations for $n-K$.

It is instructive to mention that the reflection and transmission coefficients that we derived above were based in using a real refractive index, that is neglecting losses. We can still use the reflection and transmission coefficients if we simply use the complex refractive index N instead of n . For example, consider a light wave traveling in free space incident on a material at normal incidence ($\theta_i = 90^\circ$). The reflection coefficient is now

Reflection coefficient

$$r = \frac{1 - n + jK}{1 + n - jK} \quad (1.8.10)$$

The reflectance is then

Reflectance

$$R = \left| \frac{n - jK - 1}{n - jK + 1} \right|^2 = \frac{(n - 1)^2 + K^2}{(n + 1)^2 + K^2} \quad (1.8.11)$$

which reduce to the usual forms when the extinction coefficient $K = 0$.

The optical properties n and K can be determined by measuring the reflectance from the surface of a material as a function of polarization and the angle of incidence (based on Fresnel's equations).

EXAMPLE 1.8.1 Complex refractive index of InP

An InP crystal has a refractive index (real part) n of 3.549 and an extinction coefficient K of 0.302 at a wavelength of 620 nm (photon energy of 2 eV). Calculate the absorption coefficient α of InP at this wavelength and the reflectance of the air-InP crystal surface.

Solution

The absorption coefficient is

$$\alpha = 2k'' = 2k_oK = 2[(2\pi)/(620 \times 10^{-9} \text{ m})](0.302) = 6.1 \times 10^6 \text{ m}^{-1}$$

²⁶In physics, these would be *optical phonons*. The EM wave would be interacting with optical phonons, and passing its energy onto these phonons.

²⁷In addition, the material system should be passive—contain no sources of energy.

The reflectance R is given by

$$R = \frac{(n - 1)^2 + K^2}{(n + 1)^2 + K^2} = \frac{(3.549 - 1)^2 + (0.302)^2}{(3.549 + 1)^2 + (0.302)^2} = 0.317 \text{ or } 31.7\%$$

EXAMPLE 1.8.2 Reflectance of CdTe around resonance absorption

CdTe is also an optical material used in various applications such as lenses, wedges, prisms, beam splitters, antireflection coatings, and optical windows operating typically in the infrared region up to 25 μm . It is used as an optical material for low power CO₂ laser applications. Figure 1.26 shows the infrared refractive index n and the extinction coefficient K of CdTe. Calculate the absorption coefficient α and the reflectance R of CdTe at the Reststrahlen peak, and also at 50 μm and at 100 μm . What is your conclusion?

Solution

At the Reststrahlen peak, $\lambda \approx 70 \mu\text{m}$, $K \approx 6$, and $n \approx 4$, so that the corresponding free-space propagation constant is

$$k_o = 2\pi/\lambda = 2\pi/(70 \times 10^{-6} \text{ m}) = 9.0 \times 10^4 \text{ m}^{-1}$$

The absorption coefficient α is $2k''$ as in Eq. (1.8.6) so that

$$\alpha = 2k'' = 2k_o K = 2(9.0 \times 10^4 \text{ m}^{-1})(6) = 1.08 \times 10^6 \text{ m}^{-1}$$

which corresponds to an *absorption depth* $1/\alpha$ of about 0.93 μm . The reflectance is

$$R = \frac{(n - 1)^2 + K^2}{(n + 1)^2 + K^2} = \frac{(4 - 1)^2 + 6^2}{(4 + 1)^2 + 6^2} = 0.74 \text{ or } 74\%$$

At $\lambda = 50 \mu\text{m}$, we first read $K \approx 0.02$, and $n \approx 2$ from Figure 1.26. Repeating the above calculations we get $\alpha \approx 5.08 \times 10^3 \text{ m}^{-1}$, and $R = 0.11$ or 11%. There is a sharp increase in the reflectance from 11% to 72% as we approach the Reststrahlen peak.

At $\lambda = 100 \mu\text{m}$, we first read $K \approx 0.06$, and $n \approx 3.5$ from Figure 1.26. Repeating the above calculations, we find $\alpha = 7.5 \times 10^3 \text{ m}^{-1}$, and $R = 0.31$ or 31%, which is again smaller than the peak reflectance. R is maximum around the Reststrahlen peak. Those materials that strongly absorb light at a given wavelength inside the bulk of the sample will also strongly reflect light at the same wavelength when that light is incident on the surface of the sample.

1.9 TEMPORAL AND SPATIAL COHERENCE

When we represent a traveling EM wave by a pure sinusoidal wave, for example by

$$E_x = E_o \sin(\omega_o t - k_o z) \quad (1.9.1)$$

A sinusoidal wave

with a well-defined angular frequency $\omega_o = 2\pi\nu_o$ and a propagation constant k_o , we are assuming that the wave extends infinitely over all space and exists at all times as illustrated in Figure 1.27 (a) inasmuch as a sine function extends periodically over all values of its argument. Such a sine wave is *perfectly coherent* because all points on the wave are predictable. **Perfect coherence** is therefore understood to mean that we can predict the phase of any portion of the wave from any other portion of the wave. Temporal coherence measures the extent to which two points,

such as P and Q , separated in time at a given location in space can be correlated; that is, one can be reliably predicted from the other. At a given location, for a pure sine wave as in Figure 1.27 (a), any two points such as P and Q separated by any time interval are always correlated because we can predict the phase of one (Q) from the phase of the other (P) for any temporal separation.

Any time-dependent arbitrary function $f(t)$ can be represented by a sum of pure sinusoidal waves with varying frequencies, amplitudes, and phases. The **spectrum** of a function $f(t)$ represents the amplitudes of various sinusoidal oscillations that constitute the function. All these pure sine waves are added, with the right amplitude and phase, to make up $f(t)$. Mathematically, the spectrum of $f(t)$ is found by taking the Fourier transform of $f(t)$.²⁸ We need only one sinusoidal wave at a well-defined frequency $v_o = \omega_o/2\pi$ to make up Eq. (1.9.1) as illustrated in Figure 1.27 (a).

A pure sine wave is an idealization far from reality and in practice a wave can exist only over a finite time duration Δt , which corresponds to a finite wave train of length $l = c\Delta t$ as illustrated in Figure 1.27 (b). This duration Δt may be the result of the radiation emission process, modulation of the output from a laser, or some other process (indeed, in practice, the amplitude will not be constant over Δt). It is clear that we can only correlate points in the wave train within the duration or over the spatial extent $l = c\Delta t$. This wave train has a coherence time Δt and a coherence length $l = c\Delta t$. Since it is not a perfect sine wave, it contains a number of different frequencies in its spectrum. A proper calculation shows that most of the significant frequencies that constitute this finite wave train lie, as expected, centered around v_o over a range Δv as shown in Figure 1.27 (b). The spread Δv is the spectral width of the wave train and depends on the temporal coherence length Δt . In this particular case of a truncated sinusoidal wave, $\Delta v = 2/\Delta t$.

In general, the radiation will not be a perfect truncated sine wave, and the exact relationship between Δv and Δt will depend on the envelope function of the oscillations, which depends on the

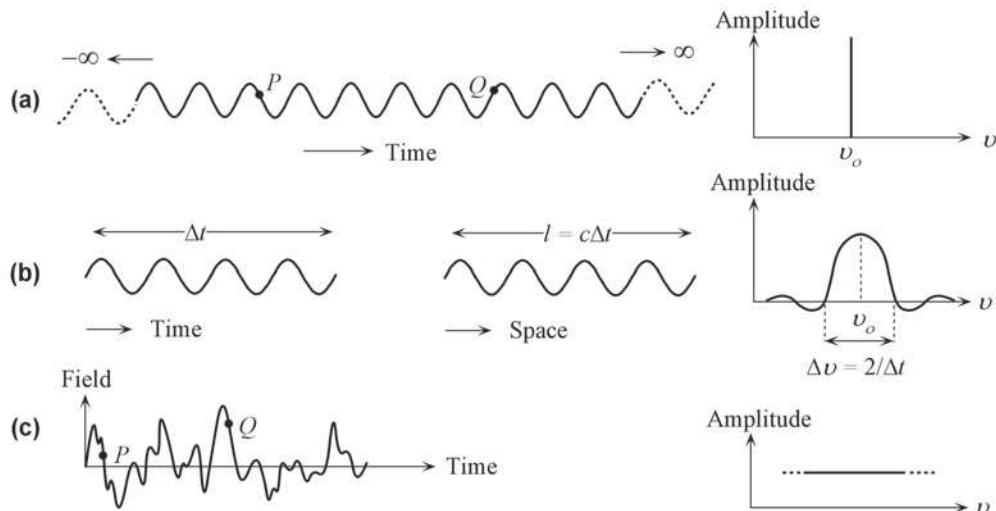


FIGURE 1.27 (a) A sine wave is perfectly coherent and contains a well-defined frequency v_o . (b) A finite wave train lasts for a duration Δt and has a length l . Its frequency spectrum extends over $\Delta v = 2/\Delta t$. It has a coherence time Δt and a coherence length l . (c) White light exhibits practically no coherence.

²⁸Fourier transformation of $f(t)$ involves integrating $f(t)\exp(-j2\pi\nu t)$ over time and produces a function $F(\nu)$ that depends on ν . $F(\nu)$ is called the **spectrum** of $f(t)$, and represents the pure harmonic waves (such as sine waves) that we need to reconstruct $f(t)$. $F(\nu)$ specifies the amplitudes and phases of these harmonics.

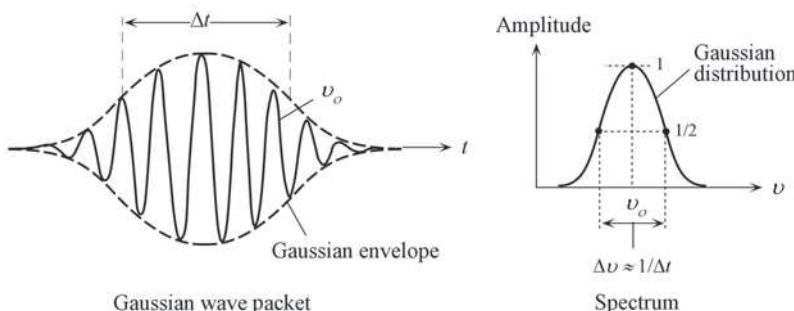


FIGURE 1.28 A Gaussian wave packet which has sinusoidal oscillations at a frequency ν_o and a Gaussian envelope (amplitude variation) over time. Its coherence time Δt between half maximum points is Δt . Its frequency spectrum is Gaussian, centered at the oscillation frequency ν_o , and extends over $\Delta\nu \approx 1/\Delta t$.

way the oscillations were generated or emitted. For example, if one takes the envelope of the oscillations to be Gaussian as shown in Figure 1.28 so that the wave is a *Gaussian wave packet*, then the Fourier transform will also be a Gaussian, centered at the oscillation frequency ν_o . The coherence length and coherence time then usually refer to some suitable widths of the respective Gaussian envelopes. If $\Delta\nu$ and Δt are the full widths at half maximum (FWHM) spreads in the frequency and time domains, as in Figure 1.28, then approximately²⁹

$$\Delta\nu \approx \frac{1}{\Delta t} \quad (1.9.2)$$

Spectral width and coherence time

Equation (1.9.2) is usually known as the **bandwidth theorem**. Although a Gaussian wave packet is useful in representing finite-length light wave packets, it still needs to be suitably truncated to be able to represent more practical pulses since a Gaussian function extends over all times. This truncation does not limit the use of Eq. (1.9.2) since it is meant to apply approximately to any wave packet. Although the exact relationship in Eq. (1.9.2) depends on the shape of the wave packet, Eq. (1.9.2) is still widely used for approximately relating $\Delta\nu$ and Δt for various wave packets.

Coherence and spectral width are therefore intimately linked. For example, the orange radiation at 589 nm (both D-lines together) emitted from a sodium lamp has spectral width $\Delta\nu \approx 5 \times 10^{11}$ Hz. This means that its coherence time is $\Delta t \approx 2 \times 10^{-12}$ s or 2 ps, and its coherence length is 6×10^{-4} m or 0.60 mm. On the other hand, the red lasing emission from a He-Ne laser operating in multimode has a spectral width around 1×10^9 Hz, which corresponds to a coherence length of 30 cm. Furthermore, a continuous wave laser operating in a single mode will have a very narrow linewidth and the emitted radiation will perhaps have a coherence length of several hundred meters. Typically light waves from laser devices have substantial coherence lengths and are therefore widely used in wave-interference studies and applications such as interferometry, holography, and laser Doppler anemometry. Suppose that standing at one location in space we measure the field vs. time behavior shown in Figure 1.27 (c) in which the zero crossing of the signal occurs randomly. Given a point P on this “waveform,” we cannot predict the “phase” or the signal at any other point Q . Thus P and Q are not in any way correlated for any temporal separation except Q coinciding with P (or being very close to it by an infinitesimally short time interval). There is no coherence in this “white” light signal and the signal essentially represents white noise;

²⁹ FWHM is the width of the Gaussian function between the half maximum points (see Appendix A). The bandwidth theorem for a Gaussian wave train is actually $\Delta\nu\Delta t = 0.88$ which can be derived by using Fourier transforms.

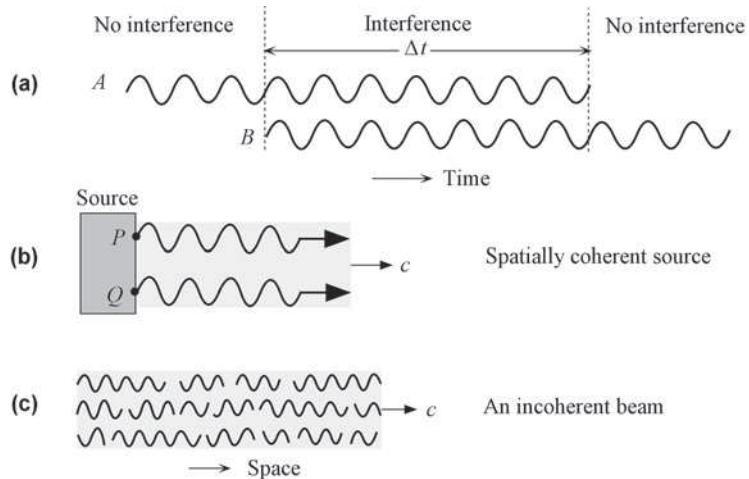


FIGURE 1.29 (a) Two waves can only interfere over the time interval Δt .
 (b) Spatial coherence involves comparing the coherence of waves emitted from different locations on the source.
 (c) An incoherent beam.

its spectrum typically contains a wide range of frequencies. White light, like an ideal sine wave, is an idealization because it assumes all frequencies are present in the light beam. However, radiation emitted from atoms typically has a central frequency and a certain spectral width, that is, a degree of coherence. The light in the real world lies between (a) and (c) in Figure 1.27.

Coherence between two waves is related to the extent of correlation between two waves. The waves A and B in Figure 1.29 (a) have the same frequency v_o but they coincide only over the time interval Δt and hence they can only give rise to interference phenomena over this time. They therefore have **mutual temporal coherence** over the time interval Δt . This situation can arise, for example, when two identical wave trains each of coherence length l travel different optical paths. When they arrive at the destination, they can interfere only over a space portion $c\Delta t$. Since interference phenomena can occur only for waves that have mutual coherence, interference experiments, such as Young's double slit experiments, can be used to measure mutual coherence between waves.

Spatial coherence describes the extent of coherence between waves radiated from different locations on a light source as shown in Figure 1.29 (b). If the waves emitted from locations P and Q on the source are in phase, then P and Q are spatially coherent. A spatially coherent source emits waves that are in phase over its entire emission surface. These waves, however, may have partial temporal coherence. A light beam emerging from a spatially coherent light source will hence exhibit spatial coherence across the beam cross-section, that is, the waves in the beam will be in phase over coherence length $c\Delta t$ in which Δt is the temporal coherence. A mostly incoherent beam will contain waves that have very little correlation with each other. The incoherent beam in Figure 1.29 (c) contains waves (across the beam cross-section) whose phases change randomly at random times. (Note, however, that there may be a very short time interval over which there is a little bit of temporal coherence.) The quantitative analysis of coherence requires mathematical techniques based on correlation functions and may be found in more advanced textbooks.

EXAMPLE 1.9.1 Coherence length of LED light

Light emitting diodes (LEDs) emit light that is seemingly of a well-defined color, or wavelength. However, they have much broader spectral widths than lasers. For example, a red LED emitting at 650 nm will have a spectral line width of 22 nm. Find the coherence time and length of the emitted radiation.

Solution

The frequency and wavelength are related through $v = c/\lambda$, so that differentiating the latter we can find the frequency width $\Delta\nu$ from the wavelength width $\Delta\lambda$

$$\frac{\Delta\nu}{\Delta\lambda} \approx \left| \frac{dv}{d\lambda} \right| = \left| -\frac{c}{\lambda^2} \right|$$

so that

$$\begin{aligned}\Delta\nu &= \Delta\lambda(c/\lambda^2) = (22 \times 10^{-9} \text{ m})(3 \times 10^8 \text{ m s}^{-1})/(650 \times 10^{-9} \text{ m})^2 \\ &= 1.562 \times 10^{13} \text{ Hz}\end{aligned}$$

Thus, the coherence time is

$$\Delta t \approx 1/\Delta\nu = 1/(1.562 \times 10^{13} \text{ Hz}) = 6.40 \times 10^{-14} \text{ s} \quad \text{or} \quad 64.0 \text{ fs}$$

The coherence length is

$$l_c = c\Delta t = 1.9 \times 10^{-5} \text{ m} \quad \text{or} \quad 19 \text{ microns}$$

The above very short coherence length explains why LEDs are not used in interferometry.

1.10 SUPERPOSITION AND INTERFERENCE OF WAVES

Optical interference involves the superposition of two or more electromagnetic waves in which the electric field vectors are added; the fields add vectorially. The waves are assumed to be nearly monochromatic, and have to have the same frequency. Two waves can only interfere if they exhibit *mutual temporal coherence* as in Figure 1.29 (a) at a point in space where they interact. Indeed, interference phenomena can be used to infer on the mutual coherence of the waves. When two waves with the same frequency with fields \mathbf{E}_1 and \mathbf{E}_2 interfere, they generate a resultant field \mathbf{E} that corresponds to the superposition of **individual fields**, that is, $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$. Consider two linearly polarized plane waves that originate from O_1 and O_2 , as schematically shown in Figure 1.30, so that the field oscillations at some arbitrary point of interest P is given by

$$\mathbf{E}_1 = \mathbf{E}_{o1} \sin(\omega t - kr_1 - \phi_1) \quad \text{and} \quad \mathbf{E}_2 = \mathbf{E}_{o2} \sin(\omega t - kr_2 - \phi_2) \quad (1.10.1)$$

where r_1 and r_2 are the distances from O_1 and O_2 to P . These waves have the same ω and k . Due to the process that generates the waves, there is a constant phase difference between them given

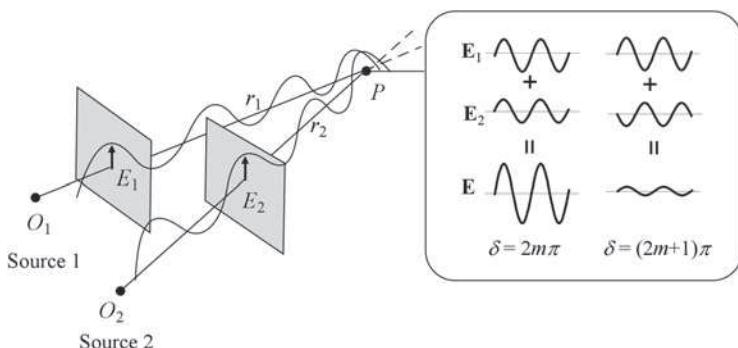


FIGURE 1.30 Interference of two mutually coherent waves of the same frequency originating from sources O_1 and O_2 . We examine the resultant field \mathbf{E} depends on the phase angle δ which depends on the optical path difference $k(r_2 - r_1)$.

by $\phi_2 - \phi_1$. The resultant field at P will be the sum of these two waves, that is, $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$. Its irradiance depends on the time average of $\mathbf{E} \cdot \mathbf{E}$, that is, $\overline{\mathbf{E} \cdot \mathbf{E}}$, so that

$$\overline{\mathbf{E} \cdot \mathbf{E}} = \overline{(\mathbf{E}_1 + \mathbf{E}_2) \cdot (\mathbf{E}_1 + \mathbf{E}_2)} = \overline{\mathbf{E}_1^2} + \overline{\mathbf{E}_2^2} + 2\overline{\mathbf{E}_1 \cdot \mathbf{E}_2}$$

It is clear that the interference effect is in the $2\overline{\mathbf{E}_1 \cdot \mathbf{E}_2}$ term. We can simplify the above equation a little further by assuming \mathbf{E}_{o1} and \mathbf{E}_{o2} are parallel with magnitudes E_{o1} and E_{o2} . Further, irradiance of the interfering waves are $I_1 = \frac{1}{2}c\varepsilon_o E_{o1}^2$ and $I_2 = \frac{1}{2}c\varepsilon_o E_{o2}^2$ so that the resultant irradiance is given by the sum of individual irradiances, I_1 and I_2 , and has an additional third term I_{21} , that is,

Interference

$$I = I_1 + I_2 + 2(I_1 I_2)^{1/2} \cos \delta \quad (1.10.2)$$

where the last term is usually written as $2(I_1 I_2)^{1/2} \cos \delta = I_{21}$, and δ is a phase difference given by

Interference
for mutually
coherent
beams

$$\delta = k(r_2 - r_1) + (\phi_2 - \phi_1) \quad (1.10.3)$$

Since we are using nearly monochromatic waves, $(\phi_2 - \phi_1)$ is constant, and the interference therefore depends on the term $k(r_2 - r_1)$, which represents the phase difference between the two waves as a result of the *optical path difference* between the waves. As we move point P , $k(r_2 - r_1)$ will change because the optical path difference between the two waves will change; and the interference will therefore also change.

Suppose $(\phi_2 - \phi_1) = 0$, the two waves are emitted from a spatially coherent source. Then, if the path difference $k(r_2 - r_1)$ is 0, 2π or a multiple of 2π , that is, $2m\pi$, $m = 0, \pm 1, \pm 2 \dots$, then the interference intensity I will be maximum; such interference is defined as **constructive interference**. If the path difference $k(r_2 - r_1)$ is π or 3π or an odd multiple of π , $(2m + 1)\pi$, then the waves will be 180° out of phase, and the interference intensity will be minimum; such interference is defined as **destructive interference**; both constructive and destructive intensities are shown in Figure 1.30. The maximum and minimum irradiances are given by

Constructive
and
destructive
interference

$$I_{\max} = I_1 + I_2 + 2(I_1 I_2)^{1/2} \quad \text{and} \quad I_{\min} = I_1 + I_2 - 2(I_1 I_2)^{1/2} \quad (1.10.4)$$

If the interfering beams have equal irradiances, then $I_{\max} = 4I_1$ and $I_{\min} = 0$.

It is important to emphasize that we have considered the interference of two nearly monochromatic waves that exhibited mutual temporal and spatial coherence. If the waves do not have any mutual coherence, that is, they are incoherent, then we cannot simply superimpose the electric fields as we did above, and the detector at P , which averages the measurement over its response time, will register an irradiance that is simply the sum of individual irradiances,

Irradiance
of two
super-
imposed
incoherent
beams

$$I = I_1 + I_2 \quad (1.10.5)$$

One of the most described interference experiments is **Young's two slit experiment** that generates an interference fringe. In the modern version of this, a coherent beam of light, as available from a laser, is incident on two parallel slits S_1 and S_2 . There is a screen far away from the slits on which the waves emanating from the slits interfere, as shown in Figure 1.31. The result is an interference pattern that is composed of light and dark regions, corresponding to I_{\max} and I_{\min} . Since S_1 and S_2 are excited by the same wavefront, they emit coherent waves, and we can take $\phi_2 - \phi_1 = 0$. Consider a point P at a distance y on the screen. The phase difference δ at P is then $k(r_2 - r_1)$. As we move P along y , $\delta = k(r_2 - r_1)$ changes and the irradiance on the screen goes through minima and maxima with distance y , following

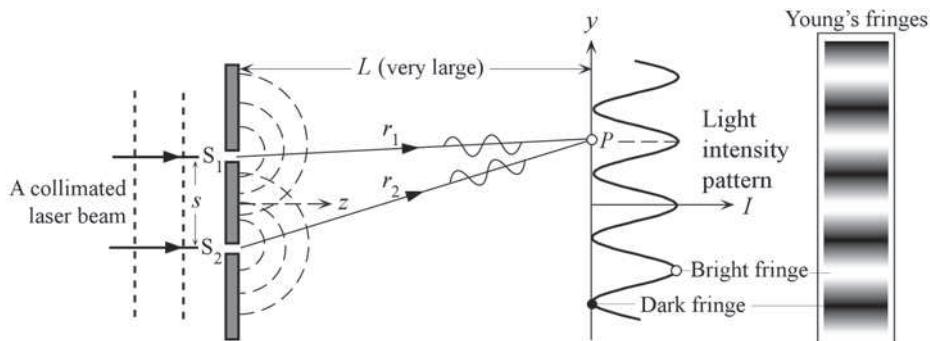


FIGURE 1.31 Young's two slit experiment. Slits S_1 and S_2 , separated by s are illuminated at the same time by coherent (nearly monochromatic) collimated laser beam. The irradiance at the screen shows bright and dark fringes due to the interference of waves emanating from the two slits. The screen is assumed to be far away at a distance L from the slits ($L \gg s$).

Eq. (1.10.2), generating dark and light bands or fringes. The maxima occur when $\delta = 2m\pi$, and minima when $\delta = (2m + 1)\pi$, where $m = 0, \pm 1, \pm 2$. It is not difficult to show that, if the screen is far away, $(r_2 - r_1) \approx (s/L)y$ in which L is the distance from the slits to the screen, and s is the separation of the slits, so that δ is proportional to y . Assuming equal irradiances are emitted from S_1 and S_2 , $I_1 + I_2 = I_o$, the brightness on the screen changes with y periodically as

$$I = I_o \{1 + \cos[(s/L)ky]\} \quad (1.10.6) \text{ Interference}$$

The resulting periodic interference pattern on the screen, as shown in Figure 1.31, is often called **Young's interference fringes**. In a better treatment one needs to consider not only the coherence length of the waves from S_1 and S_2 , but also the diffraction that takes place at each slit due to the finite width of the slit. (See diffraction in Section 1.12.)

1.11 MULTIPLE INTERFERENCE AND OPTICAL RESONATORS



Charles Fabry (1867–1945), left, and Alfred Perot (1863–1925), right, were the first French physicists to construct an optical cavity for interferometry. (Perot: The Astrophysical Journal, Vol. 64, November 1926, p. 208, courtesy of the American Astronomical Society. Fabry: Courtesy of Library of Congress Prints and Photographs Division, Washington, DC 20540, USA.)

An electrical resonator such as a parallel inductor-capacitor (*LC*) circuit allows only electrical oscillations at the resonant frequency f_o (determined by L and C) within a narrow bandwidth around f_o . Such an *LC* circuit thereby stores energy at the same frequency. We know that it also acts as a filter at the resonant frequency f_o , which is how we tune in our favorite radio stations. An **optical resonator** is the optical counterpart of the electrical resonator, storing energy or filtering light only at certain frequencies (wavelengths).

When two flat mirrors are perfectly aligned to be parallel as in Figure 1.32 (a) with free space between them, light wave reflections between the two mirrors M_1 and M_2 lead to constructive and destructive interference of these waves within the cavity. Waves reflected from M_1 traveling toward the right interfere with waves reflected from M_2 traveling toward the left. The result is a series of allowed **stationary or standing EM waves** in the cavity as in Figure 1.32 (b) (just like the stationary waves of a vibrating guitar string stretched between two fixed points). Since the electric field at the mirrors (assume metal coated) must be zero, we can only fit in an integer number m of half-wavelengths, $\lambda/2$, into the cavity length L ,

Fabry–Perot
cavity modes

$$m\left(\frac{\lambda}{2}\right) = L; \quad m = 1, 2, 3, \dots \quad (1.11.1)$$

Each particular allowed λ , labeled as λ_m , satisfying Eq. (1.11.1) for a given m defines a **cavity mode** as depicted in Figure 1.32 (b). Inasmuch as light frequency ν and wavelength λ are related by $\nu = c/\lambda$, the corresponding frequencies ν_m of these modes are the **resonant frequencies** of the cavity

Cavity
resonant
frequencies

$$\nu_m = m\left(\frac{c}{2L}\right) = mv_f; \quad v_f = c/(2L) \quad (1.11.2)$$

in which v_f is the lowest frequency corresponding to $m = 1$, the fundamental mode, and also the frequency separation of two neighboring modes, $\Delta\nu_m = \nu_{m+1} - \nu_m = v_f$. It is known as the **free spectral range**. Figure 1.32 (c) illustrates schematically the intensity of the allowed modes as a function of frequency. If there are no losses from the cavity, that is, the mirrors are perfectly reflecting, then the peaks at frequencies ν_m defined by Eq. (1.11.2) would be sharp lines. If the mirrors are not perfectly reflecting so that some radiation escapes from the cavity, then the mode peaks are not as sharp and have a finite width, as indicated in Figure 1.32 (c). It is apparent that this simple optical cavity serves to “store” radiation energy only at certain frequencies and it is called a **Fabry–Perot optical resonator**. The optical resonator does not have to have two

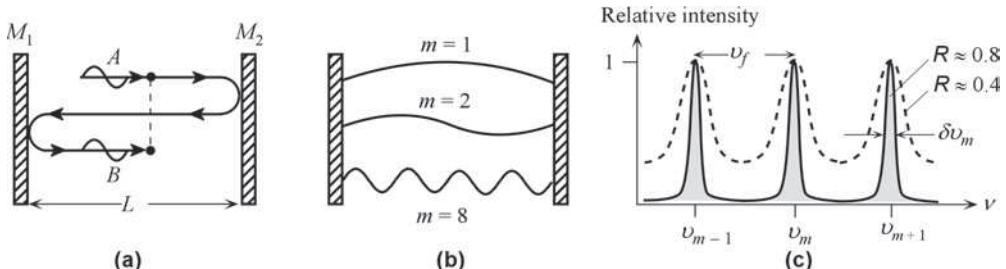


FIGURE 1.32 Schematic illustration of the Fabry–Perot optical cavity and its properties. (a) Reflected waves interfere. (b) Only standing EM waves, modes, of certain wavelengths are allowed in the cavity. (c) Intensity vs. frequency for various modes. R is the mirror reflectance and lower R means higher loss from the cavity.

mirrors and free space between them as shown in Figure 1.32 (a). It can be a solid medium, such as a dielectric (*e.g.*, glass) plate or a rod, whose ends are used to reflect light. The ends of the solid can even have thin film coatings, or dielectric mirrors, to enhance the reflection. Optical resonators are also called **etalons**.

Consider an arbitrary wave such as A traveling toward the right at some instant as shown in Figure 1.32 (a). After one round trip this wave would be again traveling toward the right but now as wave B , which has a phase difference and a different magnitude due to non-perfect reflections. If the mirrors M_1 and M_2 are identical with a reflection coefficient of *magnitude* r , then B has one round-trip phase difference of $k(2L)$ and a magnitude r^2 (two reflections) with respect to A . When A and B interfere, the result is

$$A + B = A + Ar^2 \exp(-j2kL)$$

Of course, just like A , B will continue on and will be reflected twice, and after one round trip it would be going toward the right again and we will now have three waves interfering and so on. After infinite round-trip reflections, the resultant field E_{cavity} is due to infinite such interferences.

$$\begin{aligned} E_{\text{cavity}} &= A + B + \dots \\ &= A + Ar^2 \exp(-j2kL) + Ar^4 \exp(-j4kL) + Ar^6 \exp(-j6kL) + \dots \end{aligned}$$

The sum of this geometric series is easily evaluated as

$$E_{\text{cavity}} = \frac{A}{1 - r^2 \exp(-j2kL)}$$

Once we know the field in the cavity we can calculate the intensity $I_{\text{cavity}} \propto |E_{\text{cavity}}|^2$. Further, we can use *reflectance* $R = r^2$ to further simplify the expression. The final result after algebraic manipulation is

$$I_{\text{cavity}} = \frac{I_o}{(1 - R)^2 + 4R \sin^2(kL)} \quad (1.11.3) \quad \begin{matrix} \text{Cavity} \\ \text{intensity} \end{matrix}$$

in which $I_o \propto A^2$ is the original intensity. The intensity in the cavity is maximum I_{max} whenever $\sin^2(kL)$ in the denominator of Eq. (1.11.3) is zero, which corresponds to (kL) being $m\pi$, in which m is an integer. Thus, the intensity vs. k , or equivalently, the intensity vs. frequency spectrum, peaks whenever $kL = m\pi$, as in Figure 1.32 (c). These peaks are located at $k = k_m$ that satisfy $k_m L = m\pi$, which leads directly to Eqs. (1.11.1) and (1.11.2) that were derived intuitively. For those resonant k_m values, Eq. (1.11.3) gives

$$I_{\text{max}} = \frac{I_o}{(1 - R)^2}; \quad k_m L = m\pi \quad (1.11.4) \quad \begin{matrix} \text{Maximum} \\ \text{cavity} \\ \text{intensity} \end{matrix}$$

A smaller mirror reflectance R means more radiation loss from the cavity, which affects the intensity distribution in the cavity. We can show from Eq. (1.11.3) that smaller R values result in broader mode peaks and a smaller difference between the minimum and maximum intensity in the cavity as schematically illustrated in Figure 1.32 (c). The **spectral width**³⁰ $\delta\nu_m$ of the Fabry–Perot etalon is the full width at half maximum (FWHM) of an individual mode

³⁰The spectral width is also called the fringe width and m the fringe order.

intensity, as defined in Figure 1.32 (c). It can be calculated in a straightforward fashion when $R > 0.6$ from

Spectral width and Finesse

$$\delta\nu_m = \frac{v_f}{F}; \quad F = \frac{\pi R^{1/2}}{1 - R} \quad (1.11.5)$$

in which F is called the **finesse** of the resonator, which increases as losses decrease (R increases). Large finesse lead to sharper mode peaks. Finesse is the ratio of mode separation ($\Delta\nu_m$) to spectral width ($\delta\nu_m$).

We can also define a quality factor Q for the optical resonant cavity in a similar fashion to defining a *Q-factor* for an *LC* oscillator, that is,

Q-factor and Finesse

$$\text{Quality factor, } Q = \frac{\text{Resonant frequency}}{\text{Spectral width}} = \frac{v_m}{\delta\nu_m} = mF \quad (1.11.6)$$

The *Q*-factor is a measure of the frequency selectiveness of a resonator; the higher the *Q*-factor, the more selective the resonator, or narrower the spectral width. It is also a measure of the energy stored in the resonator per unit energy dissipated (due to losses such as from the reflecting surfaces) per cycle of oscillation.

The Fabry–Perot optical cavities are widely used in laser, interference filter, and spectroscopic applications. Consider a light beam that is incident on a Fabry–Perot cavity as in Figure 1.33. The optical cavity is formed by partially transmitting and reflecting plates. Part of the incident beam enters the cavity. We know that only special cavity modes are allowed to exist in the cavity since other wavelengths lead to destructive interference. Thus, if the incident beam has a wavelength corresponding to one of the cavity modes, it can sustain oscillations in the cavity and hence lead to a transmitted beam. The output light is a fraction of the light intensity in the cavity and is proportional to Eq. (1.11.3). Commercial interference filters are based on this principle except that they typically use two cavities in series formed by dielectric mirrors (a stack of quarter wavelength layers); the structure is more complicated than in Figure 1.33. Further, adjusting the cavity length L provides a “tuning capability” to scan different wavelengths.

Equation (1.11.3) describes the intensity of the radiation in the cavity. The intensity of the transmitted radiation in Figure 1.33 can be calculated, as above, by considering that each time a wave is reflected at the right mirror, a portion of it is transmitted, and that these transmitted waves can interfere only constructively to constitute a transmitted beam when $kL = m\pi$. Intuitively, if I_{incident} is the incident light intensity, then a fraction $(1 - R)$ of this would enter

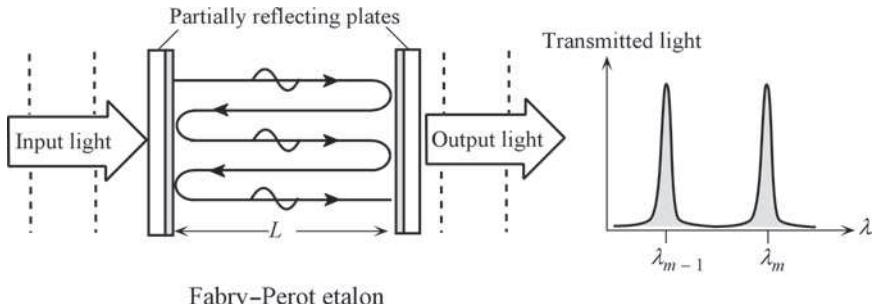


FIGURE 1.33 Transmitted light through a Fabry–Perot optical cavity.



Left: Fused silica etalon. Right:
A 10 GHz air spaced etalon with
3 zerodur spacers. (Courtesy of
Light Machinery Inc.)

the cavity to build up into I_{cavity} in Eq. (1.11.3), and a fraction $(1 - R)$ of I_{cavity} would leave the cavity as the transmitted intensity $I_{\text{transmitted}}$. Thus,

$$I_{\text{transmitted}} = I_{\text{incident}} \frac{(1 - R)^2}{(1 - R)^2 + 4R \sin^2(kL)} \quad (1.11.7) \quad \begin{matrix} \text{Transmitted} \\ \text{mode} \\ \text{intensities} \end{matrix}$$

which is again maximum just as for I_{cavity} whenever $kL = m\pi$ as shown in terms of wavelength in Figure 1.33.³¹

The ideas above can be readily extended to a medium with a refractive index n by using nk for k or λ/n for λ where k and λ are the free-space propagation constant and wavelength, respectively. Equations (1.11.1) and (1.11.2) become

$$m \left(\frac{\lambda}{2n} \right) = L \quad m = 1, 2, 3, \dots \quad (1.11.8) \quad \begin{matrix} \text{Fabry-Perot} \\ \text{cavity modes} \\ \text{in a medium} \end{matrix}$$

$$v_m = m \left(\frac{c}{2nL} \right) = mv_f; \quad v_f = c/(2nL) \quad (1.11.9) \quad \begin{matrix} \text{Cavity} \\ \text{resonant} \\ \text{frequencies} \end{matrix}$$

Further, if the angle of incidence θ at the etalon face is not normal, then we can resolve k to be along the cavity axis; that is use $k \cos \theta$ instead of k in the discussions above.

The two mirrors in Figure 1.32 (a) were assumed to have the same reflectance R . Suppose that R_1 and R_2 are the reflectances of the mirrors M_1 and M_2 . Then, we can continue to use the above equations by using an average geometric reflectance, that is $R = (R_1 R_2)^{1/2}$.

EXAMPLE 1.11.1 Resonator modes and spectral width of a semiconductor Fabry-Perot cavity

Consider a Fabry-Perot optical cavity made of a semiconductor material with mirrors at its ends. (The mirrors have been obtained by coating the end of the semiconductor crystal.) The length of the semiconductor, and hence the cavity, is 250 μm and mirrors at the ends have a reflectance of 0.90. Calculate the cavity mode nearest to the free-space wavelength of 1310 nm. Calculate the separation of the modes, finesse, spectral width of each mode in frequency and wavelength, and the Q -factor.

³¹The term on the right multiplying I_{incident} in Eq. (1.11.7) is usually known as the *Airy function*.

Solution

The wavelength of radiation inside the cavity is λ/n , where n is the refractive index of the medium, which is 3.6, and λ is the free-space wavelength. We need to use Eq. (1.11.8), so that the mode number for the wavelength 1310 nm is

$$m = \frac{2nL}{\lambda} = \frac{2(3.6)(250 \times 10^{-6})}{(1310 \times 10^{-9})} = 1374.05$$

which must be an integer (1374) so that the actual mode wavelength is

$$\lambda_m = \frac{2nL}{m} = \frac{2(3.6)(250 \times 10^{-6})}{(1374)} = 1310.04 \text{ nm}$$

Thus, for all practical purposes the mode wavelength is 1310 nm.

The mode frequency is $v_m = c/\lambda_m$ so that the separation of the modes, from Eq. (1.11.9), is

$$\Delta v_m = v_f = \frac{c}{2nL} = \frac{(3 \times 10^8)}{2(3.6)(250 \times 10^{-6})} = 1.67 \times 10^{11} \text{ Hz} \quad \text{or} \quad 167 \text{ GHz}$$

The finesse is

$$F = \frac{\pi R^{1/2}}{1 - R} = \frac{\pi 0.90^{1/2}}{1 - 0.90} = 29.8$$

and the spectral width of each mode is

$$\delta v_m = \frac{v_f}{F} = \frac{1.67 \times 10^{11}}{29.8} = 5.59 \times 10^9 \text{ Hz} \quad \text{or} \quad 5.59 \text{ GHz}$$

The mode spectral width δv_m will correspond to a certain spectral wavelength width $\delta \lambda_m$. The mode wavelength $\lambda_m = 1310 \text{ nm}$ corresponds to a mode frequency $v_m = c/\lambda_m = 2.29 \times 10^{14} \text{ Hz}$. Since $\lambda_m = c/v_m$, we can differentiate this expression to relate small changes in λ_m and v_m ,

$$\delta \lambda_m = \delta \left(\frac{c}{v_m} \right) = \left| -\frac{c}{v_m^2} \right| \delta v_m = \frac{(3 \times 10^8)}{(2.29 \times 10^{14})^2} (5.59 \times 10^9) = 3.2 \times 10^{-11} \text{ m} \quad \text{or} \quad 0.032 \text{ nm}$$

The Q -factor is

$$Q = mF = (1374)(29.8) = 4.1 \times 10^4$$

An optical cavity like this is used in so-called Fabry–Perot semiconductor laser diodes, and will be discussed further in Chapter 4.

1.12 DIFFRACTION PRINCIPLES

A. Fraunhofer Diffraction

An important property of waves is that they exhibit diffraction effects; for example, sound waves are able to bend (deflect around) corners and a light beam can similarly “bend” around an obstruction (though the bending may be very small). Figure 1.34 shows an example of a collimated light beam passing through a circular aperture (a circular opening in an opaque screen). The passing beam is found to be divergent and to exhibit an intensity pattern that has bright and dark rings, called *Airy rings*.³² The passing beam is said to be diffracted and its light intensity pattern is called

³²Sir George Airy (1801–1892), Astronomer Royal of Britain from 1835 to 1881.

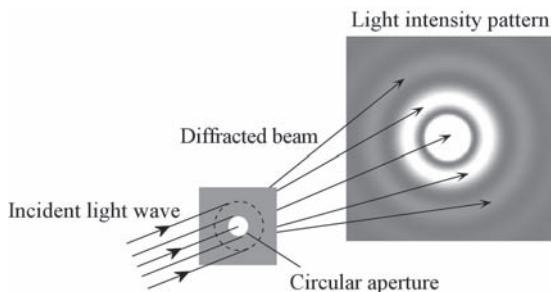


FIGURE 1.34 A collimated light beam incident on a small circular aperture becomes diffracted and its light intensity pattern after passing through the aperture is a diffraction pattern with circular bright rings (called Airy rings). If the screen is far away from the aperture, this would be a Fraunhofer diffraction pattern.

a **diffraction pattern**. Clearly, the light pattern of the diffracted beam does *not* correspond to the geometric shadow of the circular aperture. Diffraction phenomena are generally classified into two categories. In **Fraunhofer diffraction**,³³ the incident light beam is a plane wave (a collimated light beam) and the observation or detection of the light intensity pattern (by placing a photographic screen, *etc.*) is done far away from the aperture so that the waves received also look like plane waves. Inserting a lens between the aperture and the photographic screen enables the screen to be closer to the aperture. In **Fresnel diffraction**, the incident light beam and the received light waves are not plane waves but have significant wavefront curvatures. Typically, the light source and the photographic screen are both close to the aperture so that the wavefronts are curved. Fraunhofer diffraction is by far the most important; and it is mathematically easier to treat.

Diffraction can be understood in terms of the interference of multiple waves emanating from the aperture in the obstruction.³⁴ We will consider a plane wave incident on a one-dimensional slit of length a . According to the Huygens-Fresnel principle,³⁵ *every unobstructed point of a wavefront, at a given instant in time, serves as a source of spherical secondary waves (with the same frequency as that of the primary wave). The amplitude of the optical field at any point beyond is the superposition of all these wavelets (considering their amplitudes and relative phases).* Figures 1.35 (a) and (b) illustrates this point pictorially showing that, when the plane wave reaches the aperture, points in the aperture become sources of coherent spherical secondary waves. These spherical waves interfere to constitute the new wavefront (the new wavefront is the envelope of the wavefronts of these secondary waves). These spherical waves can interfere constructively not just in the forward direction as in (a) but also in other appropriate directions, as in (b), giving rise to the observed bright and dark patterns (rings for a circular aperture) on the observation screen.

We can divide the unobstructed width a of the aperture into a very large number N of coherent “point sources” each of extent $\delta y = a/N$ (obviously δy is sufficiently small to be nearly a point), as in Figure 1.36 (a). Since the aperture a is illuminated uniformly by the plane wave, the strength (amplitude) of each point source would be proportional to $a/N = \delta y$. Each would be a source of spherical waves. In the forward direction ($\theta = 0$), they would all be in phase and constitute a forward wave, along the z -direction. But they can also be in phase at some angle θ to the z -direction and hence give rise to a diffracted wave along this direction. We will evaluate the intensity of the received wave at a point on the screen as the *sum of all*

³³Joseph von Fraunhofer (1787–1826) was a German physicist who also observed the various dark lines in Sun’s spectrum due to hydrogen absorption.

³⁴“No one has been able to define the difference between interference and diffraction satisfactorily” [R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics* (Addison-Wesley, 1963)].

³⁵Eugene Hecht, *Optics*, 4th Edition (Pearson Education, 2002), Ch. 10, p. 444.

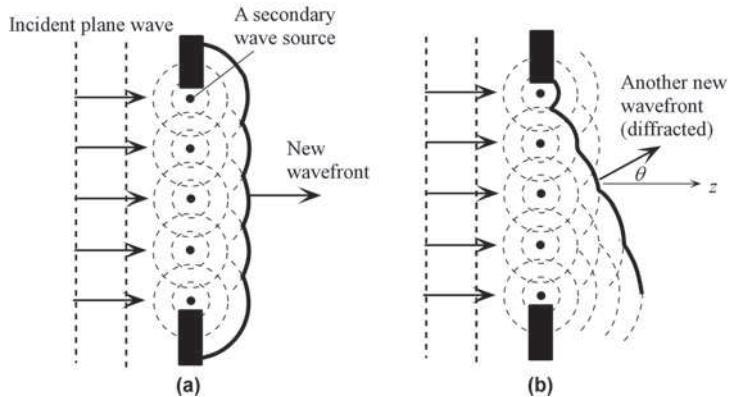


FIGURE 1.35 (a) Huygens-Fresnel principle states that each point in the aperture becomes a source of secondary waves (spherical waves). The spherical wavefronts are separated by λ . The new wavefront is the envelope of all these spherical wavefronts. (b) Another possible wavefront occurs at an angle θ to the z -direction, which is a diffracted wave.

waves arriving from all point sources in the aperture. The screen is far away from the aperture so the waves arrive almost parallel at the screen (alternatively a lens can be used to focus the diffracted parallel rays to form the diffraction pattern).

Consider an arbitrary direction θ , and consider the phase of the emitted wave (Y) from an arbitrary point source at y with respect to the wave (A) emitted from source at $y = 0$ as shown in Figure 1.36 (a). If k is the propagation constant, $k = 2\pi/\lambda$, the wave Y is out of phase with respect to A by $ky \sin \theta$. Thus the wave emitted from the point source at y has a field δE ,

$$\delta E \propto (\delta y) \exp(-jk y \sin \theta) \quad (1.12.1)$$

All of these waves from point sources from $y = 0$ to $y = a$ interfere at the screen at a point P that makes an angle θ at the slit, and the resultant field at the screen at this point P is

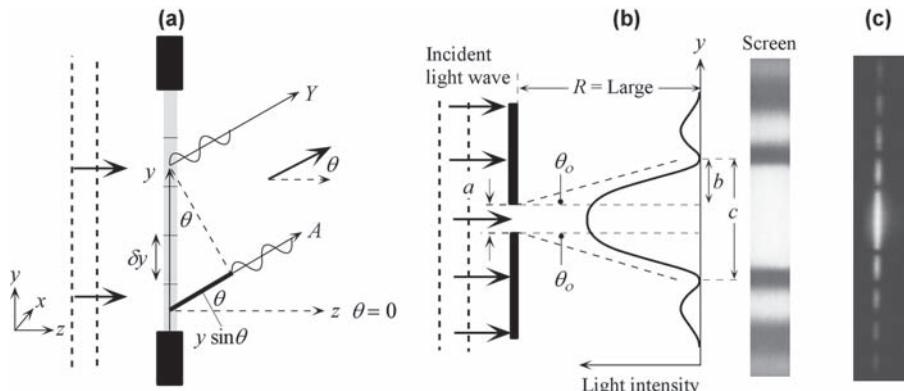


FIGURE 1.36 (a) The aperture has a finite width a along y , but it is very long along x so that it is a one-dimensional slit. The aperture is divided into N number of point sources each occupying δy with amplitude proportional to δy since the slit is excited by a plane electromagnetic wave. (b) The intensity distribution in the received light at the screen far away from the aperture: the diffraction pattern. Note that the slit is very long along x so that there is no diffraction along this dimension. The incident wave illuminates the whole slit. (c) Typical diffraction pattern using a laser pointer on a single slit. The difference from the pattern in (b) is due to the finite size of the laser pointer beam along x that is smaller than the length of the slit.

their sum. Because the screen is far away, a point on the screen is at the same distance from anywhere in the aperture. This means that all the spherical waves from the aperture experience the same phase change and decrease in amplitude in reaching the screen. This simply scales δE at the screen by an amount that is the same for all waves coming from the aperture. Thus, the resultant field $E(\theta)$ at point P at the screen is

$$E(\theta) = C \int_{y=0}^{y=a} \delta y \exp(-jk y \sin \theta) \quad (1.12.2)$$

in which C is a constant. Integrating Eq. (1.12.2) we get

$$E(\theta) = \frac{Ce^{-j\frac{1}{2}ka \sin \theta} a \sin\left(\frac{1}{2}ka \sin \theta\right)}{\frac{1}{2}ka \sin \theta}$$

The light intensity I at a point at P at the screen is proportional to $|E_\theta|^2$, and thus

$$I(\theta) = \left[\frac{C' a \sin\left(\frac{1}{2}ka \sin \theta\right)}{\frac{1}{2}ka \sin \theta} \right]^2 = I(0) \text{sinc}^2(\beta); \quad \beta = \frac{1}{2}(ka \sin \theta) \quad (1.12.3)$$

Single slit diffraction equation

in which C' is a constant and β is a convenient new variable representing θ , and sinc ("sink") is a function that is defined by $\text{sinc}(\beta) = \sin(\beta)/(\beta)$.

If we were to plot Eq. (1.12.3) as a function of θ at the screen we would see the intensity (diffraction) pattern schematically depicted in Figure 1.36 (b). First, observe that the pattern has bright and dark regions, corresponding to constructive and destructive interference of waves emanating from the aperture. Second, the center bright region is wider than the aperture width a , which mean that the transmitted beam must be *diverging*. The zero intensity occurs when, from Eq. (1.12.3),

$$\sin \theta = \frac{m\lambda}{a}; \quad m = \pm 1, \pm 2, \dots \quad (1.12.4)$$

Zero intensity points

The angle θ_o for the first zero, corresponding to $m = \pm 1$, is given by $\theta_o = \pm \lambda/a$, where we assumed that the divergence is small (usually the case) so that $\sin \theta_o \approx \theta_o$. Thus, the divergence $\Delta\theta$, the angular spread, of the diffracted beam is given by

$$\Delta\theta = 2\theta_o \approx \frac{2\lambda}{a} \quad (1.12.5)$$

Divergence from single slit of width a

A light wave at a wavelength 1300 nm, diffracted by a slit of width $a = 100 \mu\text{m}$ (about the thickness of this page), has a divergence $\Delta\theta$ of about 1.5° . From Figure 1.36 (b), it is apparent that, using geometry, we can easily calculate the width c of the central bright region of the intensity pattern, given θ_o from Eq. (1.12.5) and the distance R of the screen from the aperture.

The diffraction patterns from two-dimensional apertures such as rectangular and circular apertures are more complicated to calculate but they use the same principle based on the multiple interference of waves emitted from all point sources in the aperture. The diffraction pattern of a rectangular aperture is shown in Figure 1.37. It involves the multiplication of two individual single slit (sinc) functions, one slit of width a along the horizontal axis, and the other of width b along the vertical axis. (Why is the diffraction pattern wider along the horizontal axis?)

The diffraction pattern from a circular aperture, known as **Airy rings**, was shown in Figure 1.34, and can be roughly visualized by rotating the intensity pattern in Figure 1.36 (b)

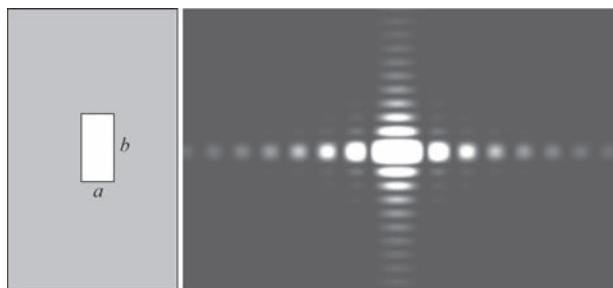


FIGURE 1.37 The rectangular aperture of dimensions $a \times b$ on the left gives the diffraction pattern on the right (b is twice a).

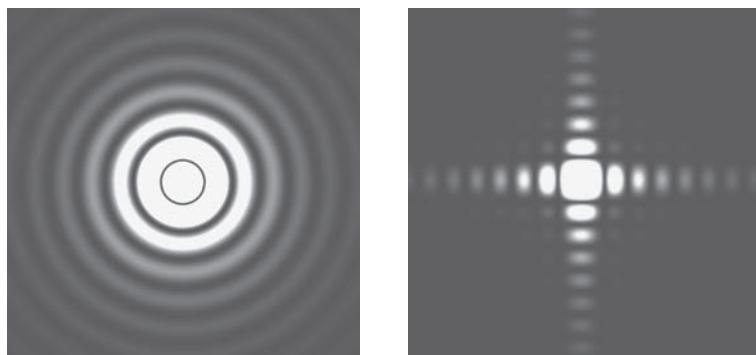
about the z -axis. We can, as we did for the single slit, sum all waves emanating from every point in the circular aperture, taking into account their relative phases when they arrive at the screen to obtain the actual intensity pattern at the screen. The result is that the diffraction pattern is a Bessel function of the first kind,³⁶ and not a simply rotated sinc function. The central white spot is called the **Airy disk**; its radius corresponds to the radius of the first dark ring. We can still use Figures 1.36 (a) and (b) to imagine how diffraction occurs from a circular aperture by taking this as a cut through the aperture so that a is now the diameter of the aperture, denoted as D . The angular position θ_o of the first dark ring, as defined as in Figure 1.36 (b), is determined by the diameter D of the aperture and the wavelength λ , and is given by

Angular
radius of
Airy disk

$$\sin \theta_o = 1.22 \frac{\lambda}{D} \quad (1.12.6)$$

The divergence angle from the aperture center to the Airy disk circumference is $2\theta_o$. If R is the distance of the screen from the aperture, then the radius of the Airy disk, approximately b , can be calculated from the geometry in Figure 1.36 (b), which gives $b/R = \tan \theta_o \approx \theta_o$. If a lens is used to focus the diffracted light waves onto a screen, then $R = f$, focal length of the lens.

It is worth commenting on the Gaussian beam at this point. Suppose we now examine a Gaussian beam with a waist $2w_o$ that is the same as the aperture size D . The far field



Diffraction pattern far away from a circular aperture.

Diffraction pattern far away from a square aperture.

³⁶Bessel functions are special mathematical functions, which can be looked up in mathematics handbooks. They are used in various engineering problems.

half-divergence angle θ of this Gaussian beam would be $\theta = (2/\pi)(\lambda/D)$ or $0.637(\lambda/D)$ as in Eq. (1.1.7). The Gaussian beam has a smaller divergence than the diffracted beam from a circular aperture. The difference is due the fact that each point in the circular aperture emits with the same intensity because the aperture is illuminated by a plane wave. If we were to change the emission intensity within the aperture to follow a Gaussian distribution, we would see a Gaussian beam as the “diffracted beam.” The Gaussian beam is a self-diffracted beam and has the smallest divergence for a given beam diameter.

EXAMPLE 1.12.1 Resolving power of imaging systems

Consider what happens when two neighboring point light sources are examined through an imaging system with an aperture of diameter D (this may even be a lens). The two sources have an angular separation of $\Delta\theta$ at the aperture. The aperture produces a diffraction pattern of the sources S_1 and S_2 , as shown in Figure 1.38. As the points get closer, their angular separation becomes narrower and the diffraction patterns overlap more. According to the **Rayleigh criterion**, the two spots are just resolvable when the principal maximum of one diffraction pattern coincides with the minimum of the other, which is given by the condition

$$\sin(\Delta\theta_{\min}) = 1.22 \frac{\lambda}{D} \quad (1.12.7)$$

Angular limit of resolution

The human eye has a pupil diameter of about 2 mm. What would be the minimum angular separation of two points under a green light of 550 nm and their minimum separation if the two objects are 30 cm from the eye? The image will be a diffraction pattern in the eye, and is a result of waves in this medium. If the refractive index $n \approx 1.33$ (water) in the eye, then Eq. (1.12.7) is

$$\sin(\Delta\theta_{\min}) = 1.22 \frac{\lambda}{nD} = 1.22 \frac{(550 \times 10^{-9} \text{ m})}{(1.33)(2 \times 10^{-3} \text{ m})}$$

giving

$$\Delta\theta_{\min} = 0.0145^\circ$$

Their minimum separation s would be

$$s = 2L \tan(\Delta\theta_{\min}/2) = 2(300 \text{ mm}) \tan(0.0145^\circ/2) = 0.076 \text{ mm} = 76 \mu\text{m}$$

which is about the thickness of a human hair (or this page).

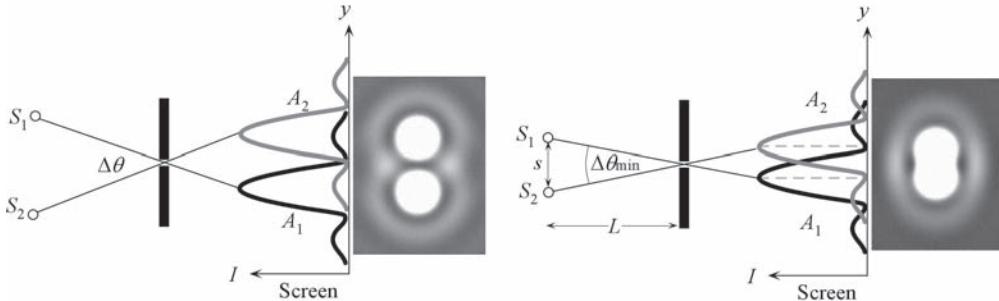


FIGURE 1.38 Resolution of imaging systems is limited by diffraction effects. As points S_1 and S_2 get closer, eventually the Airy patterns overlap so much that the resolution is lost. The Rayleigh criterion allows the minimum angular separation of two of the point sources to be determined. (Schematic illustration inasmuch as the side lobes are actually much smaller than the center peak.)

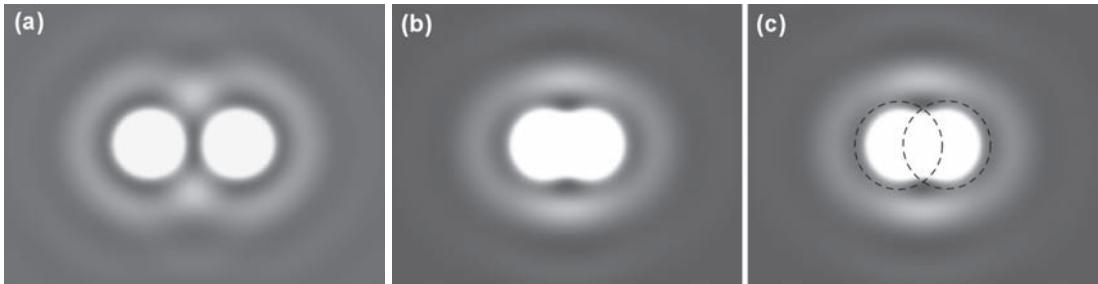


Image of two-point sources captured through a small circular aperture. (a) The two points are fully resolved since the diffraction patterns of the two sources are sufficiently separated. (b) The two images are near the Rayleigh limit of resolution. (c) The first dark ring of one image passes through the center of the bright Airy disk of the other. (Approximate.)

B. Diffraction Grating

A **diffraction grating** in its simplest form is an optical device that has a periodic series of slits in an opaque screen as shown in Figure 1.39 (a). An incident beam of light is diffracted in certain well-defined directions that depend on the wavelength λ and the grating properties. Figure 1.39 (b) shows a typical intensity pattern in the diffracted beam for a finite number of slits. There are “strong beams of diffracted light” along certain directions (θ) and these are labeled according to their occurrence: zero-order (center), first-order, either side of the zero-order, and so on. If there are an infinite number of slits then the diffracted beams have the same intensity. In reality, any periodic variation in the refractive index would serve as a diffraction grating and we will discuss other types later. As in Fraunhofer diffraction we will assume that the observation screen is far away, or that a lens is used to focus the diffracted parallel rays on to the screen (the lens in the observer’s eye does it naturally).

We will assume that the incident beam is a plane wave so that the slits become coherent (synchronous) sources. Suppose that the width a of each slit is much smaller than the separation

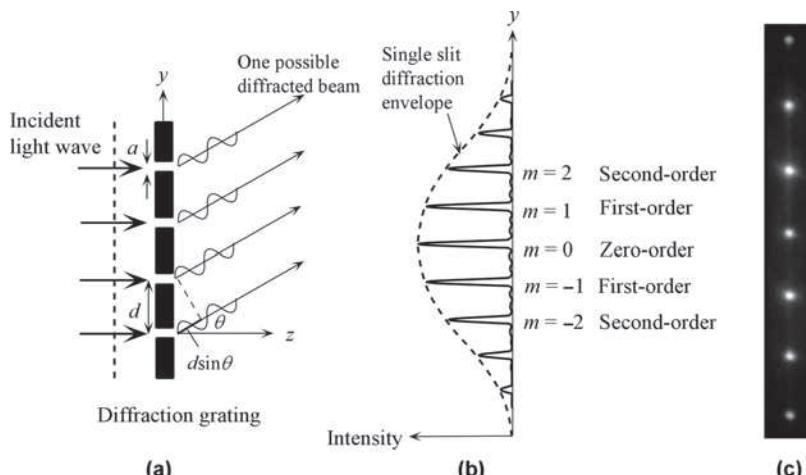


FIGURE 1.39 (a) A diffraction grating with N slits in an opaque screen. Slit periodicity is d and slit width is a ; $a \ll d$. (b) The diffracted light pattern. There are distinct, that is diffracted, beams in certain directions (schematic). (c) Diffraction pattern obtained by shining a beam from a red laser pointer onto a diffraction grating. The finite size of the laser beam results in the dot pattern. (The wavelength was 670 nm, red, and the grating has 200 lines per inch.)

d of the slits as shown in Figure 1.39 (a). Waves emanating at an angle θ from two neighboring slits are out of phase by an amount that corresponds to an optical path difference $d \sin \theta$. Obviously, all such waves from pairs of slits will interfere constructively when this is a multiple of the whole wavelength

$$d \sin \theta = m\lambda; \quad m = 0, \pm 1, \pm 2, \dots \quad (1.12.8)$$

Grating equation

which is the well-known **grating equation**, also known as the **Bragg³⁷ diffraction condition**. The value of m defines the diffraction order; $m = 0$ being zero-order, $m = \pm 1$ being first-order, etc. If the grating in Figure 1.39 (a) is in a medium of refractive index n , that is, the incident and diffracted beams are all in the same medium of index n , then we should use λ/n for the wavelength in Eq. (1.12.8), where λ is the free-space wavelength, that is, $d \sin \theta = m\lambda/n$.

The problem of determining the actual intensity of the diffracted beam is more complicated as it involves summing all such waves at the observer and, at the same time, including the diffraction effect of each individual narrow slit. With a smaller than d as in the Figure 1.39 (a), the amplitude of the diffracted beam is modulated by the diffraction amplitude of a single slit since the latter is spread substantially, as illustrated in Figure 1.39 (b). It is apparent that the diffraction grating provides a means of deflecting an incoming light by an amount that depends on its wavelength—the reason for their use in *spectroscopy*.

The diffraction grating in Figure 1.40 (a) is a **transmission grating**. The incident and diffracted beams are on opposite sides of the grating. Typically, parallel thin grooves on a glass plate would serve as a transmission grating as in Figure 1.40 (a). A **reflection grating** has the incident beam and the diffracted beams on the same side of the device as in Figure 1.40 (b). The surface of the device has a periodic reflecting structure, easily formed by etching parallel grooves in a metal film, etc. The reflecting unetched surfaces serve as synchronous secondary sources that interfere along certain directions to give diffracted beams of zero-order, first-order, etc. Among transmission gratings, it is customary to distinguish between *amplitude gratings* in which the transmission amplitude is modulated, and so-called *phase gratings* where only the refractive index is modulated, without any losses.

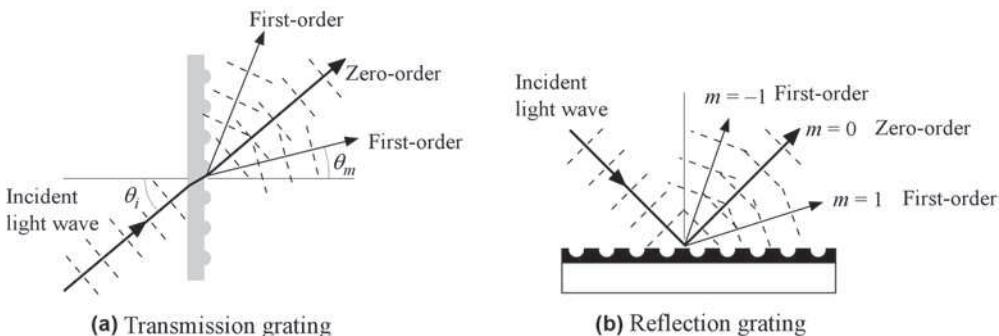


FIGURE 1.40 (a) Ruled periodic parallel scratches on a glass serve as a *transmission grating*. (The glass plate is assumed to be very thin.) (b) A *reflection grating*. An incident light beam results in various “diffracted” beams. The zero-order diffracted beam is the normal reflected beam with an angle of reflection equal to the angle of incidence.

³⁷William Lawrence Bragg (1890–1971), Australian-born British physicist, won the Nobel Prize with his father, William Henry Bragg, for his “famous equation” when he was only 25 years old.

When the incident beam is not normal to the diffraction grating, then Eq. (1.12.8) must be modified. If θ_i is the angle of incidence with respect to the normal to the grating, then the diffraction angle θ_m for the m -th mode is given by

Grating equation

$$d(\sin \theta_m - \sin \theta_i) = m\lambda; \quad m = 0, \pm 1, \pm 2, \dots \quad (1.12.9)$$

The same equation can be used for transmission and reflection gratings provided that we define the angles θ_i and θ_m as positive on either side of the normal as in Figure 1.40 (b).³⁸ Consider a grating with N slits. The slit width is a (very narrow), and d is the periodicity as before. The detector is at a distance L , far away from the grating. While the periodicity in the slits gives rise to the diffracted beams, the diffraction at each narrow slit defines the envelope of the diffracted intensities as shown in Figure 1.39 (b). If the incident plane wave is normal to the grating, the intensity distribution along y at the screen is given by

Grating diffraction pattern

$$I(y) = I_o \left[\frac{\sin \left(\frac{1}{2} k_y a \right)}{\frac{1}{2} k_y a} \right]^2 \left[\frac{\sin \left(\frac{1}{2} N k_y d \right)}{N \sin \left(\frac{1}{2} k_y d \right)} \right]^2 \quad (1.12.10)$$

where k_y is the scattering wave vector defined by $k_y = (2\pi/\lambda)(y/L) = (2\pi/\lambda)\sin \theta$, and I_o is the maximum intensity along $\theta = 0$. The second term represents the oscillations in the intensity due to interference from different slits. The first term is the envelope of the diffraction pattern, and is the diffraction pattern of a single slit.

The **resolvance** or the **resolving power** R of a diffraction grating is its ability to be able to separate out adjacent wavelengths. If $\lambda_2 - \lambda_1 = \Delta\lambda$ is the minimum wavelength separation that can be measured, as determined by the Rayleigh criterion (the maximum of the intensity distribution at λ_1 is at the first minimum of the intensity distribution at λ_2), and λ is the average wavelength $(1/2)(\lambda_1 + \lambda_2)$ in $\Delta\lambda$, then the **resolving power** is defined by

Resolving power

$$R = \lambda / \Delta\lambda \quad (1.12.11)$$

The separation $\Delta\lambda$ is also called the spectral resolution. If N grooves on a grating are illuminated and the order of diffraction is m , the theoretical resolving power is given simply by $R = mN$. The resolving power is also called the **chromatic resolving power** since it refers to the separation of wavelengths.

Diffraction gratings are widely used in spectroscopic applications because of their ability to provide light deflection that depends on the wavelength. In such applications, the undiffracted light that corresponds to the zero-order beam (Figure 1.40) is clearly not desirable because it wastes a portion of the incoming light intensity. Is it possible to shift this energy to a higher order? Robert William Wood (1910) was able to do so by ruling grooves on glass with a controlled shape as in Figure 1.41 (a) where the surface is angled periodically with a spatial period d . The diffraction condition in Eq. (1.12.9) applies with respect to the normal to the grating plane, whereas the first-order reflection corresponds to reflection from the flat surface, which is at an angle γ . Thus it is possible to “blaze” one of the higher orders (usually $m = 1$) by appropriately choosing γ . Most modern diffraction gratings are of this type. If the angle of incidence is θ_i with respect to the grating normal, then specular reflection occurs at an angle $(\gamma + \theta_i)$ with respect to the face normal and $(\gamma + \theta_i) + \gamma$ with respect to the grating normal. This reflection at $(\gamma + \theta_i) + \gamma$ should occur at diffraction angle θ_m so that

Blazing angle

$$2\gamma = \theta_m - \theta_i \quad (1.12.12)$$

³⁸ Some books use $d(\sin \theta_m + \sin \theta_i) = m\lambda$ for a transmission grating but the angles become positive on the incidence side and negative on the transmitted side with respect to the normal. It is a matter of sign convention.

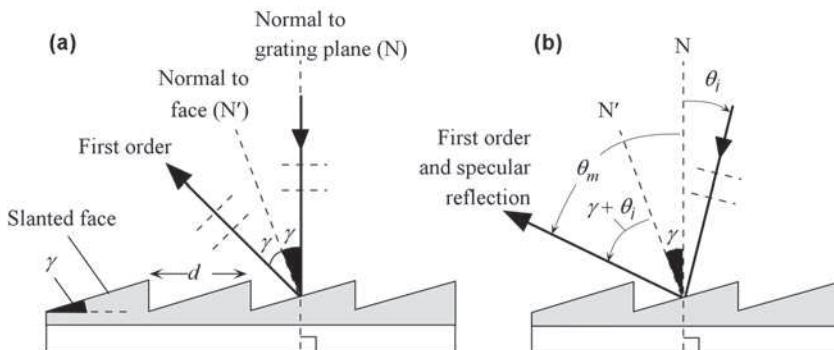
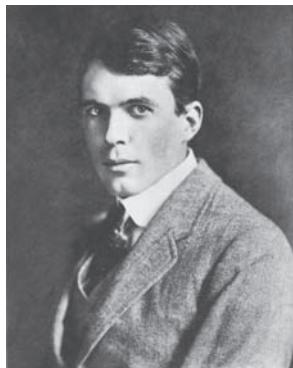


FIGURE 1.41 (a) A blazed grating. Triangular grooves have been cut into the surface with a periodicity d . The side of a triangular groove makes an angle γ to the plane of the diffraction angle. For normal incidence, the angle of diffraction must be 2γ to place the specular reflection on the diffracted beam. (b) When the incident beam is not normal, the specular reflection will coincide with the diffracted beam when $(\gamma + \theta_i) + \gamma = \theta_m$.



William Lawrence Bragg (1890–1971), Australian-born British physicist, won the Nobel Prize with his father, William Henry Bragg, for his “famous equation” when he was only 25 years old. (*SSPL via Getty Images*.)

“The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them.”

EXAMPLE 1.12.2 A reflection grating

Consider a reflection grating with a period d that is $10 \mu\text{m}$ as in Figure 1.42 (a). Find the diffracted beams if a collimated light wave of wavelength 1550 nm is incident on the grating at an angle of 45° to its normal. What should be the blazing angle γ if we were to use a blazed grating with the same periodicity? What happens to the diffracted beams if the periodicity is reduced to $2 \mu\text{m}$?

Solution

If we put $m = 0$ in Eq. (1.12.9) we would find the zero-order diffraction, which is at an angle 45° , as expected, and shown in Figure 1.42 (a). The general Bragg diffraction condition is

$$d(\sin \theta_m - \sin \theta_i) = m\lambda$$

so that

$$(10 \mu\text{m})(\sin \theta_m - \sin(45^\circ)) = (+1)(1.55 \mu\text{m})$$

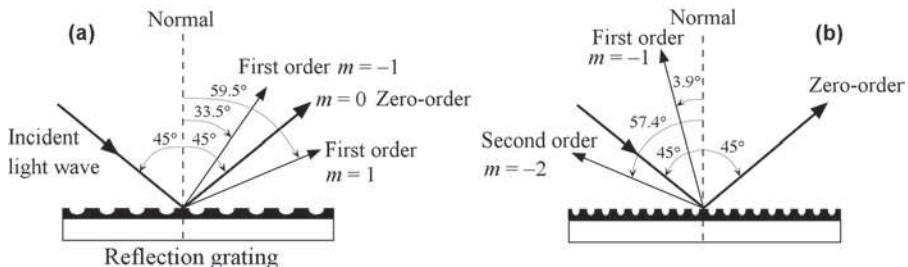


FIGURE 1.42 A light beam is incident at an angle 45° to the normal on a reflection grating. (a) The grating periodicity is $10 \mu\text{m}$. (b) The periodicity is $2 \mu\text{m}$.

and

$$(10 \mu\text{m})(\sin \theta_m - \sin(45^\circ)) = (-1)(1.55 \mu\text{m})$$

Solving these two equations, we find $\theta_m = 59.6^\circ$ for $m = 1$, and $\theta_m = 33.5^\circ$ for $m = -1$.

Consider Figure 1.41 (b) in which the secular reflection from the grooved surface coincides with the m_{th} order diffraction when $2\gamma = \theta_m - \theta_i$. Thus

$$\gamma = (1/2)(\theta_m - \theta_i) = (1/2)(59.6^\circ - 45^\circ) = 7.3^\circ$$

Suppose that we reduce d to $2 \mu\text{m}$. Recalculating the above we find $\theta_m = -3.9^\circ$ for $m = -1$ and imaginary for $m = +1$. Further, for $m = -2$, there is a second-order diffraction beam at -57.4° . Both are shown in Figure 1.42 (b). It is left as an exercise to show that if we increase the angle of incidence, for example, $\theta_i = 85^\circ$ on the first grating, the diffraction angle for $m = -1$ increases from 33.5° to 57.3° and the other diffraction peak ($m = 1$) disappears.

Additional Topics

1.13 INTERFEROMETERS

An **interferometer** is an optical instrument that uses the wave-interference phenomena to produce interference fringes (e.g., dark and bright bands or rings) which can be used to measure the wavelength of light, surface flatness, or small distances. A nearly monochromatic light wave is split into two coherent waves traveling two different paths, and then the two waves are brought together and made to interfere on a screen (or a detector array); the result is an interference pattern as in the Young's fringes in Figure 1.31. In some interferometers, the intensity of the resultant interference is measured at one location, and this intensity is monitored due to changes in one of the optical path lengths. Even a small change in the optical path, distance \times refractive index, nL , can cause a measurable shift in the diffraction pattern or a displacement in the fringes, which can be used to infer on nL . There are many types of interferometers.

The **Fabry–Perot interferometer** is a Fabry–Perot cavity–based interferometer that produces an interference ring pattern (bright and dark rings) when illuminated from a broad monochromatic light source as illustrated in Figure 1.43. The resonator consists of two parallel flat glass plates facing each other and separated by an adjustable spacer. A piezoelectric transducer can provide small changes in the spacing between the plates. The inside surface of the glass plates are coated to enhance reflections within the cavity. (Dielectric mirrors can also be used on the inner surfaces.)

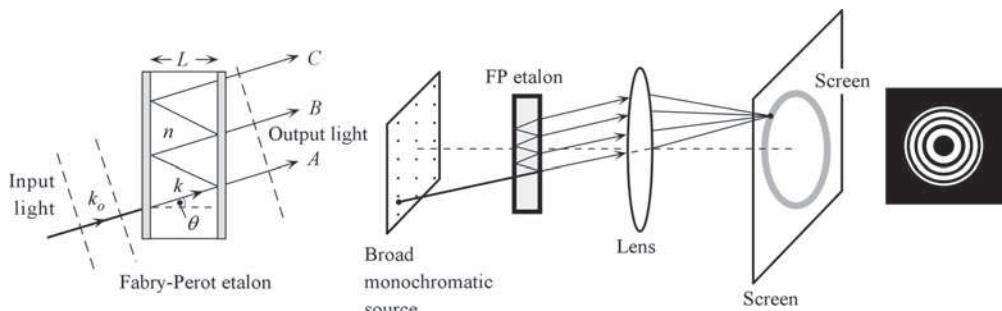


FIGURE 1.43 Fabry-Perot interferometer.

The light entering the cavity will experience multiple reflections inside the cavity. The transmitted waves A , B , C , etc., are focused by a lens onto a point P on a screen. If A , B , C , etc., are in phase then the point P will be a bright spot. Suppose that L is the separation of the plates, θ is the angle at which refracted rays inside the cavity are propagating with respect to the normal, $k = 2\pi n/\lambda$ is the propagation constant in the cavity (λ is the free-space wavelength), and n is the refractive index of the medium inside the cavity. We can then show that the phase difference between A and B is $2kL\cos\theta$, which must be $2m\pi$, where m is an integer, for constructive interference. Recall that for normal incidence this was $2kL$. Thus, P is bright if

$$2nL\cos\theta = m\lambda; \quad m = 0, \pm 1, \dots \quad (1.13.1)$$

All such points with the same θ lie on a circle about the etalon axis. It is apparent that the interference pattern consists of dark and bright rings. The interference ring diameter depends on the wavelength and the optical separation nL (refractive index \times distance) of the plates of the etalon. The interferometer can be used in spectroscopic applications such as the measurement of the source wavelength.

In Figure 1.44, showing a **Mach-Zehnder interferometer**, a coherent light beam is split by a beam splitter into two paths at O . These beams are then reflected by two mirrors M_1 and M_2 onto a second beam splitter where they are combined at C to give rise to an output light beam onto a detector D . The resultant field at C depends on the phase difference between the waves traveling the paths OAC and OCB . We can easily monitor the changes in n or the length d of a sample by placing it into one of the paths, such as in OAC . If $k_o = 2\pi/\lambda$ is the propagation constant in vacuum and $k = 2\pi n/\lambda$ is that in the sample, then the phase difference Φ between the waves arriving at C is³⁹

$$\Phi = k_o(\overline{OAC} - d) + kd - k_o\overline{OCB}$$

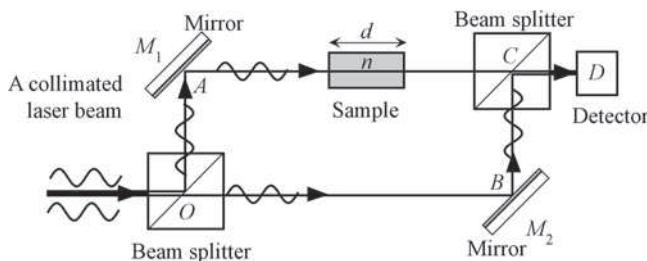


FIGURE 1.44 Mach-Zehnder interferometer.

³⁹ \overline{OAC} stands for the distance from O to A to C .

The change $\Delta\Phi$ in the phase angle due changes in the sample would be

$$\Delta\Phi = \Delta(kd) = \frac{2\pi}{\lambda} \Delta(nd)$$

Thus, the change in the detected signal represents the change in nd . On the other hand, if we can somehow modulate n or d , we have the means of modulating the intensity at the detector. The *Mach-Zehnder modulators* used in communications rely on changing n in a suitable crystal (such as lithium niobate) by the application of an electric field through the electro-optic effect as discussed later in the book.

1.14 THIN FILM OPTICS: MULTIPLE REFLECTIONS IN THIN FILMS

There is much interest in the reflection of light from, and transmission through, a thin film coating of a transparent material (called an optical coating) on a substrate such as a semiconductor crystal; or even on glass as in antireflection coated lenses. The simplest case is shown in Figure 1.45 where light traveling in a medium of refractive index n_1 is incident on a thin film coating of index n_2 on a substrate of index n_3 . There are multiple reflections in the thin film, and the problem is almost identical to the Fabry-Perot optical cavity in which there are similar multiple interferences.

Suppose the thickness of the coating is d . For simplicity, we will assume normal incidence. The phase change in traversing the coating thickness d twice is $\phi = 2 \times (2\pi/\lambda)n_2d$ where λ is the free-space wavelength. The wave has to be multiplied by $\exp(-j\phi)$ to account for this phase difference in crossing the thickness twice. The reflection and transmission coefficients for the present n_1 - n_2 - n_3 system are given by,

$$r_1 = r_{12} = \frac{n_1 - n_2}{n_1 + n_2} = -r_{21}, \quad r_2 = r_{23} = \frac{n_2 - n_3}{n_2 + n_3}, \quad (1.14.1a)$$

and

$$t_1 = t_{12} = \frac{2n_1}{n_1 + n_2}, \quad t'_1 = t_{21} = \frac{2n_2}{n_1 + n_2}, \quad t_2 = t_{23} = \frac{2n_2}{n_2 + n_3} \quad (1.14.1b)$$

where

$$1 - t_1 t'_1 = r_1^2 \quad (1.14.1c)$$

The first reflected beam, $A_1 = A_0 \times r$, the second is $A_2 = A_0 \times t_1 \times t'_1 \times r_2 \times e^{-j\phi}$, and so on.

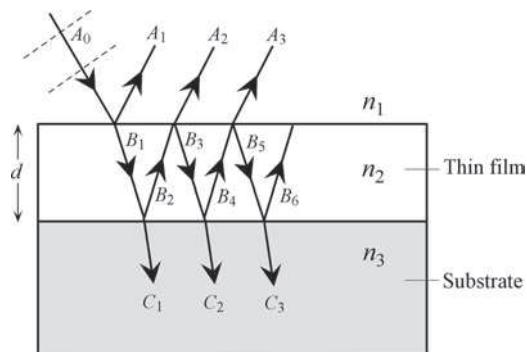


FIGURE 1.45 Light traveling in a medium of refractive index n_1 is incident on a thin film coating of index n_2 on a substrate of index n_3 .

The amplitude of the reflected beam is

$$A_{\text{reflected}} = A_1 + A_2 + A_3 + A_4 + \dots$$

that is,

$$A_{\text{reflected}}/A_0 = r_1 + t_1 t'_1 r_2 e^{-j\phi} - t_1 t'_1 r_1 r_2^2 e^{-j2\phi} + t_1 t'_1 r_1^2 r_2^3 e^{-j3\phi} + \dots \quad (1.14.2)$$

which is a geometric series. Using Eq. (1.14.1c), Eq. (1.14.2) can be conveniently summed to obtain the overall reflection coefficient r

$$r = \frac{r_1 + r_2 e^{-j\phi}}{1 + r_1 r_2 e^{-j\phi}} \quad (1.14.3)$$

Thin film reflection coefficient

Similarly, we can sum for the amplitude of the transmitted beam as

$$C_{\text{transmitted}} = C_1 + C_2 + C_3 + \dots$$

that is,

$$C_{\text{transmitted}}/A_0 = t_1 t_2 e^{-j\phi/2} - t_1 t_2 r_1 r_2 e^{-j3\phi/2} + t_1 t_2 r_1^2 r_2^2 e^{-j5\phi/2} + \dots \quad (1.14.4)$$

which is a geometric series that sums to

$$t = \frac{t_1 t_2 e^{-j\phi/2}}{1 + r_1 r_2 e^{-j\phi}} \quad (1.14.5)$$

Thin film transmission coefficient

Equations (1.14.3) and (1.14.5) describe the reflected and transmitted waves. The reflectance and transmittance are then

$$R = |r|^2 \quad T = (n_3/n_1)|t|^2 \quad (1.14.6)$$

Thin film reflectance and transmittance

Figure 1.46 (a) shows R and T as a function of ϕ for $n_1 < n_2 < n_3$. Clearly, R is minimum, and T is maximum, whenever $\phi = \pi \times (\text{odd number})$ or $\phi = 2(2\pi/\lambda)n_2 d = \pi(2m + 1)$, where $m = 0, 1, 2, \dots$. The latter leads to

$$d = \frac{\lambda}{4n_2} (2m + 1); \quad m = 0, 1, 2, \dots \quad (1.14.7)$$

Thickness for minimum reflection

which is the thickness required to minimize the reflection, and maximize the transmission of light when n_2 is intermediate between n_1 and n_3 . The oscillations in R and T with ϕ (e.g., as the wavelength is scanned) are sometime referred to as an **interference fringes** in wavelength.

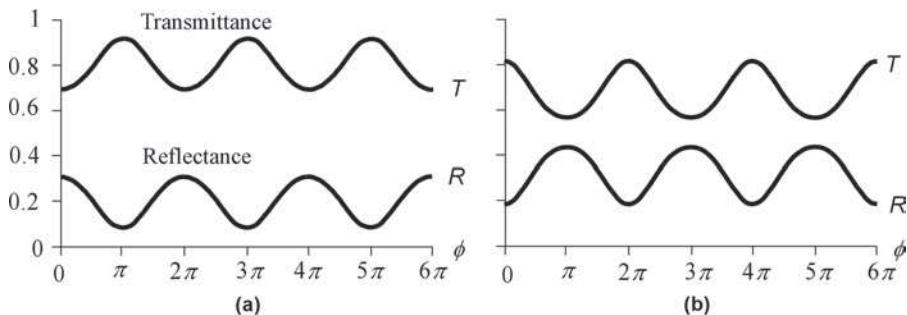


FIGURE 1.46 (a) Reflectance R and transmittance T vs. $\phi = 2n_2d/\lambda$, for a thin film on a substrate where $n_1 = 1$ (air), $n_2 = 2.5$, $n_3 = 3.5$, and $n_1 < n_2 < n_3$. (b) R and T vs. ϕ for a thin film on a substrate where $n_1 = 1$ (air), $n_2 = 3.5$, $n_3 = 2.5$, and $n_2 > n_3 > n_1$.

Figure 1.46 (b) represents the reflectance and transmittance vs. ϕ of light through a thin layer of high index material on a low index substrate where $n_1 < n_3 < n_2$; for example, a semiconductor film on a glass substrate. Notice the difference between the two cases, especially the locations of the maxima and minima. (Why is there a difference?)

The phase change ϕ , of course, depends on three factors, d , n_2 , and λ . The minimum and maximum reflectances in Figure 1.46 at those particular ϕ values can be found by using Eqs. (1.14.3) and (1.14.7). For $n_1 < n_2 < n_3$ as in Figure 1.46 (a)

*Minimum
and
maximum
reflectance*

$$R_{\min} = \left(\frac{n_2^2 - n_1 n_3}{n_2^2 + n_1 n_3} \right)^2; \quad R_{\max} = \left(\frac{n_3 - n_1}{n_3 + n_1} \right)^2 \quad (1.14.8)$$

and the transmittance can be found from $R + T = 1$. When $n_1 < n_3 < n_2$ then R_{\min} and R_{\max} equations are interchanged. While R_{\max} appears to be independent from n_2 , the index n_2 is nonetheless still involved in determining maximum reflection inasmuch as R reaches R_{\max} when $\phi = 2(2\pi/\lambda)n_2d = \pi(2m)$; when $\phi = \pi \times (\text{even number})$.

In transmission spectra measurements, a spectrophotometer is used to record the transmittance of a light beam as a function of wavelength through a sample. If the sample is a thin film on a substrate, there will be multiple reflections and interferences in the thin film, and the measured transmittance will exhibit maxima and minima as in Figure 1.46 as the wavelength (or ϕ) is scanned. The locations of the maxima and minima, with the knowledge of the substrate index (n_3), can be used to find d and n .

EXAMPLE 1.14.1 Thin film optics

Consider a semiconductor device with $n_3 = 3.5$ that has been coated with a transparent optical film (a dielectric film) with $n_2 = 2.5$, $n_1 = 1$ (air). If the film thickness is 160 nm, find the minimum and maximum reflectances and transmittances and their corresponding wavelengths in the visible range. (Assume normal incidence.)

Solution

This case corresponds to Figure 1.46 (a). Minimum reflectance R_{\min} occurs at $\phi = \pi$ or odd multiple of π , and maximum reflectance R_{\max} at $\phi = 2\pi$ or an integer multiple of 2π . From Eq. (1.14.8) we have

$$R_{\min} = \left(\frac{n_2^2 - n_1 n_3}{n_2^2 + n_1 n_3} \right)^2 = \left(\frac{2.5^2 - (1)(3.5)}{2.5^2 + (1)(3.5)} \right)^2 = 0.080 \quad \text{or} \quad 8.0\%$$

and

$$R_{\max} = \left(\frac{n_3 - n_1}{n_3 + n_1} \right)^2 = \left(\frac{3.5 - 1}{3.5 + 1} \right)^2 = 0.31 \quad \text{or} \quad 31\%$$

Corresponding transmittances are,

$$T_{\max} = 1 - R_{\min} = 0.92 \quad \text{or} \quad 92\%$$

and

$$T_{\min} = 1 - R_{\max} = 0.69 \quad \text{or} \quad 69\%$$

Without the thin film coating, the reflectance would be 31%, the maximum reflectance.

Given $\phi = 2dn_2(2\pi/\lambda)$, and $d = 160 \text{ nm}$, and the fact that the minimum reflectance corresponds to $\phi = \text{odd integer} \times \pi$, we choose $\phi = 3\pi$ (by trial and error). The wavelength λ_{\min} is then given by

$$\lambda_{\min} = 4\pi dn_2/\phi = 4\pi(160 \text{ nm})(2.5)/(3\pi) = 533 \text{ nm(green)}$$

which is in the visible. The maximum reflectance in the visible occurs when $\phi = 4\pi$, giving

$$\lambda_{\max} = 4\pi dn_2/\phi = 4\pi(160 \text{ nm})(2.5)/(4\pi) = 400 \text{ nm(violet)}$$

1.15 MULTIPLE REFLECTIONS IN PLATES AND INCOHERENT WAVES

The interference of light waves in a thin film takes place because the waves have much longer coherence lengths than the thickness of the film so that the waves exhibit mutual coherence. We can add the electric field and interference leads to bright and dark fringes. A film would be considered too thick if it does not exhibit any interference phenomena due to the coherence length of the waves being shorter than the thickness. Such cases easily arise when we pass light through a transparent (or partially transparent) plate or when the light source is incoherent. In such cases, we cannot add the electric field to find the reflected and transmitted light irradiances. The overall reflectance and transmittance would be independent of the round-trip phase change ϕ inside the plate and would not exhibit the behavior in Figure 1.46. We have to use Eq. (1.10.5).

Consider a light beam of unit intensity that is passed through a thick plate of transparent material of index n_2 in a medium of index n_1 as in Figure 1.47. The first transmitted light intensity into the plate is $(1 - R)$, and the first transmitted light out is $(1 - R) \times (1 - R)$ or $(1 - R)^2$. However, there are internal reflections as shown, so that the second transmitted light is $(1 - R) \times R \times R \times (1 - R) = R^2(1 - R)^2$, so that the transmitted intensity through the plate is

$$T_{\text{plate}} = (1 - R)^2 + R^2(1 - R)^2 + R^4(1 - R)^2 + \dots = (1 - R)^2[1 + R^2 + R^4 + \dots]$$

or

$$T_{\text{plate}} = \frac{(1 - R)^2}{1 - R^2} \quad (1.15.1)$$

Transmittance
for a thick
plate or
incoherent
light

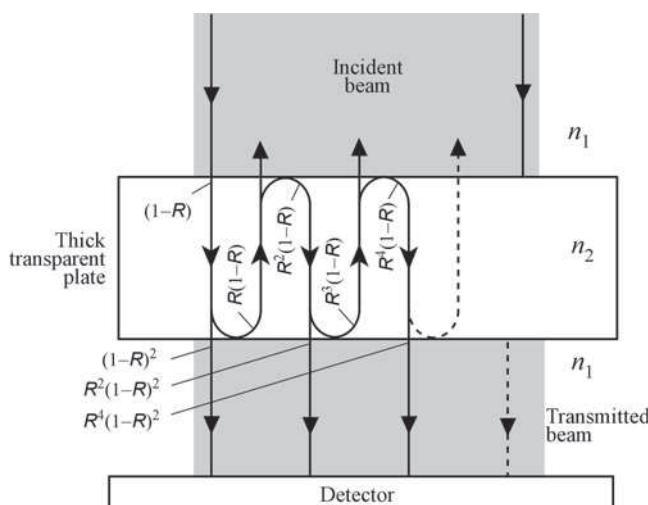


FIGURE 1.47 Transmitted and reflected light through a slab of material in which there is no interference.

Transmittance for a thick plate for incoherent light

Transmittance for a thick plate or incoherent beam of light

By substituting for R in terms of the indices, we can write this as

$$T_{\text{plate}} = \frac{2n_1 n_2}{n_1^2 + n_2^2} \quad (1.15.2)$$

For example, for a glass plate of $n_2 = 1.60$ in air ($n_1 = 1$), $T_{\text{plate}} = 89.9\%$ while the simple transmittance through an n_1-n_2 interface would give 94.7%. The overall reflectance is $1 - T_{\text{plate}}$ so that

$$R_{\text{plate}} = \frac{(n_1 - n_2)^2}{n_1^2 + n_2^2} \quad (1.15.3)$$

One of the simplest ways to determine the refractive index of a plate is to measure the transmittance T_{plate} in Eq. (1.15.1), from which we can calculate n_2 .

1.16 SCATTERING OF LIGHT

When a light beam propagates in a medium in which there are small particles or inhomogeneities, such as local changes in the refractive index of the medium, some of the power in the beam is radiated away from the direction of propagation, that is some of the power becomes scattered. **Scattering** is a process by which some of the power in a propagating electromagnetic wave is redirected as secondary EM waves in various directions away from the original direction of propagation as illustrated in Figure 1.48 (a). There are a number of scattering processes, which are usually classified in terms of the size of the scattering particles in relation to the wavelength of light that is scattered. In **Rayleigh scattering**, the scattering particle size, or the scale of inhomogeneities in a medium, is much smaller than the wavelength of light. The intensity of the scattered light at an angle θ to the original beam depends on the scattering process; the Rayleigh scattering case is shown in Figure 1.48 (b) in which the scattering is not spherically symmetric.

Consider what happens when a propagating wave encounters a molecule, an impurity in a crystal or a small dielectric particle (or region), which is smaller than the wavelength of light. The electric field in the wave polarizes the particle by displacing the lighter electrons with respect to the heavier positive nuclei. The electrons in the molecule couple and oscillate with the electric field in the wave (ac electronic polarization). The oscillation of charge “up” and “down,” or the oscillation of the induced dipole, radiates EM waves all around the molecule as illustrated in Figure 1.48 (a). We should remember that an oscillating charge is like an alternating current which always radiates EM waves (like an antenna). The net effect is that the incident

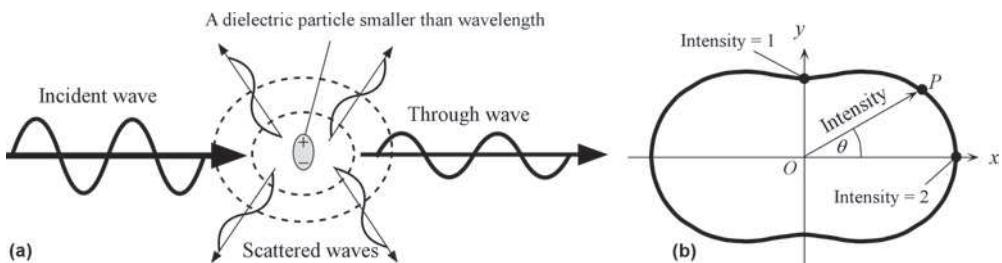


FIGURE 1.48 (a) Rayleigh scattering involves the polarization of a small dielectric particle or a region that is much smaller than the light wavelength. The field forces dipole oscillations in the particle (by polarizing it) which leads to the emission of EM waves in “many” directions so that a portion of the light energy is directed away from the incident beam. (b) A polar plot of the dependence of the intensity of the scattered light on the angular direction θ with respect to the direction of propagation x in Rayleigh scattering (in a polar plot, the radial distance OP is the intensity).



Lord Rayleigh (John William Strutt) was an English physicist (1877–1919) and a Nobel Laureate (1904) who made a number of contributions to wave physics of sound and optics. He formulated the theory of scattering of light by small particles and the dependence of scattering on $1/\lambda^4$ circa 1871. Then, in a paper in 1899 he provided a clear explanation on why the sky is blue. Ludvig Lorentz, around the same time, and independently, also formulated the scattering of waves from a small dielectric particle, though it was published in Danish (1890).⁴⁰ (© Mary Evans Picture Library/Alamy.)

wave becomes partially reradiated in different directions and hence loses intensity in its original direction of propagation. (We may think of the process as the particle absorbing some of the energy via electronic polarization and reradiating it in different directions.) It may be thought that the scattered waves constitute a spherical wave emanating from the scattering molecule, but this is not generally the case as the re-emitted radiation depends on the shape and polarizability of the molecule in different directions. We assumed a small particle so that at any time the field has no spatial variation through the particle, whose polarization then oscillates with the electric field oscillation. Whenever the size of the scattering region, whether an inhomogeneity, a small particle, a molecule, or a defect in a crystal, is much smaller than the wavelength λ of the incident wave, the scattering process is generally termed **Rayleigh scattering**. Typically, the particle size is smaller than one-tenth of the light wavelength.

Rayleigh scattering of light propagating in a glass medium is of particular interest in photonics because it results in the attenuation of the transmitted light pulses in optical fibers. The glass structure is such that there are small random spatial variations in the refractive index about some average value. There are therefore local fluctuations in the relative permittivity and polarizability, which effectively act if there are small inhomogeneities in the medium. These dielectric inhomogeneities arise from fluctuations in the relative permittivity that is part of the intrinsic glass structure. As the fiber is drawn by freezing a liquid-like flow, random thermodynamic fluctuations in the composition and structure that occur in the liquid state become frozen into the solid structure. Consequently, the glass fiber has small fluctuations in the relative permittivity which leads to Rayleigh scattering. A small inhomogeneous region acts like a small dielectric particle and scatters the propagating wave in different directions. Nothing can be done to eliminate Rayleigh scattering in glasses as it is part of their intrinsic structure.

It is apparent that the scattering process involves electronic polarization of the molecule or the dielectric particle. We know that this process couples most of the energy at ultraviolet frequencies where the dielectric loss due to electronic polarization is maximum and the loss is due to EM wave radiation. Therefore, as the frequency of light increases, the scattering becomes more severe. In other words, *scattering decreases with increasing wavelength*. The intensity of the scattered radiation is proportional to $1/\lambda^4$. For example, blue light, which has a shorter wavelength than red light, is scattered more strongly by air molecules. When we look at the sun directly, it appears yellow because the blue light has been scattered in the direct light more than

⁴⁰ Pedro Lilienfeld's "A Blue Sky History", in *Optics and Photonics News*, 15(6), 32, 2004, is highly recommended; it also provides the original references.

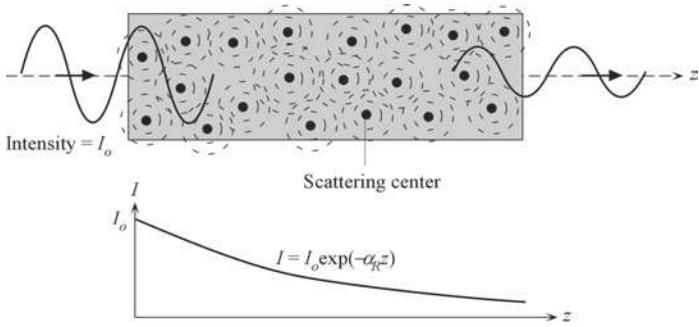


FIGURE 1.49 When a light beam propagates through a medium in which there are small particles, it becomes scattered as it propagates and loses power in the direction of propagation. The light becomes attenuated.

the red light. When we look at the sky in any direction but the sun, our eyes receive scattered light which appears blue; hence the sky is blue. At sunrise and sunset, the rays from the sun have to traverse the longest distance through the atmosphere and have the most blue light scattered, which gives the sun its red color at these times.

The intensity of a light beam in a medium with small particles (or inhomogeneities) decreases as the beam propagates due to Rayleigh scattering from these small particles as illustrated in Figure 1.49. The intensity at a position z inside the medium from the radiation receiving face is given by

Attenuation
by scattering

$$I = I_o \exp(-\alpha_R z) \quad (1.16.1)$$

where α_R is the attenuation coefficient due to Rayleigh scattering. α_R depends on the concentration of scattering particles N , their radius a , wavelength λ and the mismatch between the refractive index n of the scattering spheres and the index n_o of the medium

Rayleigh
attenuation
coefficient

$$\alpha_R \propto N \cdot a^6 \cdot \frac{1}{\lambda^4} \cdot \left(\frac{n^2 - n_o^2}{n^2 + n_o^2} \right)^2 \quad (1.16.2)$$

Equation (1.16.2) has been written in terms of products to identify those factors that affect α_R .

Mie scattering refers to the scattering of light from scatterers that have dimensions comparable with, or greater than, the wavelength of light. For example, Mie scattering would occur for light scattering from long organic molecules in a solution, or scattering from various particulate pollutants (as in smog) in the atmosphere, including dust particles. The scatterers are assumed to have a refractive index significantly different from that of the surrounding medium. The scattering depends on the ratio of the scattering particle diameter to the wavelength of light, and favors scattering in the forward direction. The dependence on the wavelength is weaker than Rayleigh scattering.

The scattering of light from various particles suspended in a liquid is usually referred to as **turbidity**. For example, small solid particles suspended in water will scatter light and would make the water appear cloudy. The quality of water is often monitored for small particles by measuring the turbidity of water.

1.17 PHOTONIC CRYSTALS

The dielectric mirror, or the Bragg reflector, in Figure 1.24 is a stack of alternating layers of high and low refractive index material in a periodic manner. The one-dimensional periodic variation in the refractive index n in Figure 1.24 represents one of the simplest periodic structures within

a general class of optical materials called **photonic crystals**. A **photonic crystal** (PC) is a material that has been structured to possess a periodic modulation of the refractive index n , just like in Figure 1.24, so that the structure influences the propagation and confinement of light within it. The periodicity can be in one- (1D), two- (2D), or three-dimensional (3D); Figures 1.50 (a)–(c) illustrates 1D, 2D, and 3D simple photonic crystals as examples. In fact, quite complicated structures can be constructed that have very interesting optical properties. The dielectric mirror or the Bragg reflector can be viewed as one of the simplest 1D photonic crystals.⁴¹ For the structure to influence the propagation of the EM wave, it has to diffract the wave, which means that the scale of periodic variations must be on the wavelength scale. As apparent from Figure 1.24, the 1D PC has a band of frequencies over which it reflects the light and, conversely, there is a **stop band** over which no transmission is possible through the dielectric stack. There is a band of frequencies that represent waves that are not allowed to go through this periodic structure in the direction of refractive index variation, along z in Figure 1.50; this band is called an **optical or photonic bandgap**.

The periodic variation in n in Figure 1.50 is normally assumed to extend indefinitely, whereas in practice, the PCs have a finite size, for example, a certain number of layers in the dielectric mirror, not infinite. As in normal crystals, the periodic structures in Figure 1.50 have a **unit cell**, which repeats itself to generate the whole lattice—that is, the whole crystal structure. For the 1D PC in Figure 1.50 (a), for example, two adjacent layers, n_1n_2 , form the unit cell. We can move this unit cell along z by a distance Λ , the **period** (or **periodicity**), many times to generate the whole 1D photonic crystal.

The periodicity of a photonic crystal implies that any property at a location z will be the same at $z \pm \Lambda$, $z \pm 2\Lambda$ and so on; that is, there is *translational symmetry* along z (in 1D).

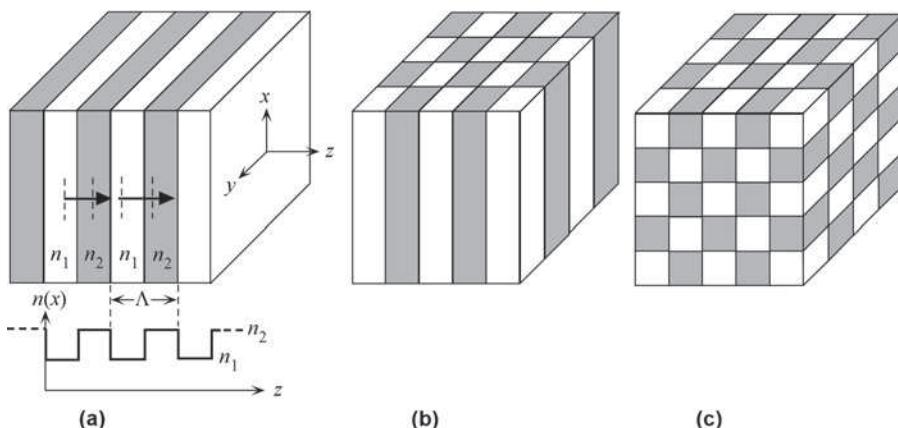


FIGURE 1.50 Photonic crystals in (a) 1D, (b) 2D, and (c) 3D, D being the dimension. Gray and white regions have different refractive indices and may not necessarily be the same size. Λ is the periodicity. The 1D photonic crystal in (a) is the well-known Bragg reflector, a dielectric stack.

⁴¹The propagation of light through such a one-dimensional (1D) periodic variation of n has been well-known, dating back to the early work of Lord Rayleigh in 1887. Eli Yablonovitch has suggested that the name “photonic crystal” should really only apply to 2D and 3D periodic structures with a large dielectric (refractive index) difference. (E. Yablonovitch, “Photonic crystals: What’s in a name?,” *Opt. Photon. News*, 18, 12, 2007.) Nonetheless, the dielectric mirror in Figure 1.50 (a) is often considered as the simplest 1D photonic crystal to derive the concept of a “photonic bandgap,” essentially a stop band.

The EM waves that are allowed to propagate along z through the periodic structure are called the **modes** of the photonic crystal. They have a special waveform that must bear the periodicity of the structure, and are called **Bloch waves**. Such a wave for the field E_x , for example, has the form $E_x(z, t) = A(z)\exp(-jkz)$, which represents a traveling wave along z and $A(z)$ is an amplitude function that has the periodicity of the structure, that is, it is periodic along z with a period Λ . $A(z)$ depends on the periodic refractive index function $n(z)$. As we will see in Chapter 3, the electron motion in a semiconductor crystal is also described by Bloch waves (electron wavefunctions).

As in the case of the dielectric mirror, the 1D PC has a band of frequencies over which there can be no propagation along z . In a homogeneous medium of refractive index n , the relationship between the frequency ω and the propagation constant k is simple, that is, $c/n = \omega/k$, where k is the propagation constant inside the medium. The dispersion behavior of the medium, that is, ω vs. k , is a straight line with a slope c/n . The dispersion characteristic of a 1D PC for waves along z in Figure 1.50 (a) is shown in Figure 1.51 (a). We notice several important characteristics. At low frequencies (long wavelengths), the waves propagate as if they are in a homogenous medium with a constant phase velocity (dashed straight line). As expected, there is a band of frequencies $\Delta\omega = \omega_2 - \omega_1$ (between S_1 and S_2) over which no propagation along z is allowed, which is a **photonic bandgap** along z . The whole ω vs. k curves are periodic in k with a period $2\pi/\Lambda$. The point P is equivalent to P' because $k = k' + (2\pi/\Lambda)$. We only need to consider k -values from $-\pi/L$ to π/L . This region is called the **first Brillouin zone**.

At low frequencies or long wavelengths (small k values) in Figure 1.51 (a), the wavelength is so much longer than the variations in n that the propagating wave experiences essentially some average refractive index, n_{av} , that is $(n_1 + n_2)/2$ if the n_1 and n_2 layers have the same thickness, and propagates through the structure as if the structure was a homogenous medium with some effective refractive index, n_{av} . Its phase velocity is ω/k , which is c/n_{av} , and its group velocity, the slope of the ω vs. k characteristic, is the same as c/n_{av} , in this region (ignoring the wavelength dependence of n_1 and n_2). As the wavelength decreases (ω increases), partial reflections of the waves from the boundaries become important and interfere with propagation. At sufficiently small wavelengths, diffraction becomes important as all these partially reflected waves interfere with each other and give rise to significant reflection. Eventually, a critical wavelength (corresponding to ω_1) is reached where the waves become fully diffracted or reflected backwards. A backward traveling (in $-z$ direction) wave experiences the same reflection. These forward and backward diffracted waves give rise to a standing wave in the structure; indeed, only the latter can exist and waves cannot propagate, that is, travel freely.

Eli Yablonovitch at the University of California at Berkeley, and Sajeev John (shown later in this chapter) at the University of Toronto, carried out the initial pioneering work on photonic crystals. Eli Yablonovitch has suggested that the name “photonic crystal” should apply to 2D and 3D periodic structures with a large dielectric (refractive index) difference. (E. Yablonovitch, “Photonic crystals: What’s in a name?,” *Opt. Photon. News*, 18, 12, 2007.) Their original papers were published in the same volume of *Physical Review Letters* in 1987. According to Eli Yablonovitch, “Photonic Crystals are semiconductors for light.” (Courtesy of Eli Yablonovitch.)



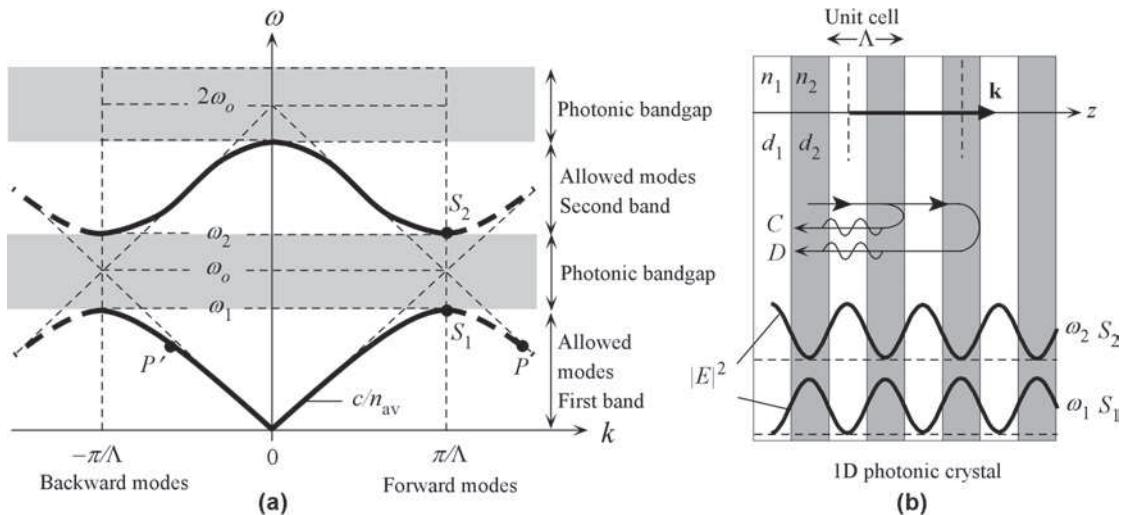


FIGURE 1.51 (a) Dispersion relation, ω vs. k , for waves in a 1D PC along the z -axis. There are allowed modes and forbidden modes. Forbidden modes occur in a band of frequencies called a photonic bandgap. (b) The 1D photonic crystal corresponding to (a), and the corresponding points S_1 and S_2 with their stationary wave profiles at ω_1 and ω_2 .

Consider what happens when reflections such as C and D from two successive unit cells interfere constructively and give rise to a backward diffracted (reflected) wave as illustrated in Figure 1.51 (b). Reflections from two successive unit cells must be in phase for full reflection. This means that the phase difference between C and D ($2\pi/\lambda(2n_1d_1 + 2n_2d_2)$) must be $2m\pi$, where $m = 1, 2, \dots$ is an integer. If we define $n_{av} = (n_1d_1 + n_2d_2)/\Lambda$, and $k = n_{av}(2\pi/\lambda)$ then reflection occurs when $2k\Lambda = 2m\pi$, that is, $k = m\pi/\Lambda$. These are the wave vectors or propagation constants of the waves that cannot be propagated. The waves suffer Bragg reflection in 1D. What happens to these diffracted waves?

The diffracted waves in $+z$ and $-z$ directions set up a standing wave in the structure. If we write the two reflected waves in opposite directions as $A \exp(jkz)$ and $A \exp(-jkz)$ they would add as $A \exp(jkz) \pm A \exp(-jkz)$, which shows that there are two possibilities S_1 and S_2 . One of them, S_1 , has most of its energy inside the n_2 , high refractive index layers, and hence has a lower frequency ω_1 . S_2 has most of its energy in the n_1 layers and has a higher frequency ω_2 ; these are shown in Figure 1.51 (b). There are no waves in the $\omega_2 - \omega_1$ interval, which we know as the **photonic bandgap** (PBG). The bandgap $\Delta\omega = \omega_2 - \omega_1$ increases linearly with the index difference $\Delta n = n_2 - n_1$.

It is clear that the 1D crystal exhibits a photonic bandgap for light propagation along z . It should also be apparent that there would be no PBGs for propagation along the x and y directions along which there are no periodic n -variations. Since we can resolve any \mathbf{k} -vector along x , y , and z directions, overall there is no *net* PBG for light propagation in a 1D PC. The PBG for a 1D crystal is called a **pseudo PBG**.

The above ideas can be readily extended to 2D and 3D periodic structures. Again, in principle, there is no full PBG in the 2D PC. In the case of 3D periodic structures, there would be photonic bandgaps along x , y , and z directions, and for different polarizations of the electric field. If the refractive index contrast and the periodicity in the 3D structure are such that these photonic bandgaps overlap in *all directions* and for *all polarizations* of light, as schematically depicted in Figure 1.52, the overlap frequency range $\Delta\omega$ becomes a **full photonic bandgap** in all directions for all polarizations of light—no light can propagate through this structure over this

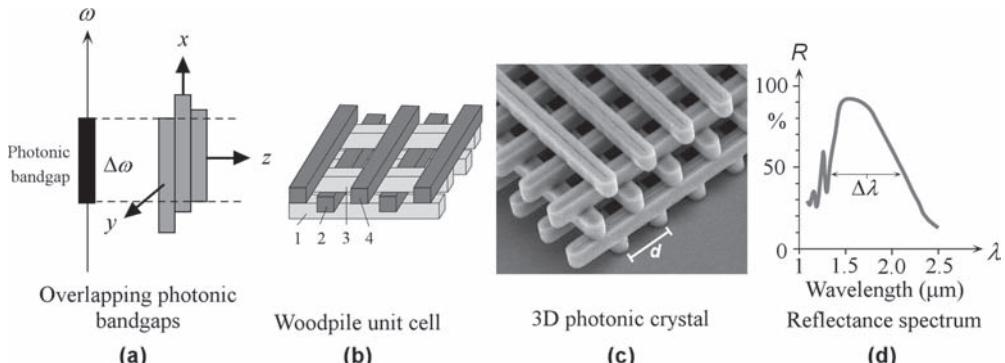
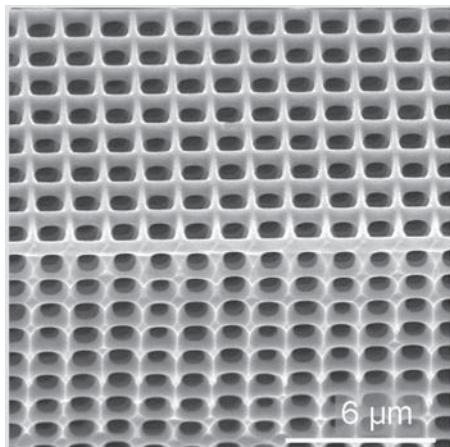
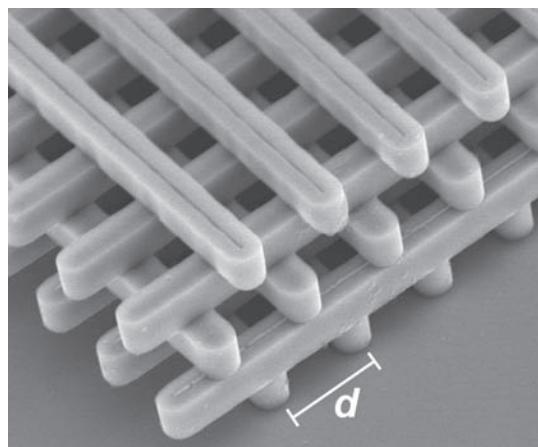


FIGURE 1.52 (a) The photonic bandgaps along x , y , and z overlap for all polarizations of the field, which results in a full photonic bandgap $\Delta\omega$ (an intuitive illustration). (b) The unit cell of a *woodpile* photonic crystal. There are 4 layers, labeled 1–4 in the figure, with each later having parallel “rods.” The layers are at right angles to each other. Notice that layer 3 is shifted with respect to 1, and 4 with respect to 2. (c) An SEM image of a woodpile photonic crystal based on polycrystalline Si; the rod-to-rod pitch d is on the micron scale. (*Courtesy of Sandia National Laboratories.*) (d) The optical reflectance of a woodpile photonic crystal showing a photonic bandgap between 1.5 μm and 2 μm. The photonic crystal is similar to that in (c) with five layers and $d \approx 0.65$ μm. (*Source:* The reflectance spectrum was plotted using the data appearing in Fig. 3 in S-Y. Lin and J.G. Fleming, *J. Light Wave Technol.*, 17, 1944, 1999.)

frequency range, $\Delta\omega$. The structure then has a **full photonic bandgap**. As it turns out not *any* 3D periodic structure results in a full photonic bandgap. Only certain 3D periodic structures allow full photonic bandgaps to develop. One such structure is the *wood pile* periodic structure shown in Figure 1.52 (b). The unit cell has four layers of rods. The rods are parallel in each layer and the layers are



An SEM image of a 3D photonic crystal made from porous silicon in which the lattice structure is close to being simple cubic. The silicon squares, the unit cells, are connected at the edges to produce a cubic lattice. This 3D PC has a photonic bandgap centered at 5 μm and about 1.9 μm wide. (*Courtesy of Max-Planck Institute for Microstructure Physics.*)



An SEM image of a 3D photonic crystal that is based on the *wood pile* structure. The rods are polycrystalline silicon. Although five layers are shown, the unit cell has four layers, *e.g.*, the four layers starting from the bottom layer. Typical dimensions are in microns. In one similar structure with rod-to-rod pitch $d = 0.65$ μm with only a few layers, the Sandia researchers were able to produce a photonic bandgap $\Delta\lambda$ of 0.8 μm centered around 1.6 μm within the telecommunications band. (*Courtesy of Sandia National Laboratories.*)

at 90° to each other. An SEM image of a woodpile photonic crystal is shown in Figure 1.52 (c). The optical reflectance, that is, 1 – transmittance, of this woodpile PC is shown in Figure 1.52 (d). It is apparent that there is a photonic bandgap, a stop band, over a range of frequencies, or wavelengths, $\Delta\lambda$, around 3 μm . The width and the location of the reflectance or the transmittance band depends on the structure of the photonic crystal, that is, the periodicity, unit cell structure and refractive index contrast. The colors of certain butterflies and insects arise not from pigments or colorants but from a photonic crystal effect with the right periodicity.

An important practical aspect of PCs, analogous to semiconductor crystals, is the importance of defects. Point and line defects that occur in normal crystals also occur in PCs, as illustrated in Figure 1.53. They are intentionally introduced to endow the PC structure with certain optical properties. A defect is a discontinuity in the periodicity of the PC. For example, if we upset the periodicity by removing a unit cell, we create a so-called **point defect**. This void can act as an *optical cavity*, trapping the EM radiation within the cavity as illustrated in Figure 1.53. We can, of course, remove a group of unit cells, or modify the refractive index over a few unit cells, which would create an optical microcavity. Defects introduce localized electromagnetic modes with frequencies within the photonic bandgap. Defects can tightly confine a mode of light in a very small cavity volume. We can also enhance the refractive index locally, which would also classify as a point defect.

Line defects are formed when a long row of unit cells are missing, or the refractive index stays constant over a long line in the crystal. Such a line in which the index is constant allows propagating EM modes within the photonic bandgap. Since EM waves can propagate within and along the line defect, the line defect acts as an **optical wave guide**, guiding the radiation along its length as shown in Figure 1.53. The EM wave cannot spread into the perfect photonic crystal region since, in this region, the frequency falls into the stop band.

One very important property of photonic crystals is their ability to *suppress* or *inhibit spontaneous emission*. We can understand this effect intuitively by considering a 3D PC made from a semiconductor. In spontaneous emission, an electron falls from the conduction band

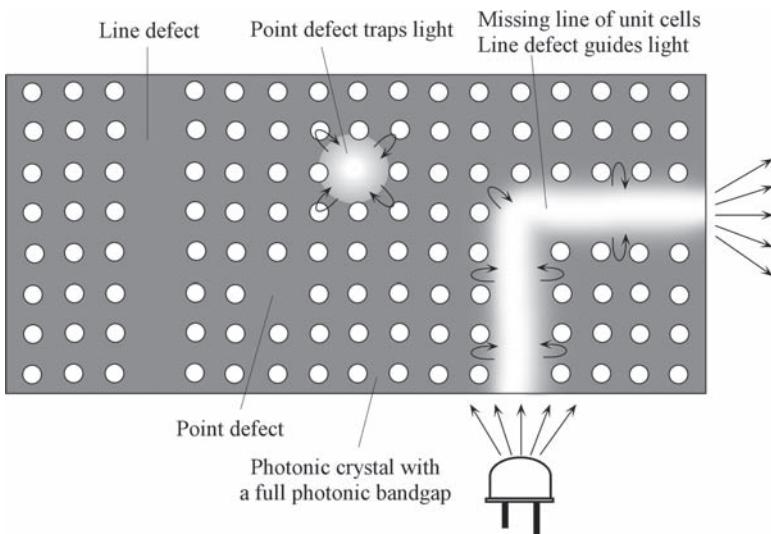


FIGURE 1.53 Schematic illustration of point and line defects in a photonic crystal. A point defect acts as an optical cavity, trapping the radiation. Line defects allow the light to propagate along the defect line. The light is prevented from dispersing into the bulk of the crystal since the structure has a full photonic bandgap. The frequency of the propagating light is in the bandgap, that is, in the stop band.

Sajeev John, at the University of Toronto, along with Eli Yablonovitch (shown earlier in the chapter) carried out the initial pioneering work in the development of the field of photonics crystals. Sajeev John was able to show that it is possible to trap light in a similar way the electron is captured, that is localized, by a trap in a semiconductor. Defects in photonic crystals can confine or localize electromagnetic waves; such effects have important applications in quantum computing and integrated photonics. (*Courtesy of Sajeev John.*)



to the valence band spontaneously and emits a photon of energy $\hbar\omega$ that corresponds to the bandgap energy E_g . However, if the photon frequency $\hbar\omega$ falls into the bandgap of the PC, then this photon is not allowed to propagate or “exist” in the structure. It is prevented from being emitted—the photons have no place to go.

Photonic crystals have also been shown to exhibit a so-called *superprism effect*. Under appropriate conditions, the dispersion of light by a prism-shaped photonic crystal is considerably enhanced over that corresponding to a homogeneous prism having the same average refractive index as the photonic crystal. This originates from the strong curvature of the $\omega-k$ curve near the edge of the Brillouin Zone as apparent in Figure 1.51 (a). The latter may also be viewed as a high refractivity variation with wavelength.

Questions and Problems

1.1 Maxwell's wave equation and plane waves

- (a) Consider a traveling sinusoidal wave of the form $E_x = E_o \cos(\omega t - kz + \phi_o)$. The latter can also be written as $E_x = E_o \cos[k(vt - z) + \phi_o]$, where $v = \omega/k$ is the velocity. Show that this wave satisfies Maxwell's wave equation, and show that $v = 1/(\mu_o \epsilon_o \epsilon_r)^{1/2}$.
- (b) Consider a traveling function of any shape, even a very short delta pulse, of the form $E_x = f[k(vt - z)]$, where f is any function, which can be written as $E_x = f(\phi)$, $\phi = k(vt - z)$. Show that this traveling function satisfies Maxwell's wave equation. What is its velocity? What determines the form of the function f ?

1.2 Propagation in a medium of finite small conductivity

An electromagnetic wave in an isotropic medium with a dielectric constant ϵ_r and a finite conductivity σ and traveling along z obeys the following equation for the variation of the electric field E perpendicular to z

$$\frac{d^2E}{dz^2} - \epsilon_o \epsilon_r \mu_o \frac{\partial^2 E}{\partial t^2} = \mu_o \sigma \frac{\partial E}{\partial t} \quad (\text{P1.1})$$

Show that one possible solution is a plane wave whose amplitude decays exponentially with propagation along z , that is, $E = E_o \exp(-\alpha' z) \exp[j(\omega t - kz)]$. Here $\exp(-\alpha' z)$ causes the envelope of the amplitude to decay with z (attenuation) and $\exp[j(\omega t - kz)]$ is the traveling wave portion. Show that in a medium in which α is small, the wave velocity and the attenuation coefficient of the field are given by

$$v = \frac{\omega}{k} = \frac{1}{\sqrt{\mu_o \epsilon_o \epsilon_r}} \quad \text{and} \quad \alpha' = \frac{\sigma}{2\epsilon_o c n}$$

where n is the refractive index ($n = \epsilon_r^{1/2}$). What is the attenuation coefficient α that describes the decay of the light intensity? (Metals with high conductivities are excluded.)

- 1.3 Point light source** What is the irradiance measured at a distance of 1 m and 2 m from a 1 W light point source?

- 1.4 Gaussian beam** Estimate the divergence and Rayleigh range of a Gaussian beam from a He-Ne Laser with $\lambda = 633$ nm and a beam width of 1.00 nm at $z = 0$. After traversing 10 m through vacuum, what will the beam width be?

- 1.5 Gaussian beam in a cavity with spherical mirrors** Consider an optical cavity formed by two aligned spherical mirrors facing each other as shown in Figure 1.54. Such an optical cavity is called a *spherical mirror*

resonator, and is most commonly used in gas lasers. Sometimes, one of the reflectors is a plane mirror. The two spherical mirrors and the space between them form an optical resonator because only certain light waves with certain frequencies can exist in this optical cavity. The radiation inside a spherical mirror cavity is a *Gaussian beam*. The actual or particular Gaussian beam that fits into the cavity is that beam whose wavefronts at the mirrors match the curvature of the mirrors. Consider the symmetric resonator shown in Figure 1.54 in which the mirrors have the same radius of curvature R . When a wave starts at A , its wavefront is the same as the curvature of A . In the middle of the cavity it has the minimum width and at B the wave again has the same curvature as B . Such a wave in the cavity can replicate itself (and hence exist in the cavity) as it travels between the mirrors provided that it has right beam characteristics, that is the right curvature at the mirrors. The radius of curvature R of a Gaussian beam wavefront at a distance z along its axis is given by

$$R(z) = z[1 + (z_o/z)^2]; \quad z_o = \pi w_o^2/\lambda \text{ is the Rayleigh range}$$

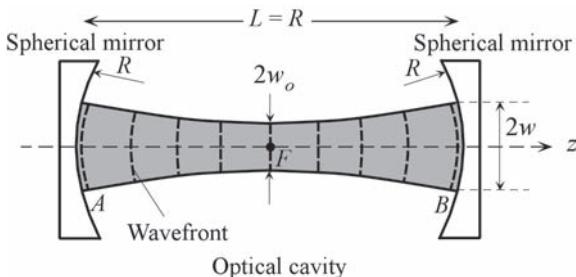


FIGURE 1.54 Two spherical mirrors reflect waves to and from each other. The optical cavity contains a Gaussian beam. This particular optical cavity is symmetric and confocal; the two focal points coincide at F .

Consider a confocal symmetric optical cavity in which the mirrors are separated by $L = R$.

- (a) Show that the cavity length L is $2z_o$, that is, it is the same as twice the Rayleigh range, which is the reason the latter is called the **confocal length**.
- (b) Show that the waist of the beam $2w_o$ is fully determined only by the radius of curvature R of the mirrors, and given by

$$2w_o = (2\lambda R/\pi)^{1/2}$$

- (c) If the cavity length $L = R = 50$ cm, and $\lambda = 633$ nm, what is the waist of the beam at the center and also at the mirrors?

1.6 Cauchy dispersion equation Using the Cauchy coefficients and the general Cauchy equation, calculate refractive index of a silicon crystal at wavelengths of 200 μm and at 2 μm , over two orders of magnitude wavelength change. What is your conclusion?

1.7 Sellmeier dispersion equation Using the Sellmeier equation and the coefficients, obtain a graph of the refractive index of fused silica (SiO_2) versus its wavelength in the range of 500 nm to 1550 nm.

1.8 Sellmeier dispersion equation The Sellmeier dispersion coefficient for pure silica (SiO_2) and 86.5% SiO_2 -13.5% GeO_2 are given in Table 1.2. Write a program on your computer or calculator, or use a math software package or even a spreadsheet program (e.g., Excel) to obtain the refractive index n as a function of λ from 0.5 μm to 1.8 μm for both pure silica and 86.5% SiO_2 -13.5% GeO_2 . Obtain the group index, N_g , vs. wavelength for both materials and plot it on the same graph. Find the wavelength at which the material dispersion, defined as the derivative of the group velocity with respect to the wavelength, becomes zero in each material.

1.9 The Cauchy dispersion relation for zinc selenide ZnSe is a II-VI semiconductor and a very useful optical material used in various applications such as optical windows (especially high power laser windows), lenses, and prisms. It transmits over 0.50–19 μm . n in the 1–11 μm range is described by a Cauchy expression of the form

$$n = 2.4365 + \frac{0.0485}{\lambda^2} + \frac{0.0061}{\lambda^4} - 0.0003\lambda^2$$

in which λ is in μm . What are the n_{-2} , n_0 , n_2 and n_4 coefficients? What is ZnSe's refractive index n and group index N_g at 5 μm ?

1.10 Refractive index, reflection, and the Brewster's angle

- Consider light of free-space wavelength 1300 nm traveling in pure silica medium. Calculate the phase velocity and group velocity of light in this medium. Is the group velocity ever greater than the phase velocity?
- What is the Brewster angle (the polarization angle θ_p) and the critical angle (θ_c) for total internal reflection when the light wave traveling in this silica medium is incident on a silica-air interface. What happens at the polarization angle?
- What is the reflection coefficient and reflectance at normal incidence when the light beam traveling in the silica medium is incident on a silica-air interface?
- What is the reflection coefficient and reflectance at normal incidence when a light beam traveling in air is incident on an air-silica interface? How do these compare with part (c) and what is your conclusion?

1.11 Snell's law and lateral beam displacement What is the lateral displacement when a laser beam passes through two glass plates, each of thickness 1 mm, with refractive indices of $n_1 = 1.570$ and $n_2 = 1.450$ respectively, if the angle of incidence is 45° ?

1.12 Snell's law and lateral beam displacement An engineer wants to design a refractometer (an instrument for measuring the refractive index) using the lateral displacement of light through a glass plate. His initial experiments involve using a plate of thickness L , and measuring the displacement of a laser beam when the angle of incidence θ_i is changed, for example, by rotating (tilting) the sample. For $\theta_i = 40^\circ$ he measures a displacement of 0.60 mm, and when $\theta_i = 80^\circ$ he measures 1.69 mm. Find the refractive index of the plate and its thickness. (Note: You need to solve a nonlinear equation for n numerically.)

1.13 Snell's law and prisms Consider the prism shown in Figure 1.55 that has an apex angle $\alpha = 60^\circ$. The prism has a refractive index of n and it is in air.

- What are Snell's law at interfaces at A (incidence and transmittance angles of θ_i and θ_t) and B (incidence and transmittance angles of θ'_i and θ'_t)?
- Total deflection $\delta = \delta_1 + \delta_2$ where $\delta_1 = \theta_t - \theta_i$ and $\delta_2 = \theta'_t - \theta'_i$. Now, $\beta + \theta'_i + \theta_t = 180^\circ$ and $\alpha + \beta = 180^\circ$. Find the deflection of the beam for an incidence angle of 45° for the following three colors at which n is known: Blue, $n = 1.4634$ at $\lambda = 486.1$ nm; yellow, $n = 1.4587$ at $\lambda = 589.2$ nm; red, $n = 1.4567$ at $\lambda = 656.3$ nm. What is the separation in distance between the rays if the rays are projected on a screen 1 m away.

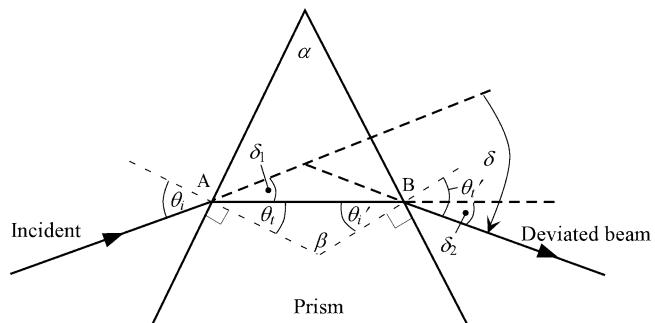


FIGURE 1.55 A light beam is deflected by a prism through an angle δ . The angle of incidence is θ_i . The apex angle of the prism is α .

1.14 Fermat's principle of least time Fermat's principle of least time in simple terms states that *when light travels from one point to another it takes a path that has the shortest time*. In going from a point A in some medium with a refractive index n_1 to a point B in a neighboring medium with refractive index n_2 as in Figure 1.56, the light path AOB involves refraction at O that satisfies Snell's law. The time it takes to travel from A to B is minimum only for the path AOB such that the incidence and refraction angles θ_i and θ_t satisfy Snell's law. Let's draw a straight line from A to B cutting the x -axes at O' . The line $AO'B$ will be our reference line and we will place the origin of x and y coordinates at O' . Without invoking Snell's law, we will vary point O along the x -axis (hence OO' is a variable labeled x), until the time it takes to travel AOB is minimum, and thereby derive Snell's law. The time t it takes for light to travel from A to B through O is

$$t = \frac{AO}{c/n_1} + \frac{OB}{c/n_2} = \frac{[(x_1 - x)^2 + y_1^2]^{1/2}}{c/n_1} + \frac{[(x_2 + x)^2 + y_2^2]^{1/2}}{c/n_2} \quad (\text{P1.2})$$

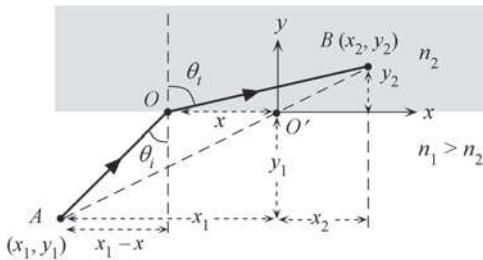


FIGURE 1.56 Consider a light wave traveling from point $A(x_1, y_1)$ to $B(x_2, y_2)$ through an arbitrary point O at a distance x from O' . The principle of least time from A to B requires that O is such that the incidence and refraction angles obey Snell's law.

The incidence and transmittance angles are given by

$$\sin \theta_i = \frac{x_1 - x}{[(x_1 - x)^2 + y_1^2]^{1/2}} \quad \text{and} \quad \sin \theta_r = \frac{(x_2 + x)}{[(x_2 + x)^2 + y_2^2]^{1/2}} \quad (\text{P1.3})$$

Differentiate Eq. (P1.2) with respect to x to find the condition for the “least time” and then use Eq. (P1.3) in this condition to derive Snell’s law.



Pierre de Fermat (1601–1665) was a French mathematician who made many significant contributions to modern calculus, number theory, analytical geometry, and probability. (*Courtesy of Mary Evans Picture Library/Alamy*.)

1.15 Antireflection (AR) coating

- (a) A laser beam of wavelength 1550 nm from air is launched to a single mode optical fiber with a core refractive index $n_1 = 1.45$. Estimate the refractive index and thickness of film required for an anti-reflecting coating on this fiber.
- (b) A Ge photodiode is designed to operate at 1550 nm, and it is required to have AR coatings to minimize reflected light. Two possible materials are available for AR coating: SiO_2 with a refractive index of 1.46, and TiO_2 with a refractive index of 2.2. Which would be better suited? What would be the thickness for the AR coating on this photodiode? The refractive index of Ge is about 4.
- (c) Consider a Ge photodiode that is designed for operation around 1200 nm. What are the best AR coating refractive index and thickness if the refractive index of Ge is about 4.0?

1.16 Single- and double-layer antireflection V-coating

For a single-layer AR coating of index n_2 on a material with index $n_3 (> n_2 > n_1)$, as shown in Figure 1.57 (a), the minimum reflectance at normal incidence is given by

$$R_{\min} = \left[\frac{n_2^2 - n_1 n_3}{n_2^2 + n_1 n_3} \right]^2$$

when the reflections A, B, \dots all interfere as destructively as possible. $R_{\min} = 0$ when $n_2 = (n_1 n_3)^{1/2}$. The choice of materials may not always be the best for a single-layer AR coating. *Double-layer AR coatings*, as shown in

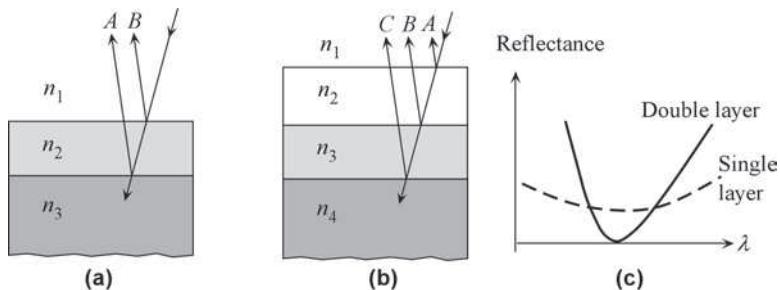


FIGURE 1.57 (a) A single-layer AR coating. (b) A double-layer AR coating and (c) its V-shaped reflectance spectrum over a wavelength range.

Figure 1.57 (b), can achieve lower and sharper reflectance at a specified wavelength as in Figure 1.57 (c). To reduce the reflection of light at the n_1-n_4 interface, two layers n_2 and n_3 , each quarter wavelength in the layer (λ/n_2 and λ/n_3) are interfaced between n_1 and n_4 . The reflections A , B , and C for normal incidence result in a minimum reflectance given by

$$R_{\min} = \left[\frac{n_3^2 n_1 - n_4 n_2^2}{n_3^2 n_1 + n_4 n_2^2} \right]^2$$

The double-layer reflectance vs. wavelength behavior usually has a V-shape, and such coatings are called *V-coatings*.

(a) Show that double-layer reflectance vanishes when

$$(n_2/n_3)^2 = n_1/n_4$$

- (b) Consider an InGaAs, a semiconductor crystal with an index 3.8, for use in a photodetector. What is the reflectance without any AR coating?
- (c) What is the reflectance when InGaAs is coated with a thin AR layer of Si_3N_4 ? Which material in Table 1.3 would be ideal as an AR coating?

TABLE 1.3 Typical AR materials and their approximate refractive indices over the visible wavelengths

	MgF_2	SiO_2	Al_2O_3	CeF_3	Sb_2O_3	Si_3N_4	SiO	ZrO_2	ZnS	TiO_2	CdS
n	1.38	1.46	1.65	1.65	1.9–2.1	1.95	2.0	2.05	2.35	2.35	2.60

- (d) What two materials would you choose to obtain a V-coating? Note: The choice of an AR coating also depends on the technology involved in depositing the AR coating and its effects on the interface states between the AR layer and the semiconductor. $\text{Si}_{1-x}\text{N}_x$ is a common AR coating on devices inasmuch as it is a good passive dielectric layer, its deposition technology is well established and changing its composition (x) changes its index.

- 1.17 Single-, double-, and triple-layer antireflection coatings** Figure 1.58 shows the reflectance of an uncoated glass, and glass that has a single- (1), double- (2) and triple- (3) layer AR coatings? The coating details are in the figure caption. Each layer in single- and double-layer AR coatings has a thickness of $\lambda/4$, where λ is the wavelength in the layer. The triple-layer AR layer has three coatings with thicknesses $\lambda/4$, $\lambda/2$, and $\lambda/4$. Can you qualitatively explain the results by using interference? What applications would need single-, double-, and triple-layer coatings?

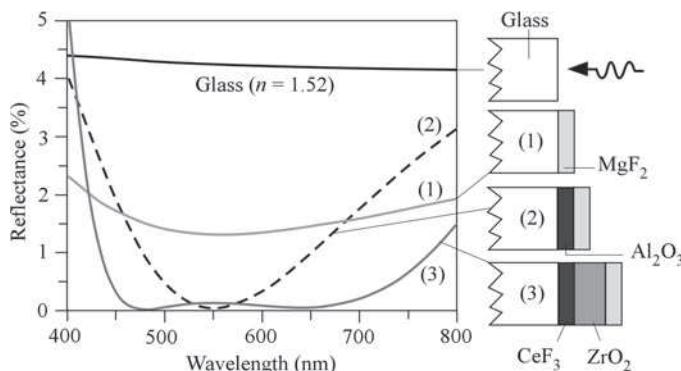


FIGURE 1.58 Reflectance vs. wavelength for a glass plate, $n = 1.52$, with and without AR coatings. (1) Single-layer AR coating is a quarter wavelength ($\lambda/4$) thick MgF_2 , $n = 1.38$. (2) Double-layer coating is $\lambda/4$ thick MgF_2 and $\lambda/4$ thick Al_2O_3 , $n = 1.69$. (3) Triple-layer coating is $\lambda/4$ thick MgF_2 , $\lambda/2$ thick ZrO_2 , $n = 2.05$, and a $\lambda/4$ thick CeF_3 , $n = 1.64$. (Source: Plotted from data appearing in Figure 2.2 in S. Chattopadhyay *et al.*, *Mater. Sci. Engin. R*, 69, 1, 2010.)

1.18 Reflection at glass–glass and air–glass interfaces A ray of light that is traveling in a glass medium of refractive index $n_1 = 1.460$ becomes incident on a less dense glass medium of refractive index $n_2 = 1.430$. Suppose that the free-space wavelength of the light ray is 850 nm.

- What should the minimum incidence angle for TIR be?
- What is the phase change in the reflected wave when the angle of incidence $\theta_i = 85^\circ$ and when $\theta_i = 90^\circ$?
- What is the penetration depth of the evanescent wave into medium 2 when $\theta_i = 85^\circ$ and when $\theta_i = 90^\circ$?
- What is the reflection coefficient and reflectance at normal incidence ($\theta_i = 0^\circ$) when the light beam traveling in the glass medium ($n = 1.460$) is incident on a glass–air interface?
- What is the reflection coefficient and reflectance at normal incidence when a light beam traveling in air is incident on an air–glass ($n = 1.460$) interface? How do these compare with part (d) and what is your conclusion?

1.19 Dielectric mirror A dielectric mirror is made up of a quarter wave layer of GaAs with $n_H = 3.38$ and AlAs with $n_L = 3.00$ at around 1550 nm. The light is incident on the mirror from another semiconductor of refractive index $n_0 = 3.40$. Find out the number of pairs of layers N needed to get 90% reflectance. Find out the bandwidth of the reflected light.

1.20 TIR and polarization at water–air interface

- Given that the refractive index of water is about 1.33, what is the polarization angle for light traveling in air and reflected from the surface of the water?
- Consider a diver in sea pointing a flashlight towards the surface of the water. What is the critical angle for the light beam to be reflected from the water surface?

1.21 Reflection and transmission at a semiconductor–semiconductor interface A light wave with a wavelength of 890 nm (free-space wavelength) that is propagating in GaAs becomes incident on AlGaAs. The refractive index of GaAs is 3.60, that of AlGaAs is 3.30.

- Consider normal incidence. What are the reflection and transmission coefficients and the reflectance and transmittance? (From GaAs into AlGaAs.)
- What is the Brewster angle (the polarization angle θ_p) and the critical angle (θ_c) for total internal reflection for the wave in (a); the wave that is traveling in GaAs and incident on the GaAs–AlGaAs interface?
- What is the reflection coefficient and the phase change in the reflected wave when the angle of incidence $\theta_i = 79^\circ$?
- What is the penetration depth of the evanescent wave into medium 2 when $\theta_i = 79^\circ$ and when $\theta_i = 89^\circ$? What is your conclusion?

1.22 Phase changes on TIR Consider a light wave of wavelength 870 nm traveling in a semiconductor medium (GaAs) of refractive index 3.60. It is incident on a different semiconductor medium (AlGaAs) of refractive index

3.40, and the angle of incidence is 80° . Will this result in total internal reflection? Calculate the phase change in the parallel and perpendicular components of the reflected electric field.

- 1.23 Fresnel's equations** Fresnel's equations are sometimes given as follows:

$$r_{\perp} = \frac{E_{ro,\perp}}{E_{io,\perp}} = \frac{n_1 \cos \theta_i - n_2 \cos \theta_t}{n_1 \cos \theta_i + n_2 \cos \theta_t}$$

$$r_{\parallel} = \frac{E_{ro,\parallel}}{E_{io,\parallel}} = \frac{n_1 \cos \theta_i - n_2 \cos \theta_t}{n_1 \cos \theta_t + n_2 \cos \theta_i}$$

$$t_{\perp} = \frac{E_{io,\perp}}{E_{ro,\perp}} = \frac{2n_1 \cos \theta_i}{n_1 \cos \theta_i + n_2 \cos \theta_t}$$

and

$$t_{\parallel} = \frac{E_{io,\parallel}}{E_{ro,\parallel}} = \frac{2n_1 \cos \theta_i}{n_1 \cos \theta_t + n_2 \cos \theta_i}$$

Show that these reduce to Fresnel's equation given in Eqs. (1.6.6) and (1.6.7).

Using Fresnel's equations, find the reflection and transmission coefficients for normal incidence and show that

$$r_{\perp} + 1 = t_{\perp} \quad \text{and} \quad r_{\parallel} + nt_{\parallel} = 1$$

where $n = n_2/n_1$.

- 1.24 Fresnel's equations** Consider a light wave traveling in a glass medium with an index $n_1 = 1.440$ and it is incident on the glass-air interface. Using Fresnel's equations only, that is, Eqs. (1.6.6a) and (1.6.6b), calculate the reflection coefficients r_{\perp} and r_{\parallel} and hence reflectances R_{\perp} and R_{\parallel} for (a) $\theta_i = 25^\circ$ and (b) $\theta_i = 50^\circ$. In the case of $\theta_i = 50^\circ$, find the phase change ϕ_{\perp} and ϕ_{\parallel} from the reflection coefficients by writing $r = |r| \exp(-j\phi)$. Compare ϕ_{\perp} and ϕ_{\parallel} from r_{\perp} and r_{\parallel} calculations with those calculated from Eqs. (1.6.11) and (1.6.12).
- 1.25 Goos-Haenchen phase shift** A ray of light which is traveling in a glass medium (1) of refractive index $n_1 = 1.460$ becomes incident on a less dense glass medium (2) of refractive index $n_2 = 1.430$. Suppose that the free-space wavelength of the light ray is 850 nm. The angle of incidence $\theta_i = 85^\circ$. Estimate the lateral Goos-Haenchen shift in the reflected wave for the perpendicular field component. Recalculate the Goos-Haenchen shift if the second medium has $n_2 = 1$ (air). What is your conclusion? Assume that the virtual reflection occurs from a virtual plane in medium B at a distance d that is roughly the same as the penetration depth. Note that d actually depends on the polarization, the direction of the field, but we will ignore this dependence.
- 1.26 Evanescent wave** Total internal reflection of a plane wave from a boundary between a more dense medium (1) n_1 and a less dense medium (2) n_2 is accompanied by an evanescent wave propagating in medium 2 near the boundary. Find the functional form of this wave and discuss how its magnitude varies with the distance into medium 2.
- 1.27 TIR and FTIR**
- (a) By considering the electric field component in medium B in Figure 1.21, explain how you can adjust the amount of transmitted light through a thin layer between two higher refractive index media.
 - (b) What is the critical angle at the hypotenuse face of a beam splitter cube made of glass with $n_1 = 1.6$ and having a thin film of liquid with $n_2 = 1.3$. Can you use 45° prisms with normal incidence?
 - (c) Explain how a light beam can propagate along a layer of material between two different media as shown in Figure 1.59 (a). Explain what the requirements are for the indices n_1, n_2, n_3 . Will there be any losses at the reflections?
 - (d) Consider the prism coupler arrangement in Figure 1.59 (b). Explain how this arrangement works for coupling an external light beam from a laser into a thin layer on the surface of a glass substrate. Light is then propagated inside the thin layer along the surface of the substrate. What is the purpose of the adjustable coupling gap?

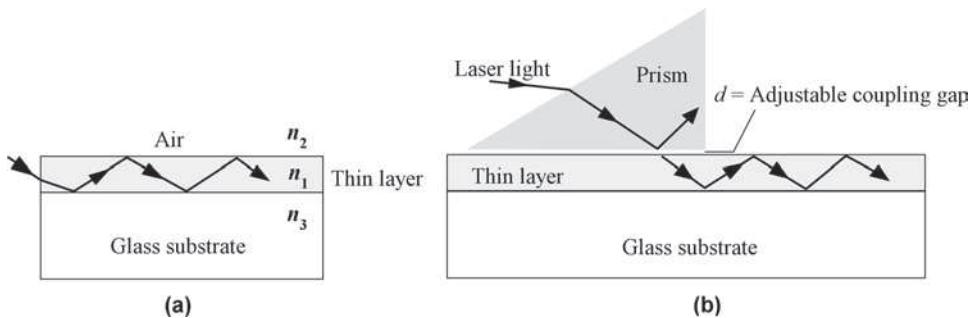


FIGURE 1.59 (a) Light propagation along an optical guide. (b) Coupling of laser light into a thin layer—optical guide—using a prism. The light propagates along the thin layer.

- 1.28 Complex refractive index and dielectric constant** The complex refractive index $N = n - jK$ can be defined in terms of the *complex relative permittivity* $\epsilon_r = \epsilon_{r1} - j\epsilon_{r2}$ as

$$N = n - jK = \epsilon_r^{1/2} = (\epsilon_{r1} - j\epsilon_{r2})^{1/2}$$

Show that

$$n = \left[\frac{(\epsilon_{r1}^2 + \epsilon_{r2}^2)^{1/2} + \epsilon_{r1}}{2} \right]^{1/2} \quad \text{and} \quad K = \left[\frac{(\epsilon_{r1}^2 + \epsilon_{r2}^2)^{1/2} - \epsilon_{r1}}{2} \right]^{1/2}$$

- 1.29 Complex refractive index** Spectroscopic ellipsometry measurements on a germanium crystal at a photon energy of 1.5 eV show that the real and imaginary parts of the complex relative permittivity are 21.56 and 2.772, respectively. Find the complex refractive index. What is the reflectance and absorption coefficient at this wavelength? How do your calculations match with the experimental values of $n = 4.653$ and $K = 0.298$, $R = 0.419$ and $\alpha = 4.53 \times 10^6 \text{ m}^{-1}$?

- 1.30 Complex refractive index** Figure 1.26 shows the infrared extinction coefficient K of CdTe. Calculate the absorption coefficient α and the reflectance R of CdTe at 60 μm and 80 μm .

- 1.31 Refractive index and attenuation in the infrared region—Reststrahlen absorption** Figure 1.26 shows the refractive index n and the extinction coefficient K as a function of wavelength λ in the infrared for a CdTe crystal due to lattice absorption, called Reststrahlen absorption. It results from the ionic polarization of the crystal induced by the optical field in the light wave. The relative permittivity ϵ_r due to positive (Cd^{2+}) and negative (Te^{2-}) ions being made to oscillate by the optical field about their equilibrium positions is given in its simplest form by

$$\epsilon_r = \epsilon'_r - j\epsilon''_r = \epsilon_{rL} + \frac{\epsilon_{rH} - \epsilon_{rL}}{\left(\frac{\omega}{\omega_T} \right)^2 - 1 + j \frac{\gamma}{\omega_T} \left(\frac{\omega}{\omega_T} \right)} \quad (\text{P1.4})$$

where ϵ_{rL} and ϵ_{rH} are the relative permittivity at low (L) and high (H) frequencies, well below and above the infrared peak, γ is a loss coefficient characterizing the rate of energy transfer from the EM wave to lattice vibrations (phonons), and ω_T is a transverse optical lattice vibration frequency that is related to the nature of bonding between the ions in the crystal. Table 1.4 provides some typical values for CdTe and GaAs. Equation (P1.4) can be used to obtain a reasonable approximation to the infrared refractive index n and extinction coefficient K due to Reststrahlen absorption. (a) Consider CdTe, and plot n and K vs. λ from 40 μm to 90 μm and compare with the experimental results in Figure 1.26 in terms of the peak positions and the width of the extinction coefficient peak. (b) Consider

TABLE 1.4 Ionic polarization resonance parameters for CdTe and GaAs

	ϵ_{rL}	ϵ_{rH}	$\omega_T (\text{rad s}^{-1})$	$\gamma (\text{rad s}^{-1})$
CdTe	10.20	7.10	2.68×10^{13}	0.124×10^{13}
GaAs	13.0	11.0	5.07×10^{13}	0.045×10^{13}

GaAs, and plot n and K vs. λ from 30 μm to 50 μm . (c) Calculate n and K for GaAs at $\lambda = 38.02 \mu\text{m}$ and compare with the experimental values $n = 7.55$ and $K = 0.629$. (You might want to use a logarithmic scale for K .)

- 1.32 Coherence length** A narrow band pass filter transmits wavelengths in the range $5000 \pm 0.5 \text{ \AA}$. If this filter is placed in front of a source of white light, what is the coherence length of the transmitted light?

1.33 Spectral widths and coherence

- (a) Suppose that frequency spectrum of a radiation emitted from a source has a central frequency v_o and a spectral width Δv . The spectrum of this radiation in terms of wavelength will have a central wavelength λ_o and a spectral width $\Delta\lambda$. Clearly, $\lambda_o = c/v_o$. Since $\Delta\lambda \ll \lambda_o$ and $\Delta v \ll v_o$, using $\lambda = c/v$, show that the line width $\Delta\lambda$ and hence the coherence length l_c are

$$\Delta\lambda = \Delta v \frac{\lambda_o}{v_o} = \Delta v \frac{\lambda_o^2}{c} \quad \text{and} \quad l_c = c\Delta t = \frac{\lambda_o^2}{\Delta\lambda}$$

- (b) Calculate $\Delta\lambda$ for a lasing emission from a He-Ne laser that has $\lambda_o = 632.8 \text{ nm}$ and $\Delta v \approx 1.5 \text{ GHz}$. Find its coherence time and length.

- 1.34 Coherence lengths** Find the coherence length of the following light sources:

- (a) An LED emitting at 1550 nm with a spectral width 150 nm;
- (b) A semiconductor laser diode emitting at 1550 nm with a spectral width 3 nm;
- (c) A quantum well semiconductor laser diode emitting at 1550 nm with a spectral width of 0.1 nm;
- (d) A multimode He-Ne laser with a spectral frequency width of 1.5 GHz;
- (e) A specially designed single mode and stabilized He-Ne laser with a spectral width of 100 MHz.

- 1.35 Fabry-Perot optical cavity** Consider an optical cavity formed between two identical mirrors, each with reflectance = 0.97. The refractive index of the medium enclosed between the mirrors is 1. Find out the minimum length of the optical cavity which can resolve spectral lines of a sodium lamp with line width $\Delta\lambda = 0.6 \text{ nm}$ and $\Delta\lambda = 589.3 \text{ nm}$. Further, estimate the mode separation in frequency and wavelength. What are the finesse F and Q factors for this cavity?

- 1.36 Fabry-Perot optical cavity from a ruby crystal** Consider a ruby crystal of diameter 1 cm and length 10 cm. The refractive index is 1.78. The ends have been silvered and the reflectances are 0.99 and 0.95 each. What is the nearest mode number that corresponds to a radiation of wavelength 694.3 nm? What is the actual wavelength of the mode closest to 694.3 nm? What is the mode separation in frequency and wavelength? What are the finesse F and Q factor for the cavity?

- 1.37 Fabry-Perot optical cavity spectral width** Consider an optical cavity of length 40 cm. Assume the refractive index is 1, and use Eq. (1.11.3) to plot the peak closest to 632.8 nm for 4 values of $R = 0.99, 0.90, 0.75$ and 0.6 . For each case find the spectral width $\delta\lambda_m$, the finesse F and Q . How accurate is Eq. (1.11.5) in predicting $\delta\lambda_m$? (You may want to use a graphing software for this problem.)

- 1.38 Diffraction** A collimated beam of light of wavelength 632.8 nm is incident on a circular aperture of 250 μm . Find out the divergence of the transmitted beam. Obtain the diameter of the transmitted beam at a distance of 10 m. What would be the divergence if the aperture is a single slit of width 250 μm ?

- 1.39 Diffraction intensity** Consider diffraction from a uniformly illuminated circular aperture of diameter D . The far field diffraction pattern is given by a Bessel function of the first kind and first order, J_1 , and the intensity at a point P on the angle θ_i with respect to the central axis through the aperture is

$$I(\gamma) = I_o \left(\frac{2J_1(\gamma)}{\gamma} \right)^2$$

where I_o is the maximum intensity, $\gamma = (1/2)kD \sin \theta$ is a variable quantity that represents the angular position θ on the screen as well as the wavelength ($k = 2\pi/\lambda$) and the aperture diameter D . $J_1(\gamma)$ can be calculated from

$$J_1(\gamma) = \frac{1}{\pi} \int_0^\pi \cos(\alpha - \gamma \sin \alpha) d\alpha$$

where α is an integration variable. Using numerical integration, or a suitable mathematics software program, plot $[J_1(\gamma)/\gamma]$ vs. γ for $\gamma = 0 - 8$ and confirm that zero-crossings occur at $\gamma = 3.83, 7.02$ and the maxima at $\gamma = 0, 5.14$. What is the intensity ratio of the first bright ring (at $\gamma = 5.14$) to that at the center of the Airy disk ($\gamma = 0$)? (You can use a very small γ instead of zero for the center intensity calculation.) Using the first zero at $\gamma = 3.83$, verify Eq. (1.12.5), $\sin \theta_o = 1.22\lambda/D$, where θ_o is the angular position of the first dark ring, as defined in Figure 1.36 (b).



George Bidell Airy (1801–1892, England).
George Airy was a professor of astronomy at Cambridge and then the Astronomer Royal at the Royal Observatory in Greenwich, England.
(© Mary Evans Picture Library/Alamy.)

- 1.40 Bragg diffraction** A reflection grating is made on the surface of a semiconductor with a periodicity of $0.5 \mu\text{m}$. If light of wavelength $1.55 \mu\text{m}$ is incident at an angle of 88° to the normal, find out the diffracted beam.
- 1.41 Diffraction grating for WDM** Consider a transmission diffraction grating. Suppose that we wish to use this grating to separate out different wavelengths of information in a WDM signal at 1550 nm . (WDM stands of wavelength division multiplexing.) Suppose that the diffraction grating has a periodicity of $2 \mu\text{m}$. The angle of incidence is 0° with respect to the normal to the diffraction grating. What is the angular separation of the two wavelength components at $1.550 \mu\text{m}$ and $1.540 \mu\text{m}$? How would you increase this separation?
- 1.42 A monochromator** Consider an incident beam on a reflection diffraction grating as in Figure 1.60. Each incident wavelength will result in a diffracted wave with a different diffraction angle. We can place a small slit and allow only one diffracted wave λ_m to pass through to the photodetector. The diffracted beam would consist of wavelengths in the incident beam separated (or fanned) out after diffraction. Only one wavelength λ_m will be diffracted favorably to pass through the slit and reach the photodetector. Suppose that the slit width is $s = 0.1 \text{ mm}$, and the slit is at a distance $R = 5 \text{ cm}$ from the grating. Suppose that the slit is placed so that it is at right angles to the incident beam: $\theta_i + \theta_m = \pi/2$. The grating has a corrugation periodicity of $1 \mu\text{m}$.

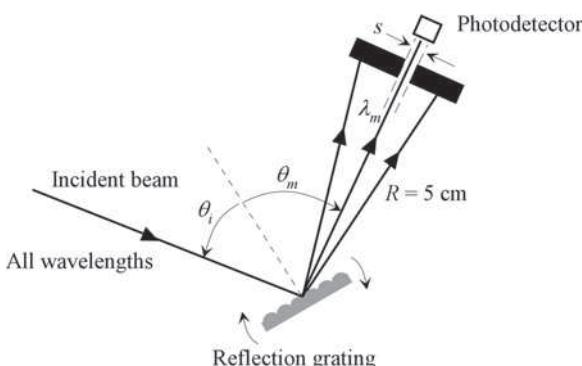


FIGURE 1.60 A monochromator based on using a diffraction grating.

- (a) What is the range of wavelengths that can be captured by the photodetector when we rotate the grating from $\theta_i = 1^\circ$ to 40° ?
- (b) Suppose that $\theta_i = 15^\circ$. What is the wavelength that will be detected? What is the resolution, that is, the range of wavelengths that will pass through the slit? How can you improve the resolution? What would be the advantage and disadvantage in decreasing the slit width s ?

1.43 Thin film optics Consider light incident on a thin film on a substrate, and assume normal incidence for simplicity.

- (a) Consider a thin soap film in air, $n_1 = n_3 = 1$, $n_2 = 1.40$. If the soap thickness $d = 1 \mu\text{m}$, plot the reflectance vs. wavelength from $0.35 \mu\text{m}$ to $0.75 \mu\text{m}$, which includes the visible range. What is your conclusion?
- (b) MgF₂ thin films are used on glass plates for the reduction of glare. Given that $n_1 = 1$, $n_2 = 1.38$, and $n_3 = 1.60$ (n for glass depends on the type of glass but 1.6 is a reasonable value), plot the reflectance as a function of wavelength from $0.35 \mu\text{m}$ to $0.75 \mu\text{m}$ for a thin film of thickness $0.10 \mu\text{m}$. What is your conclusion?

1.44 Thin film optics Consider a glass substrate with $n_3 = 1.65$ that has been coated with a transparent optical film (a dielectric film) with $n_2 = 2.50$, $n_1 = 1$ (air). If the film thickness is 500 nm , find the minimum and maximum reflectances and transmittances and their corresponding wavelengths in the visible range for normal incidence. (Assume normal incidence.) Note that the thin n_2 -film is not an AR coating, and for $n_1 < n_3 < n_2$,

$$R_{\max} = \left(\frac{n_2^2 - n_1 n_3}{n_2^2 + n_1 n_3} \right)^2 \quad \text{and} \quad R_{\min} = \left(\frac{n_3 - n_1}{n_3 + n_1} \right)^2$$

1.45 Thin film optics Consider light incident on a thin film on a substrate, and assume normal incidence for simplicity. Plot the reflectance R and transmittance T as a function of the phase change ϕ from $\phi = -4\pi$ to $+4\pi$ for the following cases

- (a) Thin soap film in air, $n_1 = n_3 = 1$, $n_2 = 1.40$. If the soap thickness $d = 1 \mu\text{m}$, what are the maxima and minima in the reflectance in the visible range?
- (b) A thin film of MgF₂ on a glass plate for the reduction of glare, where $n_1 = 1$, $n_2 = 1.38$, and $n_3 = 1.70$ (n for glass depends on the type of glass but 1.7 is a reasonable value.) What should be the thickness of MgF₂ for minimum reflection at 550 nm ?
- (c) A thin film of semiconductor on glass where $n_1 = 1$, $n_2 = 3.5$, and $n_3 = 1.55$.

1.46 Transmission through a plate Consider the transmittance of light through a partially transparent glass plate of index n in which light experiences attenuation (either by absorption or scattering). Suppose that the plate is in a medium of index n_o , the reflectance at each $n-n_o$ interface is R and the attenuation coefficient is α .

- (a) Show that

$$T_{\text{plate}} = \frac{(1 - R)^2 e^{-\alpha d}}{(1 - R^2)e^{-2\alpha d}}$$

- (b) If T is the transmittance of a glass plate of refractive index n in a medium of index n_o show that, in the absence of any absorption in the glass plate,

$$n/n_o = T^{-1} + (T^{-2} - 1)^{1/2}$$

if we neglect any losses in the glass plate.

- (c) If the transmittance of a glass plate in air has been measured to be 89.96%. What is its refractive index? Do you think this is a good way to measure the refractive index?

1.47 Scattering Consider Rayleigh scattering. If the incident light is unpolarized, the intensity I_s of the scattered light at a point at a distance r at an angle θ to the original light beam is given by

$$I_s \propto \frac{1 - \cos^2 \theta}{r^2}$$

Plot a polar plot of the intensity I_s at a fixed distance r from the scatter as we change the angle θ around the scatterer. In a polar plot, the radial coordinate (OP in Figure 1.48 (b)) is I_s . Construct a contour plot in the xy plane in which a contour represents a constant intensity. You need to vary r and θ or x and y such that I_s remains constant. Note $x = r \cos \theta$ and $y = r \sin \theta$, $\theta = \arctan(y/x)$, $r = (x^2 + y^2)^{1/2}$.

1.48 One-dimensional photonic crystal (a Bragg mirror) The 1D photonic crystal in Figure 1.50 (a), which is essentially a Bragg reflector, has the dispersion behavior shown in Figure 1.51 (a). The stop-band $\Delta\omega$ for normal incidence and for all polarizations of light is given by (R. H. Lipson and C. Lu, *Eur. J. Phys.*, 30, S33, 2009)

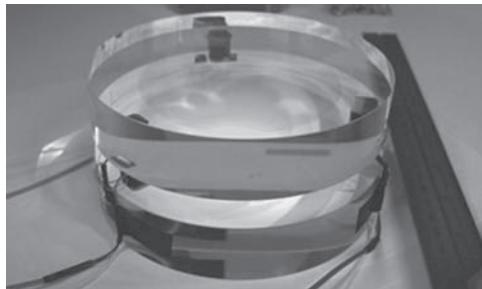
$$\frac{\Delta\omega}{\omega_o} = (4/\pi) \arcsin \left(\frac{n_2 - n_1}{n_2 + n_1} \right)$$

where $\Delta\omega$ is the stop band, ω_o is the center frequency defined in Figure 1.51 and n_2 and n_1 are the high and low refractive indices. Calculate the lowest stop band in terms of photon energy in eV, and wavelength 1550 nm for a 1D photonic crystal structure with $n_1d_1 = n_2d_2 = \lambda/4$, made up of: (i) Si ($n_{\text{Si}} = 3.5$) and SiO_2 ($n_{\text{SiO}_2} = 1.445$) pairs, and (ii) Si_3N_4 ($n_{\text{Si}_3\text{N}_4} = 2.0$) and SiO_2 pairs.

- 1.49 Photonic crystals** Concepts have been borrowed from crystallography, such as a unit cell, to define a photonic crystal. What is the difference between a unit cell used in a photonic crystal and that used in a real crystal? What is the size limit on the unit cell of a photonic crystal? Is the refractive index a microscopic or a macroscopic concept? What is the assumption on the refractive index?



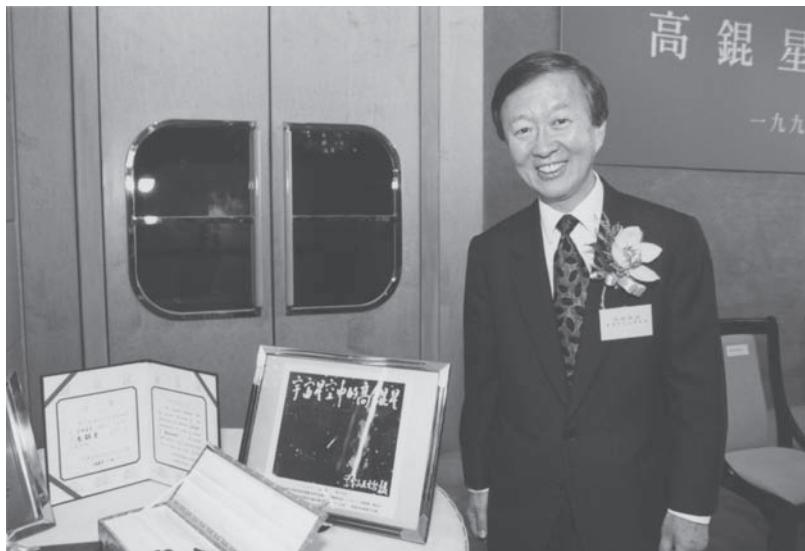
A scanning Fabry–Perot interferometer (Model SA200), used as a spectrum analyzer, which has a free spectral range of 1.5 GHz, a typical finesse of 250, spectral width (resolution) of 7.5 MHz. The cavity length is 5 cm. It uses two concave mirrors instead of two planar mirrors to form the optical cavity. A piezoelectric transducer is used to change the cavity length and hence the resonant frequencies. A voltage ramp is applied through the coaxial cable to the piezoelectric transducer to scan frequencies. (*Courtesy of Thorlabs.*)



This is a tunable large aperture (80 mm) etalon with two end plates that act as reflectors. The end plates have been machined to be flat to wavelength/110. There are three piezoelectric transducers that can tilt the end plates and hence obtain perfect alignment. (*Courtesy of Light Machinery.*)

*The introduction of optical fiber systems will revolutionize the communications network.
The low-transmission loss and the large bandwidth capability of the fiber systems allow
signals to be transmitted for establishing communications contacts over large distances
with few or no provisions of intermediate amplification.*

—Charles K. Kao¹



Charles Kao at the naming ceremony of Minor Planet (3463) "Kaokuen" by Nanjing's Purple Mountain Observatory in July 1996. Charles Kao and his colleagues carried out the early experiments on optical fibers at the Standard Telecommunications Laboratories Ltd. (the research center of Standard Telephones and Cables) at Harlow in the United Kingdom, during the 1960s. He shared the Nobel Prize in 2009 in Physics with Willard Boyle and George Smith for "groundbreaking achievements concerning the transmission of light in fibers for optical communication." In a milestone paper with George Hockam published in the IEE Proceedings in 1966 they predicted that the intrinsic losses of glass optical fibers could be much lower than 20 dB km^{-1} , which would allow their use in long-distance telecommunications. Today, optical fibers are used not only in telecommunications but also in various other technologies such as instrumentation and sensing. From 1987 to his retirement in 1996, Professor Kao was the vice-chancellor of the Chinese University of Hong Kong. (Courtesy of the Chinese University of Hong Kong.)

¹Charles K. Kao (one of the pioneers of glass fibers for optical communications) *Optical Fiber Systems: Technology, Design, and Applications* (McGraw-Hill, 1982), p. 1.

2

DIELECTRIC WAVEGUIDES AND OPTICAL FIBERS

2.1 SYMMETRIC PLANAR DIELECTRIC SLAB WAVEGUIDE

A. Waveguide Condition

To understand the general nature of light wave propagation in optical waveguides, we first consider the planar dielectric slab waveguide shown in Figure 2.1, which is the simplest waveguide in terms of tractable analysis. A slab of dielectric of thickness $2a$ and refractive index n_1 is sandwiched between two semi-infinite regions, both of refractive index n_2 ($n_2 < n_1$). The region of higher refractive index (n_1) is called the **core** and the region of lower refractive index (n_2) sandwiching the core is called the **cladding**.

A light ray can readily propagate along such a waveguide, in a zigzag fashion, provided it can undergo total internal reflection (TIR) at the dielectric boundaries. It seems that any light wave that has an angle of incidence (θ) at the core–cladding interface greater than the critical angle (θ_c) for TIR will be propagated. This, however, is true only for a very thin light beam with a diameter much less than the slab thickness, $2a$. We consider the realistic case when the whole end of the waveguide is illuminated, as illustrated in Figure 2.1. To simplify the analysis, we will assume that light is launched from a line source in a medium of refractive index n_1 . In general, the refractive index of the launching medium will be different than n_1 , but this will affect only the amount of light coupled into the guide.

Consider a hypothetical plane wave propagating in the dielectric slab waveguide as shown in Figure 2.2. We will take the electric field E to be along x , parallel to the interface and perpendicular to z . The ray is guided in a zigzag fashion along the guide axis z by reflections from the core–cladding ($n_1–n_2$) boundaries. The result is the effective propagation of the electric field E along z . The figure also shows the constant phase wavefronts, normal to direction of propagation, on this ray. This particular ray is reflected at B and then at C . Just after the reflection at C , the wavefront at C overlaps the wavefront at A on the original ray. The wave interferes with itself. Unless these wavefronts at A and C are in phase, the two will interfere destructively and destroy each other. Only certain reflection angles θ give rise to the constructive interference and hence only certain waves can exist in the guide.

The phase difference between the points A and C corresponds to an optical path length $AB + BC$. In addition, there are two total internal reflections (TIRs) at B and C and each introduces a further phase change of ϕ . Suppose that k_1 is the

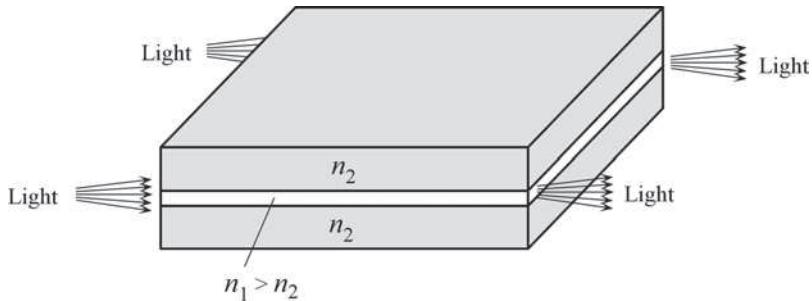


FIGURE 2.1 A planar dielectric waveguide has a central rectangular region of refractive index n_1 , which is higher than the refractive index n_2 of the surrounding region. It is assumed that the waveguide is infinitely wide and the central region is of thickness $2a$. It is illuminated at one end by a nearly monochromatic light source.

propagation constant in n_1 , that is, $k_1 = kn_1 = 2\pi n_1/\lambda$, in which k and λ are the free-space propagation constant and wavelength, respectively. The direction of k_1 is at an angle θ to the normal to the core-cladding interface as shown in Figure 2.2. For constructive interference, the phase difference $\Delta\phi_{AC}$ between A and C must be a multiple of 2π .

$$\Delta\phi_{AC} = k_1(AB + BC) - 2\phi = m(2\pi) \quad (2.1.1)$$

in which $m = 0, 1, 2, \dots$ is an integer.

We can easily evaluate the $AB + BC$ from geometrical considerations. $BC = d/\cos\theta$ and $AB = BC\cos(2\theta)$. Thus

$$AB + BC = BC\cos(2\theta) + BC = BC[2\cos^2\theta - 1] + 1 = 2d\cos\theta$$

Thus, for wave propagation along the guide we need

$$k_1(2d\cos\theta) - 2\phi = m(2\pi) \quad (2.1.2)$$

It is apparent that only certain θ and ϕ values can satisfy this equation for a given integer m . But the phase change ϕ depends on θ and also on the polarization state of the wave (direction of the electric field, that is, E_x here). In Eq. (2.1.2), for each m , there will be one allowed angle θ_m and one corresponding ϕ_m . Dividing Eq. (2.1.2) by 2 we obtain the **waveguide condition**

$$\left[\frac{2\pi n_1(2a)}{\lambda} \right] \cos\theta_m - \phi_m = m\pi \quad (2.1.3)$$

in which ϕ_m indicates that ϕ is a function of the incidence angle θ_m .

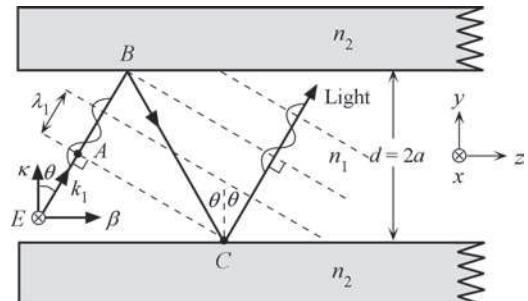


FIGURE 2.2 A light ray traveling in the guide must interfere constructively with itself to propagate successfully. Otherwise destructive interference will destroy the wave. E is parallel to x . (λ_1 and k_1 are the wavelength and the wave vector inside the core medium n_1 , that is, $\lambda_1 = \lambda/n_1$).

It may be thought that the treatment above is somewhat artificial as we took a narrow angle for θ . It turns out that Eq. (2.1.3) can be derived as a general waveguide condition for guided waves whether we use a narrow or a wider angle, one or multiple rays. We can derive the same condition if we take two arbitrary parallel rays entering the guide as in Figure 2.3. The rays 1 and 2 are initially in phase, and represent the same plane wave. Ray 1 then suffers two reflections at A and B, and is then again traveling parallel to ray 2. Unless the wavefront on ray 1 just after reflection at B is in phase with the wavefront at B' on ray 2, the two would destroy each other. Both rays initially start in phase; ray 1 at A just before reflection and ray 2 at A'. Ray 1 at B, just after two reflections, has a phase $k_1 AB - 2\phi$. Ray 2 at B' has a phase $k_1(A'B')$. The difference between the two phases must be $m(2\pi)$ and leads to the waveguide condition in Eq. (2.1.3). The waveguide condition in Eq. (2.1.3) represents those special waves, for example, the wave with rays 1 and 2 in Figure 2.3, that can propagate in the guide without self-destruction or being lost into the cladding. These waves have special angles of incidence θ_m that must satisfy Eq. (2.1.3).

We can resolve the wave vector k_1 into two propagation constants, β and κ , along and perpendicular to the guide axis z as shown in Figure 2.2. Given θ_m that satisfies the waveguide condition for a particular m , we use a subscript m to identify the particular β_m and κ_m as

$$\beta_m = k_1 \sin \theta_m = \left(\frac{2\pi n_1}{\lambda} \right) \sin \theta_m \quad (2.1.4)$$

Propagation constant along the guide

and

$$\kappa_m = k_1 \cos \theta_m = \left(\frac{2\pi n_1}{\lambda} \right) \cos \theta_m \quad (2.1.5)$$

Transverse propagation constant

The simplistic analysis as embedded in the waveguide condition in Eq. (2.1.3) shows quite clearly that only certain reflection angles are allowed within the guide corresponding to $m = 0, 1, 2, \dots$. We note that higher m values yield lower θ_m . Each choice of m leads to a different reflection angle θ_m . Each different m value leads to a different propagation constant along the guide given by Eq. (2.1.4).

In Figure 2.3 we only considered an initially upward traveling wave and followed what happens to rays 1 and 2 in this wave, and derived the condition that this wave can propagate along the guide. An initially downward traveling wave would also be allowed provided that its incidence angle θ_m also satisfied the waveguide condition. Indeed, for a given m , we would have upward and downward traveling waves, indicated as waves A and B in Figure 2.4. These upward and downward traveling waves both propagate along z with the same propagation constant β_m , but travel in opposite directions along y with equal but opposite propagation constants $\pm \kappa_m$ as shown in Figure 2.4. We know that when two identical waves traveling in opposite directions interfere, they set up a standing

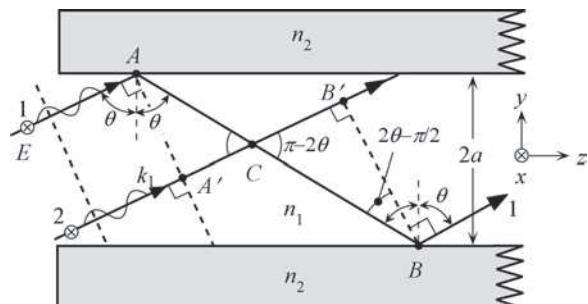
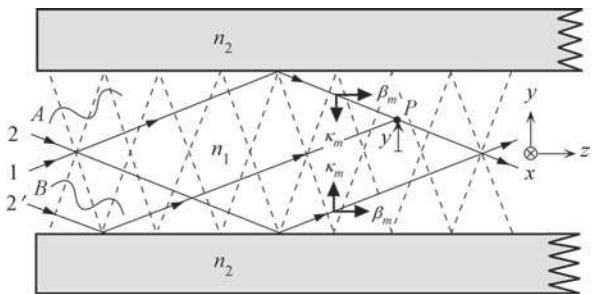


FIGURE 2.3 Two arbitrary waves 1 and 2 that are initially in phase must remain in phase after reflections from the core-cladding boundaries. Otherwise the two will interfere destructively and cancel each other.

FIGURE 2.4 We can identify upward (*A*) and downward (*B*) traveling waves in the guide which interfere to set up a standing wave along *y* and a wave that is propagating along *z*. Rays 2 and 2' initially belong to the same wavefront but 2' becomes reflected before 2. The interference of 1 and 2' determines the field at a height *y* from the guide center.



wave—in this case, a standing wave that is not propagating along *y*. We can calculate the electric field at a height *y* from the guide center by considering the interference of ray 1 on *A* and ray 2' on *B* at point *P* in Figure 2.4, which is given as a problem in Questions 2.3 and 2.4 at the end of the chapter. The result would give us the field distribution across the guide, along *y*, which would be propagating along *z*, as illustrated in Figure 2.5. We can write the field in the waveguide as

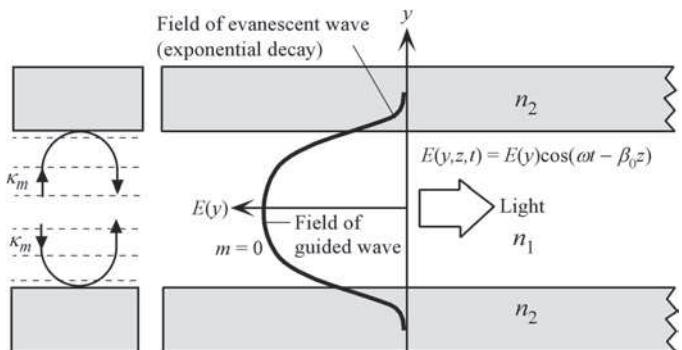
Possible waves in the guide

$$E(y, z, t) = E_m(y) \cos(\omega t - \beta_m z + \delta) \quad (2.1.6)$$

in which $E_m(y)$ is the field distribution along *y* for a given *m* and δ is a phase angle that accounts for the fact that the field may not necessarily be maximum at $t = 0$ and $z = 0$, depending on where the origin is taken. Equation (2.1.6) represents a particular field distribution $E_m(y)$ across the guide that is *traveling down the guide along z* with a wave vector β_m .

Figure 2.5 shows the field pattern for the lowest mode, $m = 0$, which has a maximum intensity at the center. This is very similar to the lowest-order standing wave in which a half-wavelength oscillation has been generated in a string stretched between two fixed points—two walls. In our case, the whole field distribution is moving along *z* with a propagation vector, β_0 . Also notice that the field penetrates into the cladding, which is due to a propagating evanescent wave in the cladding near the boundary. The field of the evanescent wave in the cladding (in medium *n*₂) decays exponentially with distance from the boundary, following an $\exp(-\alpha_2 y')$ type of behavior where *y'* is measured from the boundary (*y'* = *y* − *a*) and α_2 is the attenuation coefficient in medium *n*₂ (as derived in Chapter 1). The field pattern in the core exhibits a harmonic variation across the guide, or along *y* as it represents a partial standing wave that has resulted from the interference of two oppositely traveling waves. Figure 2.6 illustrates the field patterns for the first three modes, $m = 0, 1$, and 3 .

FIGURE 2.5 Left: The upward and downward traveling waves have equal but opposite propagation constants $\pm \kappa_m$ and interfere to set up a standing electric field pattern across the guide. Right: The electric field pattern of the lowest mode traveling wave along the guide. This mode has $m = 0$ and the lowest θ . It is often referred to as the glazing incidence ray. It has the highest phase velocity along the guide.



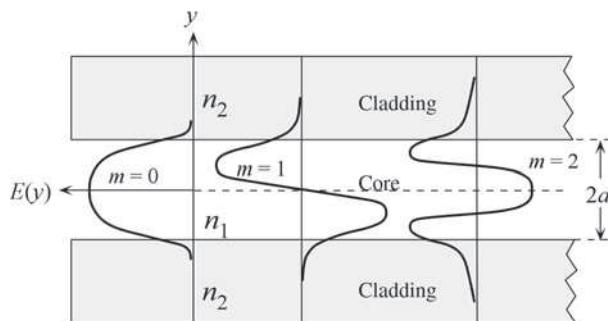


FIGURE 2.6 The electric field patterns of the first three modes ($m = 0-2$) traveling along the guide. Notice different extents of field penetration into the cladding.

We have seen that each m leads to an allowed θ_m value that corresponds to a particular traveling wave in the z -direction as described by Eq. (2.1.6) with a particular propagation constant β_m as defined by Eq. (2.1.4). Each of these traveling waves, with a distinct field pattern, $E_m(y)$, constitutes a **mode of propagation**. The integer m identifies these modes and is called the **mode number**. The light energy can be transported along the guide only by one or more of these possible modes of propagation as illustrated in Figure 2.7. Notice that the rays have been shown to penetrate the cladding, as in Figure 2.6, and reflected from an apparent plane in the cladding. Since θ_m is smaller for larger m , higher modes exhibit more reflections but they also penetrate much more into the cladding as schematically illustrated in Figure 2.7. For the lowest mode, $m = 0$, which leads to θ_m being closest to 90° , the wave is said to travel *axially*. Light that is launched into the core of the waveguide can only travel along the guide in these allowed modes specified by Eq. (2.1.3). These modes will travel down the guide at different group velocities. When they reach the end of the guide they constitute the emerging light beam. If we launch a short-duration light pulse into the dielectric waveguide, the light emerging from the other end will be a broadened light pulse because light energy would have been broken into modes and propagated at different group velocities along the guide as illustrated in Figure 2.7.² The light pulse therefore broadens as it travels along the guide; it suffers dispersion.

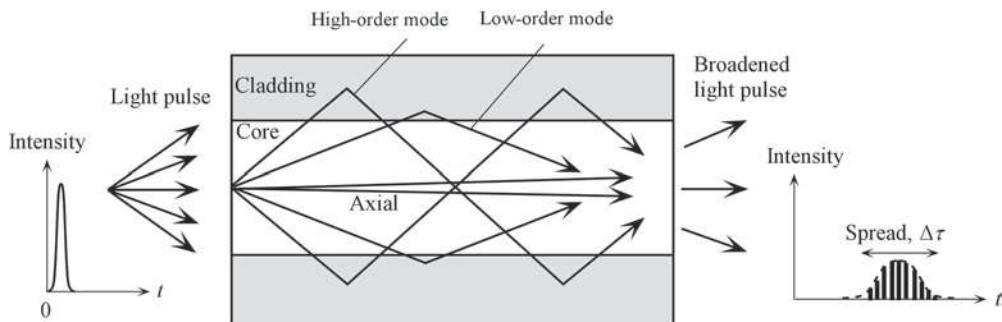


FIGURE 2.7 Schematic illustration of light propagation in a slab dielectric waveguide. Light pulse entering the waveguide breaks up into various modes which then propagate at different group velocities down the guide. At the end of the guide, the modes combine to constitute the output light pulse, which is broader than the input light pulse.

² Although the spread in Figure 2.7 has been shown as “Gaussian” looking, this is not strictly so and the exact shape of the spread cannot be derived in a simple fashion.

B. Single and Multimode Waveguides

Although the waveguide condition in Eq. (2.1.3) specifies the allowed θ_m values, θ_m must nonetheless satisfy TIR, that is, $\sin \theta_m > \sin \theta_c$. With the latter condition imposed in addition, we can only have up to a certain maximum number of modes being allowed in the waveguide. From Eq. (2.1.3) we can obtain an expression for $\sin \theta_m$ and then apply the TIR condition, $\sin \theta_m > \sin \theta_c$, to show that the mode number m must satisfy

Maximum value of m

$$m \leq \frac{2V - \phi}{\pi} \quad (2.1.7)$$

in which V , called the **V-number**, is a quantity defined by

V-number definition

$$V = \frac{2\pi a}{\lambda} (n_1^2 - n_2^2)^{1/2} \quad (2.1.8)$$

The V -number also has other names, **V-parameter**, **normalized thickness**, and **normalized frequency**. The term *normalized thickness* is more common for the present planar waveguide, whereas in optical fibers (discussed later), the V -number and normalized frequency terms are more usual. For a given free-space wavelength λ , the V -number depends on the waveguide geometry ($2a$) and waveguide properties, n_1 and n_2 . *It is therefore a characteristic parameter of the waveguide at a given wavelength.*

The question arises whether there is a value of V that makes $m = 0$ the only possibility, that is, there is only one mode propagating. Suppose that for the lowest mode, the propagation is due to a glazing incidence at $\theta_m \rightarrow 90^\circ$, then $\phi \rightarrow \pi$, and from Eq. (2.1.7), $V = (m\pi + \phi)/2$ or $\pi/2$. When $V < \pi/2$ there is only one mode propagating, which is the lowest mode with $m = 0$. From the expressions for V and ϕ , we can show that $\phi \leq 2V$ so that Eq. (2.1.7) never gives a negative m . *Thus, when $V < \pi/2$, $m = 0$ is the only possibility and only the fundamental mode ($m = 0$) propagates along the dielectric slab waveguide, which is then termed a single-mode planar waveguide.* The free-space wavelength λ_c that leads to $V = \pi/2$ in Eq. (2.1.8) is the **cutoff wavelength**, and above this wavelength, $\lambda > \lambda_c$, only one-mode, the fundamental mode, will propagate.

C. TE and TM Modes

We have shown that, for a particular mode, the variation of the field intensity along y , $E_m(y)$, is harmonic. Figures 2.8 (a) and (b) consider two of the possibilities for the electric field direction of a wave traveling toward the core-cladding boundary.

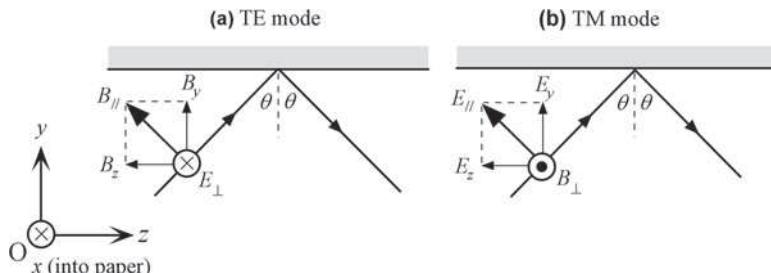


FIGURE 2.8 Possible modes can be classified in terms of (a) transverse electric (TE) and (b) transverse magnetic field (TM). Plane of incidence is the paper.

- (a) The electric field is perpendicular to the plane of incidence (plane of paper) as indicated by E_{\perp} , and shown in Figure 2.8 (a). E_{\perp} is along x , so that $E_{\perp} = E_x$.
- (b) The magnetic field is perpendicular to the plane of incidence as indicated by B_{\perp} , as shown in Figure 2.8 (b). In this case the electric field is parallel to the plane of incidence and is denoted by E_{\parallel} .

Any other field direction (perpendicular to the path of the ray) can be resolved to have electric field components along E_{\parallel} and E_{\perp} . These two fields experience different phase changes, ϕ_{\parallel} and ϕ_{\perp} , and consequently require different angles θ_m to propagate along the guide. We therefore have a different set of modes for E_{\parallel} and E_{\perp} . The modes associated with E_{\perp} (or E_x) are termed **transverse electric field modes**, denoted by TE_m , because E_{\perp} is actually perpendicular to the direction of propagation, z .

The modes associated with the E_{\parallel} field have a magnetic field B_{\perp} that is perpendicular to the direction of propagation and are termed **transverse magnetic field modes**, denoted by TM_m . It is interesting that E_{\parallel} has a field component parallel to the z -axis, shown as E_z , which is parallel to the direction of propagation. It is apparent that E_z is a propagating longitudinal electric field. In free space, it is *impossible* for such a longitudinal field to exist but within an optical guide, due to the interference phenomena, it is indeed possible to have a longitudinal field. Similarly, the TE modes, those with B_{\parallel} , have a magnetic field along z that propagates along the z -direction as a longitudinal wave.

The phase change ϕ that accompanies TIR depends on the polarization of the field and is different for E_{\parallel} and E_{\perp} . The difference, however, is negligibly small for $n_1 - n_2 \ll 1$ and thus the waveguide condition and the cutoff condition can be taken to be identical for both TE and TM modes for many practical cases.

EXAMPLE 2.1.1 Waveguide modes

Consider a planar dielectric waveguide with a core thickness 20 μm , $n_1 = 1.455$, $n_2 = 1.440$. The light to be guided has a wavelength of 900 nm. Given the waveguide condition in Eq. (2.1.3) and the expression for ϕ in TIR for the TE mode (Chapter 1)

$$\tan\left(\frac{1}{2}\phi_m\right) = \frac{\left[\sin^2\theta_m - \left(\frac{n_2}{n_1}\right)^2\right]^{1/2}}{\cos\theta_m}$$

using a graphical solution, find the allowed angles θ_m for all the modes. What is your conclusion?

Solution

The waveguide condition in Eq. (2.1.3) can be divided by 2 to obtain

$$\frac{1}{2}(2a)k_1 \cos\theta_m - \frac{1}{2}m\pi = \frac{1}{2}\phi_m$$

and taking the tangent of both sides,

$$\tan\left(ak_1 \cos\theta_m - m\frac{\pi}{2}\right) = \frac{\left[\sin^2\theta_m - \left(\frac{n_2}{n_1}\right)^2\right]^{1/2}}{\cos\theta_m} = f(\theta_m) \quad (2.1.9)$$

The left-hand side (LHS) simply reproduces itself whenever $m = 0, 2, 4, \dots$ even integer. It becomes a cotangent function whenever $m = 1, 3, \dots$ odd integer. The solutions therefore fall into

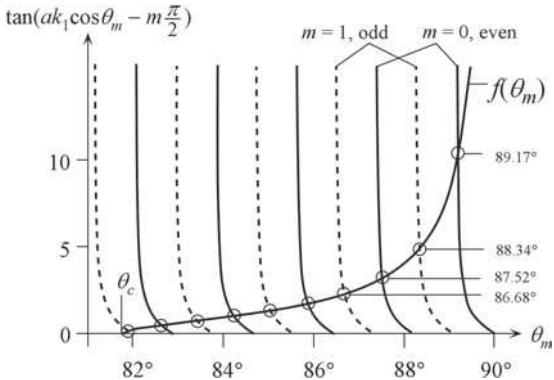


FIGURE 2.9 Modes in a planar dielectric waveguide can be determined by plotting the LHS and the RHS of Eq. (2.1.9).

odd and even m categories. Figure 2.9 shows the right-hand side (RHS), $f(\theta_m)$ vs. θ_m as well as the LHS, $\tan(ak_1 \cos \theta_m - m\frac{\pi}{2})$. Since the critical angle, $\theta_c = \arcsin(n_2/n_1) = 81.77^\circ$, we can only find solutions in the range $\theta_m = 81.77^\circ - 90^\circ$. For example, the intersections for $m = 0$ and $m = 1$ are at $89.17^\circ, 88.34^\circ$. The values of θ_m from the intersection points shown in Figure 2.9 are summarized in Table 2.1. We can also calculate the penetration δ_m of the field into the cladding using (Chapter 1)

$$\frac{1}{\delta_m} = \alpha_m = \frac{2\pi n_2 \left[\left(\frac{n_1}{n_2} \right)^2 \sin^2 \theta_m - 1 \right]^{1/2}}{\lambda} \quad (2.1.10)$$

Attenuation
in cladding

Using Eqs. (2.1.9) and (2.1.10) we find the mode angles θ_m and corresponding penetration depths δ_m of the field into the cladding shown in Table 2.1. The highest mode (highest m) has substantial field penetration into the cladding.

TABLE 2.1 Mode m , incidence angle θ_m , and penetration δ_m from solving Eq. (2.1.9) for a planar dielectric waveguide with $d = 2a = 20 \mu\text{m}$, $n_1 = 1.455$, $n_2 = 1.440$

m	0	1	2	3	4	5	6	7	8	9
θ_m	89.2°	88.3°	87.5°	86.7°	85.9°	85.0°	84.2°	83.4°	82.6°	81.9°
δ_m (μm)	0.691	0.702	0.722	0.751	0.793	0.866	0.970	1.15	1.57	3.83

It is apparent that $f(\theta_m)$ intersects the LHS tangent function once for each choice of m until $\theta_m \leq \theta_c$. There are 10 modes.

An accurate solution of Eq. (2.1.9) would show that the fundamental mode angle for the TE mode is actually 89.172° . If we were to use the phase change ϕ_m for the TM mode from Chapter 1, we would find an angle 89.170° , which is almost identical to ϕ_m for the TE mode. (The difference is in the third decimal.)

EXAMPLE 2.1.2 V-number and the number of modes

Using Eq. (2.1.7), estimate the number of modes that can be supported in a planar dielectric waveguide that is $100 \mu\text{m}$ wide, and has $n_1 = 1.490$ and $n_2 = 1.470$ at the free-space source wavelength (λ), which is $1 \mu\text{m}$. Compare your estimate with the formula

Number
of modes

$$M = \text{Int}\left(\frac{2V}{\pi}\right) + 1 \quad (2.1.11)$$

in which $\text{Int}(x)$ is the integer function; it drops the decimal fraction of x .

Solution

The phase change ϕ on TIR cannot be more than π so ϕ/π is less than 1. For a multimode guide ($V \gg 1$), we can write Eq. (2.1.7) as

$$m \leq \frac{2V - \phi}{\pi} \approx \frac{2V}{\pi}$$

We can calculate V since we are given $a = 50 \mu\text{m}$, $n_1 = 1.490$, $n_2 = 1.460$, and $\lambda = 1 \times 10^{-6} \text{ m}$ or $1 \mu\text{m}$,

$$V = (2\pi a/\lambda)(n_1^2 - n_2^2)^{1/2} = 76.44$$

Then, $m \leq 2(76.44)/\pi = 48.7$, or $m \leq 48$. There are about 49 modes, because we must also include $m = 0$ as a mode. Using Eq. (2.1.11),

$$M = \text{Int}\left[\frac{2(76.44)}{\pi}\right] + 1 = 49$$

EXAMPLE 2.1.3 Mode field width, $2w_o$

The field distribution along y penetrates into the cladding as illustrated in Figure 2.5. The extent of the electric field across the guide is therefore more than $2a$. Within the core, the field distribution is harmonic whereas from the boundary into the cladding, the field is due to the evanescent wave and it decays exponentially according to

$$E_{\text{cladding}}(y') = E_{\text{cladding}}(0) \exp(-\alpha_{\text{cladding}} y')$$

in which $E_{\text{cladding}}(y')$ is the field in the cladding at a position y' measured from the boundary and α_{cladding} is the decay constant (or attenuation) for the evanescent wave in medium 2, which is given by (Chapter 1)

$$\alpha_{\text{cladding}} = \frac{2\pi n_2}{\lambda} \left[\left(\frac{n_1}{n_2} \right)^2 \sin^2 \theta_i - 1 \right]^{1/2}$$

where λ is the wavelength in free space. For the axial mode, we can take the approximation $\theta_i \rightarrow 90^\circ$

$$\alpha_{\text{cladding}} = \frac{2\pi n_2}{\lambda} \left[\left(\frac{n_1}{n_2} \right)^2 \sin^2 \theta_i - 1 \right]^{1/2} \approx \frac{2\pi}{\lambda} (n_1^2 - n_2^2)^{1/2} = \frac{V}{a}$$

The field in the cladding decays by a factor of e^{-1} when $y' = \delta = 1/\alpha_{\text{cladding}} = a/V$. The extent of the field in the cladding is about δ . The total extent of the field across the whole guide is therefore $2a + 2\delta$, which represents the *mode field width* for the fundamental mode in the present planar dielectric waveguide, and denoted by $2w_o$. Thus,

$$2w_o \approx 2a + 2 \frac{a}{V}$$

that is,

$$2w_o \approx 2a \left[\frac{(V + 1)}{V} \right] \quad (2.1.12) \quad \begin{matrix} \text{Fundamental} \\ \text{mode field} \\ \text{width} \end{matrix}$$

We note that as V increases, the width of the mode field becomes the same as the core thickness, $2a$. In single-mode operation $V < \pi/2$ and $2w_o$ is considerably larger than $2a$. In fact at $V = \pi/2$, $2w_o$ is 1.6 times $2a$. [Equation (2.1.12) is only valid for a symmetric planar dielectric waveguide.]

2.2 MODAL AND WAVEGUIDE DISPERSION IN PLANAR WAVEGUIDES

A. Waveguide Dispersion Diagram and Group Velocity

The propagating modes that exist in a slab waveguide are determined by the waveguide condition. Each choice of m from 0 to its maximum value results in one distinct solution and one possible propagation constant β_m . We see that each mode propagates with a different propagation constant even if illumination is by monochromatic radiation. The examination of Figure 2.7 gives the impression that the axial ray has the least reflections and therefore seems to arrive more quickly (sooner) than a higher-mode ray. Higher modes zigzag more along the guide and appear to have longer ray paths. However, there are two important wrong impressions with this view. First, what is important is the group velocity v_g along the guide, the velocity at which the energy or information is transported. Second is that the higher modes penetrate more into the cladding where the refractive index is smaller and the waves travel faster.

We know from Chapter 1 that the group velocity v_g depends on $d\omega/d\beta$ in which ω is the frequency³ and β is the propagation constant. For each mode m , the mode angle θ_m is determined by the *waveguide condition* that depends on the wavelength, hence on the frequency ω , and the waveguide properties (n_1 , n_2 , and a) as apparent from Eq. (2.1.3). Thus, $\theta_m = \theta_m(\omega)$ and consequently, $\beta_m = k_1 \sin \theta_m = \beta_m(\omega)$, a function of ω . The group velocity of a given mode is therefore a function of the light frequency and the waveguide properties. We see that even if the refractive index were constant (independent of frequency or wavelength), the group velocity v_g of a given mode would still depend on the frequency by virtue of the guiding properties of the waveguide structure.

From the waveguide condition, Eq. (2.1.3), given the refractive indices (n_1 and n_2) and the guide dimension (a), we can calculate β_m for each ω and for each mode m to obtain the ω vs. β_m characteristics of the waveguide, which is called the **dispersion diagram**, as shown schematically in Figure 2.10. The slope $d\omega/d\beta_m$ at any frequency ω is the group velocity v_g . All allowed propagation modes are contained within the lines with slopes c/n_1 and c/n_2 . The cutoff frequency ω_{cutoff} corresponds to the cutoff condition ($\lambda = \lambda_c$) when $V = \pi/2$ and $\omega > \omega_{\text{cutoff}}$ leads to more than one mode. Examination of Figure 2.10 leads to two immediate consequences. First, the group velocity at one frequency changes from one mode to another. Second, for a given mode it changes with the frequency.

The group velocity therefore depends not only on the frequency but also on the waveguide properties. In fact, it is possible to relate the group velocity to the V -number as discussed later.

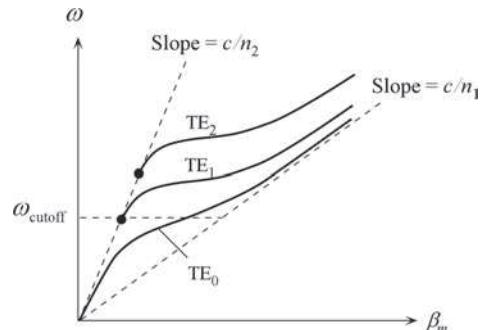


FIGURE 2.10 Schematic dispersion diagram ω vs. β for the slab waveguide for various TE_m modes. ω_{cutoff} corresponds to $V = \pi/2$. The group velocity v_g at any ω is the slope of the ω vs. β curve at that frequency.

³The term *frequency* implies angular frequency as it refers to ω .

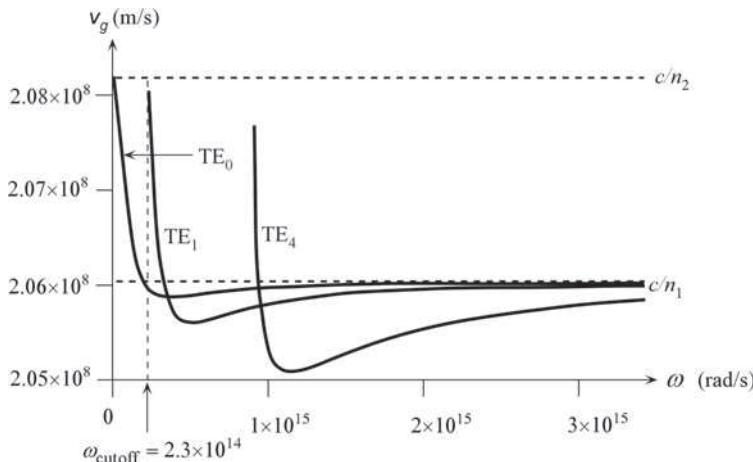


FIGURE 2.11 Group velocity v_g vs. angular frequency for three modes, $m = 0, 1$, and 4 , for a planar dielectric waveguide that has $n_1 = 1.455$, $n_2 = 1.440$, $a = 10 \mu\text{m}$. TE_0 is for $m = 0$, etc. (Results calculated by using Livemath, a math-software application.)

We can use the slope $d\omega/d\beta$ to calculate the group velocity v_g as a function of frequency ω for three different modes as in Figure 2.11. Notice that v_g starts high, with a value close to c/n_2 (the phase velocity in the cladding), on the onset of a mode, when the mode is excited, then it undershoots the c/n_1 (the core phase velocity) and eventually becomes c/n_1 at higher frequencies.

B. Intermodal Dispersion

In multimode operation, well above ω_{cutoff} , the lowest mode ($m = 0$) has the slowest group velocity, close to c/n_1 , and the highest mode has the highest group velocity. The reason is that a good portion of the field in higher modes is carried by the cladding where the refractive index is smaller (see Figure 2.6). The lowest mode is contained substantially in the core. The modes therefore take different times to travel the length of the fiber. This phenomenon is called **modal dispersion** or **intermodal dispersion**.⁴ A direct consequence of modal dispersion is the following. If a short-duration light pulse signal is coupled into the dielectric waveguide, then this pulse will travel along the guide via the excitation of various allowed modes. These modes will travel at different group velocities. The reconstruction of the light pulse at the receiving end from the various modes will result in a *broadened signal* as illustrated in Figure 2.7. It is clear that in a single-mode waveguide in which only one mode can propagate (the $m = 0$ mode), there will be no modal dispersion.

To evaluate the modal dispersion of a signal along the waveguide, we need to consider the shortest and longest times required for the signal to traverse the waveguide, which is tantamount to identifying the slowest and fastest modes excited in terms of their group velocities. If $\Delta\tau$ is the propagation time difference between the fastest and slowest modes over a distance L , then modal dispersion is defined by

$$\Delta\tau = \frac{L}{v_{g\min}} - \frac{L}{v_{g\max}} \quad (2.2.1)$$

Modal dispersion

in which $v_{g\min}$ is the minimum group velocity of the slowest mode and $v_{g\max}$ is the maximum group velocity of the fastest mode.

⁴ Also simply called **intermode dispersion**.

The lowest-order mode ($m = 0$), when $\omega > \omega_{\text{cutoff}}$, as apparent in Figures 2.10 and 2.11, has a group velocity $v_{g\min} \approx C/n_1$. The fastest propagation corresponds to the highest-order mode, which has a group velocity very roughly c/n_2 . Thus, approximately,

Modal dispersion

$$\frac{\Delta\tau}{L} \approx \frac{n_1 - n_2}{c} \quad (2.2.2)$$

Equation (2.2.2) considers only two extreme modes, the lowest and the highest, and does not consider whether some intermediate modes can have group velocities falling outside the range c/n_1 to c/n_2 (consider the slope $d\omega/d\beta_m$ in Figure 2.10 and notice that the group velocity can undershoot the lower value c/n_1). Neither does it consider how the light energy is proportioned between various modes. Taking $n_1 = 1.480$ (core) and $n_2 = 1.460$ (cladding), we find $\Delta\tau/L \approx 6.7 \times 10^{-11} \text{ s m}^{-1}$ or 67 ns km^{-1} . In general, intermodal dispersion is not as high as indicated by this estimate due to “intermode coupling,” which is the exchange of energy between the propagating modes.

The spread $\Delta\tau$ in Eq. (2.2.2) is between the two extremes of the broadened output light pulse. In optoelectronics we are frequently interested in the spread $\Delta\tau_{1/2}$ between the half intensity points that is smaller than the full width. The determination of $\Delta\tau_{1/2}$ depends on the temporal shape of the output light pulse, but as a first-order approximation, when many modes are present, we can take $\Delta\tau_{1/2} \approx \Delta\tau$.

C. Intramodal Dispersion

Figures 2.10 and 2.11 show that the lowest mode ($m = 0$) has a group velocity that depends on the frequency ω or wavelength λ . Thus, even if we use the guide in single-mode operation, as long as the excitation source has a finite spectrum, it will contain various frequencies (there is no such thing as a perfect monochromatic light wave). These frequencies will then travel with different group velocities, as apparent in Figure 2.11, and hence arrive at different times. The longer the wavelength (lower the frequency), the greater the penetration of the field into the cladding, as illustrated in Figure 2.12. Thus, a greater portion of the light energy is carried by the cladding in which the phase velocity is higher. Longer wavelengths propagate faster, even though by the same mode. This is called **waveguide dispersion** inasmuch as it results from the guiding properties of the dielectric structure and it has nothing to do with the frequency (or wavelength) dependence of the refractive index. Since increasing the wavelength decreases the V -number, a characteristic property of the guide, the dispersion can also be stated as due to the wavelength dependence of the V -number. There is no simple calculation for the waveguide

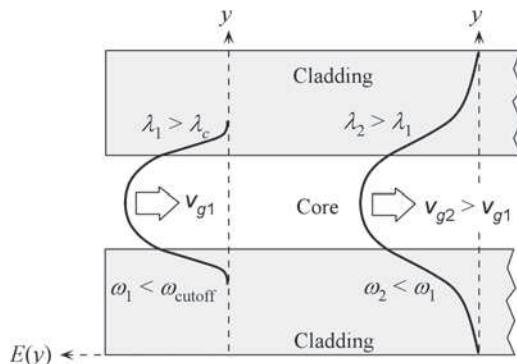
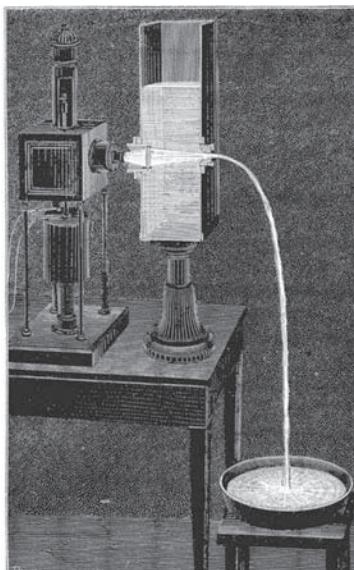


FIGURE 2.12 The electric field of TE_0 mode extends more into the cladding as the wavelength increases. As more of the field is carried by the cladding, the group velocity increases.

dispersion because we have to incorporate the spectrum of the input light and calculate the group velocity from the dispersion diagram in Figure 2.10. (This is discussed later in this chapter.)

The refractive index of the guide material will also depend on the wavelength and thus modify the ω vs. β_m behavior in Figure 2.10. The change in the group velocity of a given mode due to the n vs. λ dependence also gives rise to the broadening of a propagating light pulse. This is called **material dispersion**. Thus both waveguide and material dispersion act together to broaden a light pulse propagating *within* a given mode. Combined dispersion is called **intramode dispersion**.



Light is guided along a water jet as demonstrated by Jean-Daniel Colladon. This illustration was published in *La Nature, Revue des Sciences*, in 1884 (p. 325). His first demonstration was apparently around 1841 (*Comptes Rendus*, 15, 800–802, October 24, 1842). A similar demonstration was done by John Tyndall for the Royal Institution in London in his 1854 lecture. Apparently, Michael Faraday had originally suggested the experiment to John Tyndall though Faraday himself probably learned about it either from another earlier demonstration or through Jean-Daniel Colladon's publication. Although John Tyndall is often credited with the original discovery of a water-jet guiding light, Tyndall, himself, does not make that claim but neither does he attribute it to someone else.⁵ (*The fountain, courtesy of Cnum, Conservatoire Numérique des Arts et Métiers, France; Colladon's portrait, courtesy of Musée d'histoire des sciences, Genève, Switzerland.*)

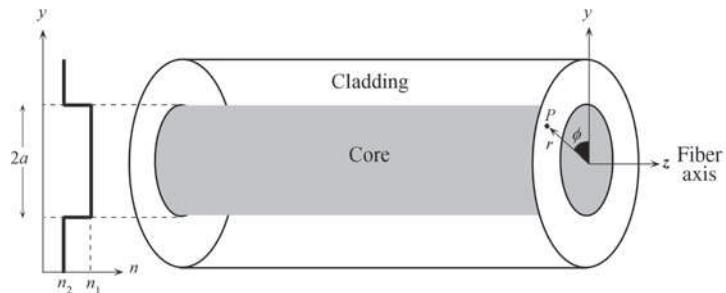
2.3 STEP-INDEX OPTICAL FIBER

A. Principles and Allowed Modes

The general ideas for guided wave propagation in a planar dielectric waveguide can be readily extended, with certain modifications, to the step-index optical fiber shown in Figure 2.13. This is essentially a cylindrical dielectric waveguide with the inner core dielectric having a refractive

⁵Jeff Hecht, “Illuminating the Origin of Light Guiding,” *Optics & Photonics News*, 10, 26, 1999, and his wonderful book *The City of Light* (Oxford University Press, 2004) describe the evolution of the optical fiber from the water jet experiments of Colladon and Tyndall to modern fibers with historical facts and references.

FIGURE 2.13 The step-index optical fiber. The central region, the core, has greater refractive index than the outer region, the cladding. The fiber has cylindrical symmetry. The coordinates r , ϕ , z are used to represent any point P in the fiber. Cladding is normally much thicker than shown.



index n_1 greater than n_2 of the outer dielectric, cladding. In practice, the cladding has a protective polymeric coating put on it as soon as the fiber is drawn, which is not shown or needed in understanding the basic fiber operation.

The **normalized index difference** Δ is defined by

Normalized
index
difference

$$\Delta = \frac{n_1 - n_2}{n_1} \quad (2.3.1)$$

For all practical fibers used in optical communications, n_1 and n_2 differ only by a small amount, typically less than a few percent so that $\Delta \ll 1$.

We recall that the planar waveguide is bounded only in one dimension so that reflections occur only in the y -direction in Figure 2.4. The requirement of constructive interference of waves then leads to the existence of distinct modes each labeled by m . The cylindrical guide in Figure 2.13 is bounded in two dimensions and the reflections occur from all the surfaces, that is, from a surface encountered along any radial direction at any angle ϕ to the y -axis in Figure 2.12. Since any radial direction can be represented in terms of x and y , reflections in both x and y directions are involved in constructive interference of waves and we therefore need two integers, l and m , to label all the allowed traveling waves or guided modes that can exist in the guide. (The description of standing waves in one dimension needs only one integer, such as m in the planar guide, whereas in two dimensions, we need two.)

We also recall that in the case of a planar waveguide we visualized a guided propagating wave in terms of rays that were zigzagging down the guide, and all these rays necessarily passed through the axial plane of the guide. Further, all waves were either TE (transverse electric) or TM (transverse magnetic). A distinctly different feature of the step-index fiber from the planar waveguide is existence of rays that zigzag down the fiber without necessarily crossing the fiber axis, so-called skew rays. A **meridional ray** enters the fiber through the fiber axis and hence also crosses the fiber axis on each reflection as it zigzags down the fiber. It travels in a plane that contains the fiber axis as illustrated in Figure 2.14 (a). On the other hand, a **skew ray** enters the fiber off the fiber axis and zigzags down the fiber without crossing the axis. When viewed looking down the fiber (its projection in a plane normal to the fiber axis), it traces out a polygon around the fiber axis as shown in Figure 2.14 (b). A skew ray therefore has a *helical path* around the fiber axis. In a step-index fiber both meridional and skew rays give rise to guided modes (allowed propagating waves) along the fiber, each with a propagation constant β along z . Guided modes resulting from meridional rays are either TE or TM type as in the case of the planar waveguide. Skew rays, on the other hand, give rise to modes that have both E_z and B_z (or H_z) components and are therefore not TE or TM waves. They are called HE or EH modes as both electric

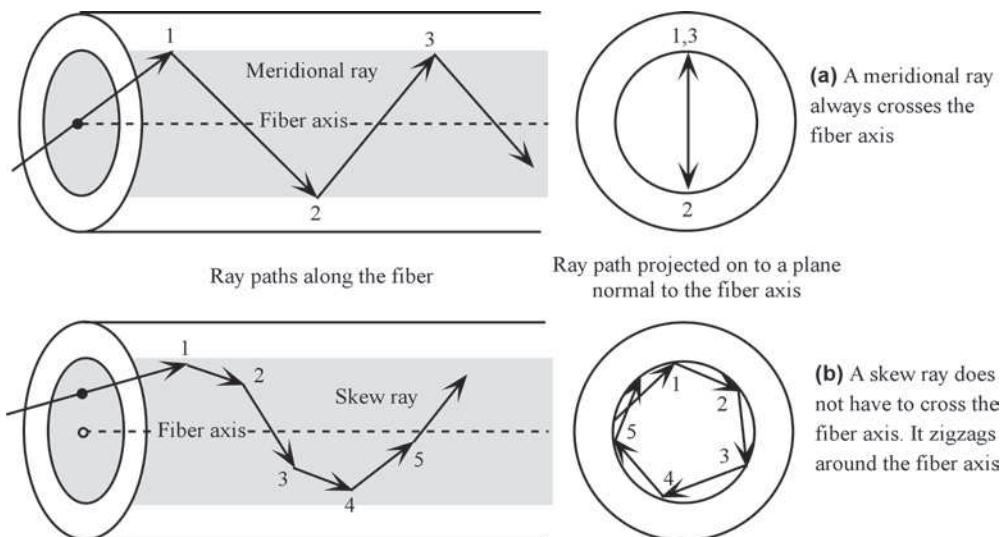


FIGURE 2.14 Illustration of the difference between (a) a meridional ray and (b) a skew ray. Numbers represent reflections of the ray.

and magnetic field can have components along z . They are called *hybrid modes*. It is apparent that guided modes in a step-index fiber cannot be as easily described as those in the planar guide.

Guided modes in a step-index fiber with $\Delta \ll 1$, called **weakly guiding fibers**, are generally visualized by traveling waves that are almost plane polarized. They have transverse electric and magnetic fields (\mathbf{E} and \mathbf{B} are perpendicular to each other and also to z), analogous to field directions in a plane wave but the field magnitudes are not constant in the plane. These waves are called **linearly polarized** (LP) and have transverse electric and magnetic field characteristics. A guided LP mode along the fiber can be represented by the propagation of an electric field distribution $E(r, \phi)$ along z . This field distribution, or pattern, is in the plane normal to the fiber axis and hence depends on r and ϕ but not on z . Further, because of the presence of two boundaries, it is characterized by two integers, l and m . The propagating field distribution in an LP mode is therefore given by $E_{lm}(r, \phi)$ and we represent the mode as LP_{lm} . Thus, an LP_{lm} mode can be described by a traveling wave along z of the form

$$E_{LP} = E_{lm}(r, \phi) \exp j(\omega t - \beta_{lm} z) \quad (2.3.2)$$

in which E_{LP} is the field of the LP mode and β_{lm} is its propagation constant along z . It is apparent that for a given l and m , $E_{lm}(r, \phi)$ represents a particular field pattern at a position z that is propagated along the fiber with an effective propagation constant, or wave vector, β_{lm} .

Figure 2.15 (a) displays the electric field pattern (E_{01}) in the **fundamental mode** of the step-index fiber that corresponds to $l = 0$ and $m = 1$, the LP_{01} mode. The field is maximum at the center of the core (or fiber axis) and penetrates somewhat into the cladding due to the accompanying evanescent wave. The extent of penetration depends on the V -number of the fiber (and hence on the wavelength). The light intensity in a mode is proportional to E^2 , which means that the intensity distribution in the LP_{01} mode has a maximum along the fiber axis as visualized in Figure 2.15 (b), brightest zone at the center and brightness decreasing toward the cladding. Figures 2.15 (c) and (d) also show the intensity distributions in the LP_{11} and LP_{21} modes. The integers l and m are related to the intensity pattern in an LP_{lm} mode. There are m number of maxima along r starting from the

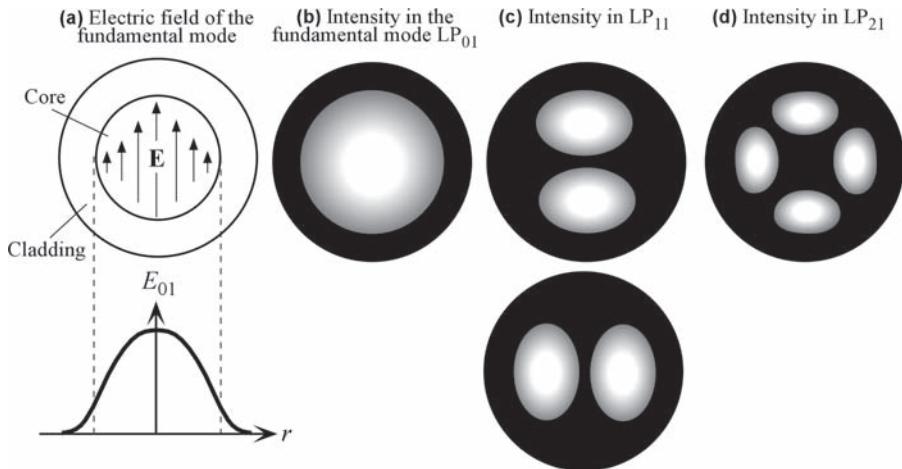


FIGURE 2.15 The electric field distribution of the fundamental mode in the transverse plane to the fiber axis z . The light intensity in the fundamental mode is greatest at the center of the fiber. Intensity patterns in LP_{01} , LP_{11} , and LP_{21} modes. (a) The field in the fundamental mode. (b)–(d) Indicative light intensity distributions in three modes— LP_{01} , LP_{11} , and LP_{21} .

core center and $2l$ number of maxima around a circumference as apparent in Figure 2.15. l is often called the **radial mode number**. Within the ray picture, l represents the extent of helical propagation, or the amount of skew ray contribution to the mode. In the fundamental mode this is zero. Further, m is directly associated with the reflection angle θ of the rays as in the planar guide.

We see from the discussions above that light travels along the fiber through various modes of propagation, with each mode having its own propagation constant β_{lm} and its own electric field pattern $E_{lm}(r, \phi)$. Each mode has its own group velocity $v_g(l, m)$ that depends on the ω vs. β_{lm} dispersion behavior. (We have to use the slope of this dispersion diagram to obtain v_g for a particular mode, l, m .) When a light pulse is fed into the fiber, it travels down the fiber through various modes. These modes propagate with different group velocities and therefore emerge at the end of the fiber with a spread of arrival times, which means that the output pulse is a broadened version of the input pulse. As in the case of the planar guide, this broadening of the light pulse is an **intermodal dispersion** phenomenon. We can, however, design a suitable fiber that allows only the fundamental mode to propagate and hence exhibits no modal dispersion.

For a step-index fiber, we define a **V-number**, or a **normalized frequency**, in a similar fashion to the planar waveguide

V-number
definition

$$V = \frac{2\pi a}{\lambda} (n_1^2 - n_2^2)^{1/2} = \frac{2\pi a}{\lambda} (2n_1 n \Delta)^{1/2} \quad (2.3.3)$$

in which a is the radius of the fiber core, λ is the free-space wavelength, n is the average refractive index of the core and cladding, that is, $n = (n_1 + n_2)/2$, and Δ is the **normalized index difference**, that is

Normalized
index
difference

$$\Delta = \frac{n_1 - n_2}{n_1} \approx \frac{n_1^2 - n_2^2}{2n_1^2} \quad (2.3.4)$$

When the V-number is less than 2.405, it can be shown that only one mode, the fundamental mode (LP_{01}) shown in Figures 2.15 (a) and (b), can propagate through the fiber core. As V is decreased further by reducing the core size, the fiber can still support the LP_{01} mode,

but the mode extends increasingly into the cladding. The finite cladding size may then result in some of the power in the wave being lost. A fiber that is designed (by the choice of a and Δ) to allow only the fundamental mode to propagate at the required wavelength is called a **single-mode fiber (SMF)**. Typically, single-mode fibers have a much smaller core radius and a smaller Δ than multimode fibers (MMFs). If the wavelength λ of the source is reduced sufficiently, a single-mode fiber will become multimode as V will exceed 2.405; higher modes will also contribute to propagation. The cutoff wavelength λ_c above which the fiber becomes single mode is given by

$$V_{\text{cutoff}} = \frac{2\pi a}{\lambda_c} (n_1^2 - n_2^2)^{1/2} = 2.405 \quad (2.3.5)$$

Single-mode cutoff wavelength

When the V -parameter increases above 2.405, the number of modes rises sharply. A good approximation to the number of modes M in a step-index multimode fiber is given by

$$M \approx \frac{V^2}{2} \quad (2.3.6)$$

Number of modes when $V \gg 2.405$

The effect of varying the various physical parameters of a step-index optical fiber on the number of propagating modes can be readily deduced from the V -number in Eq. (2.3.3). For example, increasing the core radius (a) or its refractive index (n_1) increases the number of modes. On the other hand, increasing the wavelength or the cladding refractive index (n_2) decreases the number of modes. The cladding diameter does not enter the V -number equation and so it may be thought that it plays no significant role in wave propagation. In multimode fibers, light propagates through many modes and these are all mainly confined to the core. In a step-index fiber, the field of the fundamental mode penetrates into the cladding as an evanescent wave traveling along the boundary. If the cladding is not sufficiently thick, this field will reach the end of the cladding and escape, resulting in intensity loss. For example, typically, the cladding diameter for a single-mode step-index fiber is at least 10 times the core diameter.



Narinder Singh Kapany was born in Punjab in India, studied at the Agra University, and then obtained his PhD from the Imperial College of Science and Technology, University of London, in 1955. He held a number of key positions in both academia and industry, including a Regents Professor at the University of California, Berkeley; the University of California, Santa Cruz (UCSC); and the Director of the Center for Innovation and Entrepreneurial Development at UCSC. He made significant contributions to optical glass fibers starting in the 1950s, and essentially coined the term *fiber optics* in the 1960s. His book *Fibre Optics: Principles and Applications*, published in 1967, was the first on optical fibers. (Courtesy of Dr. Narinder S. Kapany.)

EXAMPLE 2.3.1 A multimode fiber

Calculate the number of allowed modes in a multimode step-index fiber that has a core of refractive index of 1.468 and diameter 100 μm, and a cladding of refractive index of 1.447 if the source wavelength is 850 nm.

Solution

The V -number for this fiber can be calculated by substituting $a = 50 \mu\text{m}$, $\lambda = 0.850 \mu\text{m}$, $n_1 = 1.468$, and $n_2 = 1.447$ into the expression for the V -number

$$V = (2\pi a/\lambda)(n_1^2 - n_2^2)^{1/2} = (2\pi 50/0.850)(1.468^2 - 1.447^2)^{1/2} = 91.44.$$

Since $V \gg 2.405$, the number of modes is

$$M \approx V^2/2 = (91.44)^2/2 = 4181$$

which is large. There are numerous modes.

EXAMPLE 2.3.2 A single-mode fiber

What should be the core radius of a single-mode fiber that has a core of $n_1 = 1.4532$, cladding of $n_2 = 1.4483$, and is to be used with a source wavelength of 1.55 μm? What is the V -number of a fiber that has a diameter of 8.2 μm and is operating at 1.55 μm?

Solution

Single-mode propagation is achieved when $V \leq 2.405$. Thus we need

$$V = (2\pi a/\lambda)(n_1^2 - n_2^2)^{1/2} \leq 2.405$$

or

$$[2\pi a/(1.55 \mu\text{m})](1.4532^2 - 1.4483^2)^{1/2} \leq 2.405$$

which gives $a \leq 4.98 \mu\text{m}$ or a core diameter of 9.96 μm. As suspected, this is rather thin for easy coupling of the fiber to a light source or to another fiber, and special coupling techniques must be used. We also note that the size of a is on the same order of magnitude as the wavelength, which means that the geometric ray picture, strictly, cannot be used to describe light propagation. Typical single-mode fibers operating at 1.55 μm have a core diameter of 8–9 μm. (Corning's standard single-mode fiber SMF-28e⁺ has a core diameter of 8.2 μm and a cutoff at 1.260 μm.)

For a fiber that has a diameter of 8.2 μm at $\lambda = 1.55 \mu\text{m}$

$$V = (2\pi a/\lambda)(n_1^2 - n_2^2)^{1/2} = [2\pi(4.1 \mu\text{m})/(1.55 \mu\text{m})](1.4532^2 - 1.4483^2)^{1/2} = 1.982$$

which clearly is an SMF since $V < 2.405$.

B. Mode Field Diameter

As apparent in Figure 2.15, the field of the fundamental mode penetrates into the cladding. The **mode field diameter (MFD)** (denoted as $2w$) is the extent of the field distribution across an optical guide in the *fundamental mode*. The field penetrates the *cladding* so that the mode field diameter is not simply the diameter of the core. We need to know the optical power or field squared, $E(r)^2$, distribution across the fiber as a function of the radial distance r from the fiber axis.

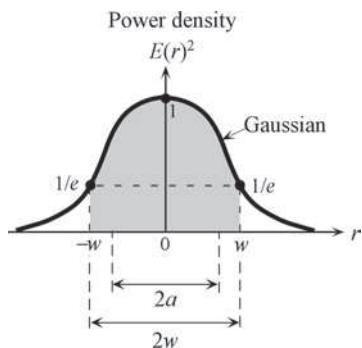


FIGURE 2.16 The power density (proportional to E^2) distribution across a single-mode fiber. $2w$ is the mode field diameter.

It is quite common in practice to represent the field distribution across the fiber as a *Gaussian* function, as shown in Figure 2.16, which can then be used to define an MFD, that is⁶

$$E(r)^2 = E(0)^2 \exp[-(r/w)^2] \quad (2.3.7) \quad \begin{matrix} \text{Fundamental} \\ \text{mode field} \end{matrix}$$

where $E(0)$ is the field at the center. At $r = w$, the power has dropped by $1/e$ and the field by $1/e^2$. One useful approximation for w in terms of the V -number is the **Marcuse MFD equation**⁷

$$2w = 2a(0.65 + 1.619V^{-3/2} + 2.879V^{-6}); \quad 0.8 < V < 2.5 \quad (2.3.8) \quad \begin{matrix} \text{Mode} \\ \text{field} \\ \text{diameter} \end{matrix}$$

Since V is typically 2, the MFD is about 1.27 times the core diameter. In the range $1.6 < V < 2.4$, the MFD can be calculated by an even simpler equation

$$2w \approx \frac{2.6(2a)}{V} \quad (2.3.9) \quad \begin{matrix} \text{Mode field} \\ \text{diameter in} \\ 1.6 < V < 2.4 \end{matrix}$$

Clearly, as V decreases, the field penetrates more into the cladding; the MFD is wider.

EXAMPLE 2.3.3 Mode field diameter

Calculate the cutoff wavelength for single-mode operation for a fiber that has a core with diameter of $8.2 \mu\text{m}$, a refractive index of 1.4532, and a cladding of refractive index of 1.4485. What is the V -number and the mode field diameter for operation at $\lambda = 1.31 \mu\text{m}$?

Solution

For single-mode operation

$$V = (2\pi a/\lambda)(n_1^2 - n_2^2)^{1/2} < 2.405$$

Substituting for a , n_1 , and n_2 and rearranging we get

$$\lambda > [2\pi(4.1 \mu\text{m})(1.4532^2 - 1.4485^2)^{1/2}]/2.405 = 1.251 \mu\text{m}$$

⁶The actual mathematical approach involves Bessel functions, which are hard to manipulate. The Gaussian representation in Eq. (2.3.7) turns out to provide a good fit, and is widely used.

⁷This is known as Marcuse's MFD equation; D. Marcuse, *Bell Syst. Tech. J.*, 56, 703, 1977.

Wavelengths shorter than $1.251 \mu\text{m}$ will result in a multimode propagation.

At $\lambda = 1.31 \mu\text{m}$

$$V = 2\pi \left[(4.1 \mu\text{m}) / (1.31 \mu\text{m}) \right] (1.4532^2 - 1.4485^2)^{1/2} = 2.30$$

The mode field diameter from Eq. (2.3.8) is

$$2w = 2(4.1) [0.65 + 1.62(2.30)^{-3/2} + 2.88(2.30)^{-6}] = 2(4.1)(1.134) = 9.30 \mu\text{m}$$

The simpler formula in Eq. (2.3.9) yields

$$2w = (2a)(2.6/V) = 2(4.1)(2.6/2.30) = 9.28 \mu\text{m}$$

whereas if we use the equation for the planar waveguide for the present fiber we get

$$2w = 2a[(V + 1)/V] = 11.8 \mu\text{m}$$

which overestimates the MFD.

C. Propagation Constant and Group Velocity

Since propagation constant β_{lm} of an LP mode depends on the waveguide properties and the source wavelength, it is convenient to describe light propagation in terms of a “normalized” propagation constant that depends only on the V -number. Given $k = 2\pi/\lambda$ (in free space) and guide indices n_1 and n_2 , the **normalized propagation constant** b is related to $\beta = \beta_{lm}$ by the definition

$$b = \frac{(\beta/k)^2 - n_2^2}{n_1^2 - n_2^2} \quad (2.3.10)$$

Normalized propagation constant

According to the definition above, the lower limit $b = 0$ corresponds to $\beta = kn_2$, propagation in the cladding material, and the upper limit $b = 1$ corresponds to $\beta = kn_1$, propagation in the core material. The dependence of b on the V -number has been calculated in the literature for various modes as shown in Figure 2.17 for a few lower-order LP modes. Such b vs. V curves are universal, that is the same, for all step-index fibers, and therefore they are very useful in predicting the nature of light propagation in such fibers. Notice that the fundamental mode (LP_{01}) exists for all V -numbers and LP_{11} is cut off at $V = 2.405$. There is a cutoff V and a corresponding cutoff wavelength for each particular LP mode higher than the fundamental mode.

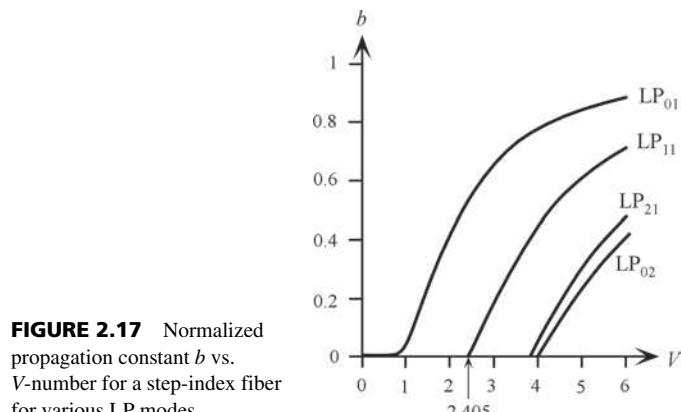


FIGURE 2.17 Normalized propagation constant b vs. V -number for a step-index fiber for various LP modes.

Given the V -parameter of the fiber we can easily find b , and hence β , for the allowed LP modes from Figure 2.17. Once we determine b from V the propagation constant is easily calculated.

$$\beta/k = [n_2^2 + b(n_1^2 - n_2^2)]^{1/2} \quad (2.3.11)$$

Propagation constant of a mode

For nearly all step-index fibers, the refractive index difference Δ is small, that is, they are weakly guiding. Equation (2.3.10) can then be approximated to a simpler form (see Question 2.14)

$$b \approx \frac{(\beta/k) - n_2}{n_1 - n_2} \quad (2.3.12)$$

Approximate normalized propagation constant

The calculation of the group velocity of the fundamental mode in a single-mode fiber involves examining the frequency dependence of b , which depends on V as in Figure 2.17. Over the range $1.5 < V < 2.5$, a simple empirical relationship gives b as⁸

$$b \approx \left(1.1428 - \frac{0.996}{V}\right)^2 \quad (2.3.13)$$

Equation (2.3.13) turns out to be quite useful in calculating the propagation characteristics of light in a single-mode fiber as shown in Example 2.3.4.

EXAMPLE 2.3.4 Group velocity and delay

Consider a single-mode fiber with core and cladding indices of 1.4480 and 1.4400, core radius of $3\text{ }\mu\text{m}$, operating at $1.5\text{ }\mu\text{m}$. Given that we can approximate the fundamental mode normalized propagation constant b by Eq. (2.3.13), calculate the propagation constant β . Change the operating wavelength to λ' by a small amount, say 0.1%, and then recalculate the new propagation constant β' . Then determine the group velocity v_g of the fundamental mode at $1.5\text{ }\mu\text{m}$, and the group delay τ_g over 1 km of fiber.

Solution

To calculate b from V in Eq. (2.3.13), we need V

$$V = (2\pi a/\lambda)(n_1^2 - n_2^2)^{1/2} = 2\pi[(3\text{ }\mu\text{m})/(1.5\text{ }\mu\text{m})](1.4480^2 - 1.4400^2)^{1/2} = 1.910088$$

Substituting the above V in Eq. (2.3.13) gives $b = 0.3860859$.

At $\lambda = 1.5\text{ }\mu\text{m}$, $k = 2\pi/\lambda = 4,188,790\text{ m}^{-1}$. In addition, $\omega = 2\pi c/\lambda = 1.255757 \times 10^{15}\text{ rad s}^{-1}$. Thus, from Eq. (2.3.11) we have

$$\begin{aligned} \beta &= k[n_2^2 + b(n_1^2 - n_2^2)]^{1/2} \\ &= (4,188,790\text{ m}^{-1})[1.4400^2 + (0.3860859)(1.4480^2 - 1.4400^2)]^{1/2} \\ &= 6,044,818\text{ m}^{-1} \end{aligned}$$

TABLE 2.2 Calculation of the group velocity at two different wavelengths

Calculation →	V	$k(\text{m}^{-1})$	$\omega(\text{rad s}^{-1})$	b	$\beta(\text{m}^{-1})$
$\lambda = 1.50000\text{ }\mu\text{m}$	1.910088	4,188,790	1.255757×10^{15}	0.3860859	6.044818×10^6
$\lambda' = 1.50150\text{ }\mu\text{m}$	1.908180	4,184,606	1.254503×10^{15}	0.3854382	6.038757×10^6

⁸See J. Senior, *Optical Fiber Communications, Principles and Practice*, 3rd Edition (Prentice-Hall, Pearson Education, 2009), Ch. 2.

Calculations are summarized in Table 2.2. We can repeat the calculations for $\lambda' = 1.5(1 + 0.001) = 1.5015\text{ }\mu\text{m}$. The group velocity is

$$v_g = \frac{\omega' - \omega}{\beta' - \beta} = \frac{(1.254503 - 1.255757) \times 10^{15}}{(6.044818 - 6.038757) \times 10^6} = 2.07 \times 10^8 \text{ m s}^{-1}$$

The group delay τ_g over 1 km is 4.83 μs . Since we were after small changes in ω and β , we kept several decimal places in the calculations.

D. Modal Dispersion in Multimode Step-Index Fibers

We know from the planar waveguide that a light pulse fed into the guide breaks into different modes and these modes propagate with different group velocities as shown in Figure 2.7, provided the guide is operating as a multimode guide (with $V > \pi/2$). Exactly the same arguments apply in the case of a multimode fiber that has $V > 2.405$, and Figure 2.7 can easily represent the situation for light propagation in a multimode fiber. Equation (2.2.2) gives the intermodal dispersion $\Delta\tau$ at the output of a planar guide of length L , that is

Modal dispersion between extreme points

$$\frac{\Delta\tau}{L} \approx \frac{n_1 - n_2}{c} = \frac{n_1\Delta}{c} \quad (2.3.14)$$

We can use the same equation to estimate the intermodal dispersion in a step-index multimode fiber, which, essentially, depends on the slowest (lowest mode) and fastest (highest mode) velocity difference, $(n_1 - n_2)/c$. A proper calculation should consider the way in which modes propagate and find the difference in the arrival times at the output, but for an MMF with many modes and for which $\Delta \ll 1$, the result leads to Eq. (2.3.14). There is, however, an important condition attached to Eq. (2.3.14). It applies between the first and last exiting modes, and therefore represents the *full spread* in the output pulse.

EXAMPLE 2.3.5 A multimode fiber and dispersion

A step-index multimode fiber for use at 850 nm has a core diameter of 100 μm and a refractive index of 1.4599. The cladding has a refractive index of 1.4565. Calculate the V -number, the number of modes and the modal dispersion per km of fiber.

Solution

The V -number of the fiber is

$$V = (2\pi a/\lambda)(n_1^2 - n_2^2)^{1/2} = [(2\pi 50)/(0.850)] [(1.4599)^2 - (1.4565)^2]^{1/2} = 36.8$$

The number of modes, $M \approx V^2/2 = 677$.

The modal dispersion is

$$\Delta\tau/L \approx (n_1 - n_2)/c = (1.4599 - 1.4565)/(3 \times 10^8 \text{ m s}^{-1}) = 11.3 \text{ ns km}^{-1}$$

Such a modal dispersion, as will be shown later, is very significant and restricts their use to low frequency work or to communications over short distances.

2.4 NUMERICAL APERTURE

Not all source radiation can be guided along an optical fiber. Only rays falling within a certain cone at the input of the fiber can normally be propagated through the fiber. Figure 2.18 shows the path of a light ray launched from the outside medium of refractive index n_0 (not necessarily air) into the fiber core. Suppose that the incidence angle at the end of the fiber core is α , and inside the waveguide the ray makes an angle θ , with the normal to the fiber axis. Then, unless the angle θ is greater than the critical angle θ_c for TIR, the ray will escape into the cladding. Thus for light propagation, the launching angle α has to be such that TIR is supported within the fiber. From Figure 2.18, it should be apparent that the maximum value of α is that which results in $\theta = \theta_c$. At the n_0-n_1 interface, Snell's law gives

$$\frac{\sin \alpha_{\max}}{\sin(\pi/2 - \theta_c)} = \frac{n_1}{n_0}$$

in which θ_c is determined by the onset of TIR, that is, $\sin \theta_c = n_2/n_1$. We can now eliminate θ_c to obtain

$$\sin \alpha_{\max} = \frac{(n_1^2 - n_2^2)^{1/2}}{n_0}$$

The **numerical aperture** (NA) is a characteristic parameter of an optical fiber defined by

$$\text{NA} = (n_1^2 - n_2^2)^{1/2} \quad (2.4.1)$$

so that in terms of NA, the **maximum acceptance angle**, α_{\max} , becomes

$$\sin \alpha_{\max} = \frac{\text{NA}}{n_0} \quad (2.4.2)$$

Definition
of numerical
aperture,
NA

Maximum
acceptance
angle

The angle $2\alpha_{\max}$ is called the **total acceptance angle** and depends on the NA of the fiber and the refractive index of the launching medium. NA is an important factor in light launching designs into the optical fiber. We should note that Eq. (2.4.2) is strictly applicable for meridional rays. Skew rays have a wider acceptance angle. Since NA is defined in terms of the refractive indices, we can relate the V-number to NA

$$V = \frac{2\pi a}{\lambda} \text{NA} \quad (2.4.3)$$

V-number
and NA

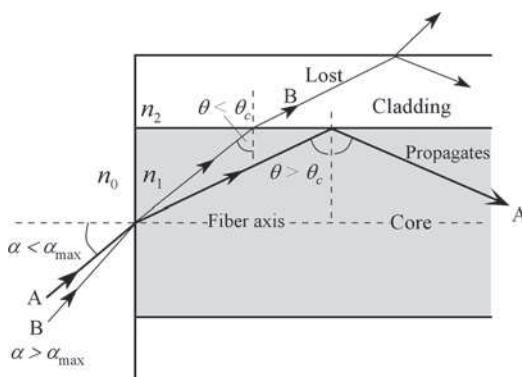


FIGURE 2.18 Maximum acceptance angle α_{\max} is that which just gives total internal reflection at the core-cladding interface, that is, when $\alpha = \alpha_{\max}$, $\theta = \theta_c$. Rays with $\alpha < \alpha_{\max}$ (e.g., ray A) experience total internal reflection at the core-cladding interface and can propagate if θ satisfies the waveguide condition. Rays with $\alpha > \alpha_{\max}$ (e.g., ray B) become refracted and penetrate the cladding and are eventually lost.

Multimode propagation in a fiber involves many modes, the majority of which are usually skew rays. The acceptance angle is larger for these skew rays than that predicted by Eq. (2.4.2), which is for meridional rays. Some authors use the NA definition based on $NA = (n_1^2 - n_2^2)^{1/2}/n_0$. In this case Eq. (2.4.2) becomes $\sin \alpha_{\max} = NA$. The preference for using the definition in Eq. (2.4.1) is that nearly all data sheets and technical specifications quote NA as defined in Eq. (2.4.1). This value then can be used in Eq. (2.4.2) with any n_0 . Further, the definition in Eq. (2.4.1) refers only to the fiber properties without any reference to the medium from which light is launched.

EXAMPLE 2.4.1 A multimode fiber and total acceptance angle

A step-index fiber has a core diameter of 100 μm and a refractive index of 1.4800. The cladding has a refractive index of 1.4600. Calculate the numerical aperture of the fiber, acceptance angle from air, and the number of modes sustained when the source wavelength is 850 nm.

Solution

The numerical aperture is

$$NA = (n_1^2 - n_2^2)^{1/2} = (1.4800^2 - 1.4600^2)^{1/2} = 0.2425$$

From $\sin \alpha_{\max} = NA/n_0 = 0.2425/1$, the acceptance angle is $\alpha_{\max} = 14^\circ$ and the total acceptance angle is 28° .

The V-number in terms of the numerical aperture can be written as

$$V = (2\pi a/\lambda)NA = [(2\pi 50)/(0.85)](0.2425) = 89.62$$

The number of modes, $M \approx V^2/2 = 4015$.

EXAMPLE 2.4.2 A single-mode fiber

A typical single-mode optical fiber has a core of diameter 8 μm and a core refractive index of 1.4600. The normalized index difference is 0.3%. The cladding diameter is 125 μm . Calculate the numerical aperture and the acceptance angle of the fiber. What is the single-mode cutoff frequency λ_c of the fiber?

Solution

The numerical aperture is

$$NA = (n_1^2 - n_2^2)^{1/2} = [(n_1 + n_2)(n_1 - n_2)]^{1/2}$$

Substituting $(n_1 - n_2) = n_1\Delta$ and $(n_1 + n_2) \approx 2n_1$, we get

$$NA \approx [(2n_1)(n_1\Delta)]^{1/2} = n_1(2\Delta)^{1/2} = 1.4600(2 \times 0.003)^{1/2} = 0.113$$

The acceptance angle is given by

$$\sin \alpha_{\max} = NA/n_0 = 0.113/1, \quad \text{or} \quad \alpha_{\max} = 6.5^\circ$$

The condition for single-mode propagation is $V \leq 2.405$, which corresponds to a minimum wavelength λ_c given by Eq. (2.4.3),

$$\lambda_c = (2\pi a NA)/2.405 = [(2\pi)(4 \mu\text{m})(0.113)]/2.405 = 1.18 \mu\text{m}$$

Excitation wavelengths shorter than 1.18 μm will result in a multimode operation.

2.5 DISPERSION IN SINGLE-MODE FIBERS

A. Material Dispersion

Since only one mode propagates in a single-mode, step-index fiber, it should be apparent that there is no intermodal dispersion of an input light pulse. This is the advantage of the single-mode fiber. However, even if light propagation occurs by a single mode, there will still be some dispersion because the input light source cannot be perfectly monochromatic and has a finite spectral width, that is, a range of λ values. For example, even excitation from a laser will result in a spectrum of source wavelengths fed into the fiber as schematically illustrated in Figure 2.19. The wavelength width $\Delta\lambda$ of the spectrum will depend on the nature of the light source but it is never zero. Each radiation with a different λ will propagate, via the fundamental mode, at a different group velocity v_g because the group index N_g of the medium (related to n_1) depends on the wavelength. The waves will therefore arrive at different times at the end of the fiber and hence will give rise to a broadened output light pulse as illustrated in Figure 2.19. The broadened light output means that we have to separate out the output pulses so that they do not overlap, which means that dispersion reduces the maximum bit rate for sending signals along a fiber. It limits the fiber bandwidth, a point that is quantified in Section 2.7.

This type of dispersion that results from the wavelength dependence of the material properties of the guide is called **material dispersion**. Its manifestation is the result of no practical light source being perfectly monochromatic so that there are waves in the guide with various wavelengths, that is, a range of λ values. The group index, N_g , for the silica glass (Chapter 1) is almost constant at wavelengths around 1.3 μm , which implies zero material dispersion around this wavelength.

Material dispersion can be evaluated by considering the dependence of the group velocity v_g on the wavelength through the N_g vs. λ behavior, and the spectrum of the light source that defines the range $\Delta\lambda$ of emitted wavelengths from the light source. We consider a very short-duration light pulse as the input signal with a spectrum of wavelengths over $\Delta\lambda$ as in Figure 2.19. Then, the output is a pulse that is broadened by $\Delta\tau$ due to the spread in the arrival times τ of these waves with different λ . $\Delta\tau$ increases with the fiber length L because slower waves fall further behind the faster waves over a longer distance. Dispersion is therefore normally expressed as spread per unit length and is given by

$$\frac{\Delta\tau}{L} = |D_m| \Delta\lambda \quad (2.5.1)$$

Material dispersion

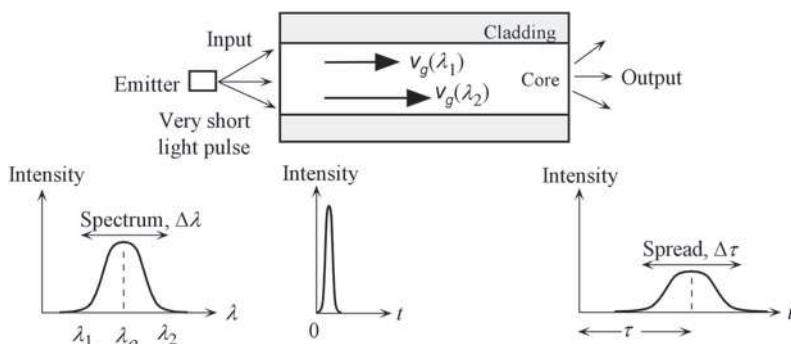


FIGURE 2.19 All excitation sources are inherently non-monochromatic and emit within a spectrum, $\Delta\lambda$, of wavelengths. Waves in the guide with different wavelengths travel at different group velocities due to the wavelength dependence of n_1 . The waves arrive at the end of the fiber at different times and hence result in a broadened output pulse.

in which D_m is a coefficient called the **material dispersion coefficient**. It is approximately given by the second derivative of the refractive index

Material dispersion coefficient

$$D_m \approx -\frac{\lambda}{c} \left(\frac{d^2 n}{d\lambda^2} \right) \quad (2.5.2)$$

Note that although D_m may be negative or positive, $\Delta\tau$ and $\Delta\lambda$ are defined as positive quantities, which is the reason for the magnitude sign in Eq. (2.5.1). To find $\Delta\tau/L$ in Eq. (2.5.1) we must evaluate D_m at the center wavelength λ_o of the input light spectrum. The dependence of D_m on λ for silica as a typical fiber core glass material is displayed in Figure 2.20. Notice that the curve passes through zero at $\lambda \approx 1.27 \mu\text{m}$. When silica (SiO_2) is doped with germania (GeO_2) to increase the refractive index for the core, the D_m vs. λ curve shifts slightly to higher wavelengths.

The transit time τ of a light pulse represents a delay of information between the output and the input. The signal delay time per unit distance, τ/L , is called the **group delay** (τ_g) and is determined by the group velocity v_g as it refers to the transit time of signals (energy). If β_{01} is the propagation constant of the fundamental mode, then by definition

Fundamental mode group delay time

$$\tau_g = \frac{1}{v_g} = \frac{d\beta_{01}}{d\omega} \quad (2.5.3)$$

which will be a function of wavelength. Material dispersion in Eq. (2.5.1) is the spread in τ_g in Eq. (2.5.3) due to the dependence of β_{01} on wavelength through N_g . The spread $\Delta\lambda$ in the input wavelengths, due to the wavelength dependence of group index N_g , leads to a spread $\Delta\tau_g$ in τ_g ,

$$\Delta\tau_g = \frac{d}{d\lambda} \left(\frac{1}{v_g} \right) \Delta\lambda = \frac{d}{d\lambda} \left(\frac{N_g}{c} \right) \Delta\lambda \approx -\frac{\lambda}{c} \left(\frac{d^2 n}{d\lambda^2} \right) \Delta\lambda$$

which leads to Eq. (2.5.2) for D_m .

B. Waveguide Dispersion

Another dispersion mechanism, called the **waveguide dispersion**, is due to the dependence of the group velocity $v_g(01)$ of the fundamental mode ($l = 1, m = 0$) on the V -number, which depends on the source wavelength λ , even if n_1 and n_2 were constant. It should be emphasized that waveguide dispersion is quite separate from material dispersion. Even if n_1 and n_2 were wavelength independent (no material dispersion), we will still have waveguide dispersion by virtue of $v_g(01)$ depending on V and V depending inversely on λ . The waveguide dispersion arises as a result of the guiding properties of the waveguide, which imposes a nonlinear $\omega-\beta_{lm}$ relationship, as discussed in Section 2.2. Therefore, a spectrum of source wavelengths will result in different V -numbers for each source wavelength and hence different propagation velocities. The result is that there will be a spread in the group delay times of the fundamental-mode waves with different λ and hence a broadening of the output light pulse as illustrated in Figure 2.19.

If we use a light pulse of very short duration as input, as in Figure 2.19, with a wavelength spectrum of $\Delta\lambda$, then the broadening or dispersion per unit length, $\Delta\tau/L$, in the output light pulse due to waveguide dispersion can be found from

Waveguide dispersion

$$\frac{\Delta\tau}{L} = |D_w| \Delta\lambda \quad (2.5.4)$$

in which D_w is a coefficient called the **waveguide dispersion coefficient**. D_w depends on the waveguide properties in a non-trivial way.

The group velocity v_g depends on $d\omega/d\beta$ and hence the group delay τ_g , inversely proportional to v_g , depends on $d\beta/d\omega$. The variation of τ_g with the wavelength, due to different excitation wavelengths being fed in, means that we are interested in $d\tau_g/d\lambda$. The next step is to consider V as depending on ω only and keeping n and Δ as wavelength (frequency) independent. We also need to substitute for β in terms of b from Eq. (2.3.12) since b vs. V is a universal curve for step-index SMFs. A proper analysis⁹ shows that the evaluation of $d^2\beta/d\omega^2$ requires that we consider $d^2\beta/dV^2$ (recall that V is the normalized frequency) and hence $d^2(bV)/dV^2$. The final result for D_w is

$$D_w = \frac{n_2 \Delta}{c \lambda} \left[V \frac{d^2(bV)}{dV^2} \right] \quad (2.5.5)$$

Waveguide dispersion coefficient

Thus, as expected, a second differential is involved, which can be evaluated by using Figure 2.17. The variation of the term in square brackets with V has been given in the literature, and one can easily find D_w . For V -numbers in the range (see footnote 9) $1.5 < V < 2.4$,

$$D_w \approx -\frac{0.025\lambda}{a^2 c n_2} \quad (2.5.6)$$

Waveguide dispersion coefficient

where D_w is in s m^{-2} , and to convert to $\text{ps nm}^{-1} \text{km}^{-1}$, we multiply D_w by 10^6 . If λ and a are both in μm , substituting for c as well,

$$D_w(\text{ps nm}^{-1} \text{km}^{-1}) \approx -\frac{83.76 \lambda(\mu\text{m})}{[a(\mu\text{m})]^2 n_2} \quad (2.5.7)$$

Waveguide dispersion coefficient

in which μm in parentheses refers to the units of λ and a in Eq. (2.5.7).

First note that the magnitude of D_w increases with the wavelength. Second, it is apparent that D_w depends on the guide geometry through the core radius a . Figure 2.20 also shows the dependence of D_w on the wavelength for a core radius of $a \approx 4.2 \mu\text{m}$. We note that D_m and D_w have opposite tendencies with wavelength.

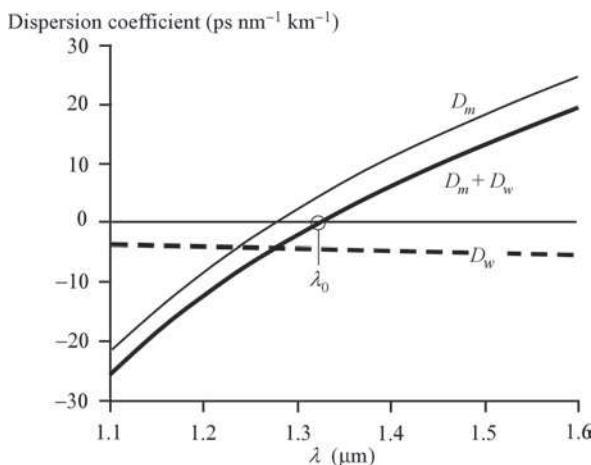


FIGURE 2.20 Material dispersion coefficient (D_m) for the core material (taken as SiO_2), waveguide dispersion coefficient (D_w) (for $a = 4.2 \mu\text{m}$), and the total or chromatic dispersion coefficient D_{ch} ($= D_m + D_w$) as a function of free-space wavelength, λ .

⁹See Question 2.18, where this is derived.

C. Chromatic Dispersion

In single-mode fibers, the dispersion of a propagating pulse arises because of the finite width $\Delta\lambda$ of the source spectrum; it is not perfectly monochromatic. Dispersion mechanism is based on the fundamental mode velocity depending on the source wavelength, λ . This type of dispersion caused by a range of source wavelengths is generally termed **chromatic dispersion** and includes both material and waveguide dispersion since both depend on $\Delta\lambda$. As a first approximation, the two dispersion effects can be simply added so that the overall dispersion per unit length becomes

Chromatic dispersion

$$\frac{\Delta\tau}{L} = |D_m + D_w|\Delta\lambda \quad (2.5.8)$$

This is shown in Figure 2.20 where $D_m + D_w$, defined as the **chromatic dispersion coefficient**, D_{ch} , passes through zero at a certain wavelength, λ_0 . For the example in Figure 2.20, chromatic dispersion is zero at around 1320 nm.

Since D_w depends on the guide geometry, as apparent in Eq. (2.5.6), it is therefore possible to shift the zero dispersion wavelength λ_0 by suitably designing the guide. For example, by reducing the core radius and increasing the core doping, λ_0 can be shifted to 1550 nm where light attenuation in the fiber is minimal. Such fibers are called **dispersion shifted fibers**.

Although chromatic dispersion, $D_m + D_w$, passes through zero, this does not mean that there would be no dispersion at all. First, we should note that $D_m + D_w$ can only be made zero for only one wavelength, λ_0 , not at every wavelength within the spectrum, $\Delta\lambda$, of the source. Since the spread in the signal delay, that is dispersion, $\Delta\tau$, along a fiber depends on $\Delta\lambda$, the linewidth of the source, we can expand $\Delta\tau$ as a Taylor series in $\Delta\lambda$. Consider the expansion at $\lambda = \lambda_0$ where $D_{ch} = 0$. The first term with $\Delta\lambda$ would have $d\Delta\tau/d\lambda$ as a coefficient, that is D_{ch} , and at λ_0 this will be zero. But, the second term with $(\Delta\lambda)^2$ has a second differential, $d^2\Delta\tau/d\lambda^2$ or $dD_{ch}/d\lambda$, that is finite. Thus, the dispersion at λ_0 would be controlled by the slope S_0 of D_{ch} vs. λ curve at λ_0 . A proper mathematical analysis gives the dispersion at λ_0 as (Question 2.19)

Chromatic dispersion at λ_0 , $D_{ch} = 0$

$$\Delta\tau = \frac{L}{2} S_0(\Delta\lambda)^2 \quad (2.5.9)$$

where S_0 , the slope of D_{ch} vs. λ at $\lambda = \lambda_0$ and is usually given as $\text{ps nm}^{-2} \text{km}^{-1}$. For the standard single-mode fiber in Figure 2.20, S_0 is about $0.9 \text{ ps nm}^{-2} \text{km}^{-1}$ so that a laser source with a linewidth of 2 nm results in a dispersion of 1.8 ps km^{-1} .

D. Profile and Polarization Dispersion Effects

Although material and waveguide dispersion are the chief sources of broadening of a propagating light pulse there are other dispersion effects. There is an additional dispersion mechanism called **profile dispersion** that arises because the group velocity, $v_g(01)$, of the fundamental mode also depends on the refractive index difference Δ , that is, $\Delta = \Delta(\lambda)$. If Δ changes with wavelength, then different wavelengths from the source would have different group velocities and experience different group delays leading to pulse broadening. It is part of chromatic dispersion because it depends on the input spectrum $\Delta\lambda$. In the case of profile dispersion

Profile dispersion

$$\frac{\Delta\tau}{L} = |D_p|\Delta\lambda \quad (2.5.10)$$

in which D_p is the **profile dispersion coefficient**, which can be calculated (though not in any simple fashion; see Question 2.19). Typically, D_p is less than $1 \text{ ps nm}^{-1} \text{ km}^{-1}$ and thus negligible compared with D_w . The overall chromatic dispersion coefficient then becomes

$$D_{ch} = D_m + D_w + D_p \quad (2.5.11) \quad \text{Chromatic dispersion}$$

It should be mentioned that the reason Δ exhibits a wavelength dependence is due to material dispersion characteristic, that is, n_1 vs. λ and n_2 vs. λ behavior, so that, in reality, profile dispersion originates from material dispersion though its effects appear through the wavelength dependence of $\Delta = \Delta(\lambda)$.

As apparent from Eq. (2.5.11), the sign of the dispersion coefficient is important, since individual dispersion coefficients, D_m , D_w , and D_p , are added to get the chromatic dispersion D_{ch} . In Figure 2.20, D_m pass through zero at λ_0 . The group index N_g of glass, from Chapter 1, has a negative slope for $\lambda < \lambda_0$ and positive for $\lambda > \lambda_0$. Consider sending two pulses along the fiber, one at a wavelength λ_1 and the other at $\lambda_2 > \lambda_1$; both less than λ_0 . The group velocity v_{g1} at λ_1 is smaller than v_{g2} at λ_2 . The pulse at λ_2 comes out first and then later the pulse at λ_1 . This situation in which longer wavelengths travel faster and come out first is assigned *negative dispersion*.¹⁰ When $\lambda > \lambda_0$, the situation for D_m is reversed, and longer wavelengths come out later; D_m is then *positive*. In the case of D_w for a simple step-index fiber, its sign is always negative as in Eq. (2.5.6)—longer wavelengths travel faster. As apparent in Figure 2.12, at longer wavelengths, a greater portion of the fundamental mode propagates in the cladding.

Polarization dispersion arises when the fiber is not perfectly symmetric and homogeneous, that is, the refractive index is *not isotropic*. When the refractive index depends on the direction of the electric field, the propagation constant of a given mode depends on its polarization. Due to various variations in the fabrication process, such as small changes in the glass composition, geometry, and induced local strains (either during fiber drawing or cabling), the refractive indices n_1 and n_2 may not be isotropic. Suppose that n_1 has the values n_{1x} and n_{1y} when the electric field is parallel to the x - and y -axes, respectively, as shown in Figure 2.21. Then the

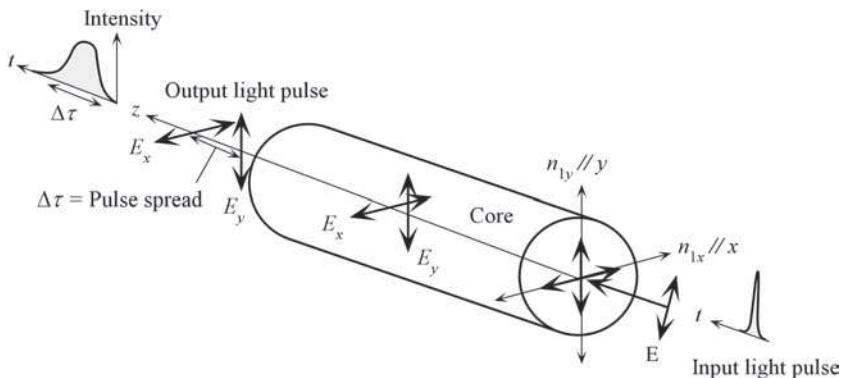


FIGURE 2.21 The core refractive index has different values along two orthogonal directions corresponding to two orthogonal electric field oscillation directions (polarizations). Take x - and y -axes along these directions. An input light will travel along the fiber with E_x and E_y polarizations having different group velocities and hence arrive at the output at different times, which leads to the dispersion of the light pulse.

¹⁰ Clearly, the assignment of positive or negative sign must be somewhat arbitrary, but the definition here is the one most commonly used.

propagation constant for fields along x -and y -would be different, $\beta_x(01)$ and $\beta_y(01)$, leading to different group delays and hence dispersion even if the source wavelength is monochromatic. The situation in reality is more complicated because n_{1x} and n_{1y} would vary along the fiber length and there will be interchange of energy between these modes as well. The final dispersion nonetheless depends on the extent of anisotropy, $n_{1x} - n_{1y}$, which is kept minimum by various manufacturing procedures (such as rotating the fiber during drawing). The term **polarization mode dispersion** (PMD) represents this type of dispersion.

When an optically anisotropic medium exhibits different refractive indices in two different orthogonal directions of field oscillations, it is said to be **birefringent**.¹¹ As mentioned, stresses experienced by the fiber during and after manufacturing can easily induce different refractive indices, n_{1x} and n_{1y} , in different directions in a fiber, that is induced birefringence, which would lead to a finite PMD. The dispersion does not scale linearly with fiber length L . It scales very roughly as $L^{1/2}$ and hence PMD is often quoted as $\text{ps km}^{-1/2}$. While it is possible to minimize or even compensate for chromatic dispersion, the control of PMD involves a high quality of control in the manufacturing process. The PMD of a fiber that is cabled can be significantly higher than when it is first manufactured. The effect of PMD on the fiber bandwidth is one of the ultimate hurdles in long-haul communications, and manufacturers are well aware of the importance of keeping PMD as low as possible, especially in cabled fibers. Its value is typically a fraction of a $\text{ps km}^{-1/2}$. For example, many commercial SMFs are reported as having a PMD less than $0.1 \text{ ps km}^{-1/2}$ for an individual fiber. The dispersion $\Delta\tau$ due to PMD can be written as

Polarization
mode
dispersion

$$\Delta\tau = D_{\text{PMD}} L^{1/2} \quad (2.5.12)$$

Note that D_{PMD} has different units than D_{ch} , and does not depend on the spectral linewidth of the input source.

EXAMPLE 2.5.1 Material dispersion

By convention, the width $\Delta\lambda$ of the wavelength spectrum of the source and the dispersion $\Delta\tau$ refer to half-intensity widths and not widths from one extreme end to the other. Suppose that $\Delta\tau_{1/2}$, called the **linewidth**, is the width of intensity vs. wavelength spectrum between the half intensity points and $\Delta\lambda_{1/2}$ is the width of the output light intensity vs. time signal between the half-intensity points.

Estimate the material dispersion effect per km of silica fiber operated from a light-emitting diode (LED) emitting at $1.55 \mu\text{m}$ with a linewidth of 100 nm. What is the material dispersion effect per km of silica fiber operated from a laser diode emitting at the same wavelength with a linewidth of 2 nm?

Solution

From Figure 2.20, at $1.55 \mu\text{m}$, the material dispersion coefficient $D_m = 22 \text{ ps nm}^{-1} \text{ km}^{-1}$. This is given as per nm of spectral linewidth and per km of fiber length. For the LED, $\Delta\lambda_{1/2} = 100 \text{ nm}$ and

$$\Delta\tau_{1/2} \approx L |D_m| \Delta\lambda_{1/2} \approx (1 \text{ km})(22 \text{ ps nm}^{-1} \text{ km}^{-1})(100 \text{ nm}) = 2200 \text{ ps}, \text{ or } 2.2 \text{ ns}$$

For the laser diode, $\Delta\lambda_{1/2} = 2 \text{ nm}$, and

$$\Delta\tau_{1/2} \approx L |D_m| \Delta\lambda_{1/2} = (1 \text{ km})(22 \text{ ps nm}^{-1} \text{ km}^{-1})(2 \text{ nm}) = 44 \text{ ps}, \text{ or } 0.044 \text{ ns}$$

¹¹Birefringence is discussed in Chapter 6. It is essentially a phenomenon in which a propagating light wave in an anisotropic optical medium experiences two different refractive indices that depend on the electric field directions and the direction of propagation of the light wave.

There is clearly a big difference between the dispersion effects of the two sources. The total dispersion, however, as indicated by Figure 2.20, will be less. Indeed, if the fiber is properly dispersion shifted so that $D_m + D_w = 0$ at $1.55 \mu\text{m}$, then dispersion due to excitation from a typical laser diode will be a few picoseconds per km (but not zero). Further, D_m vs. λ behavior depends on the composition of the core material and can be readily calculated from the wavelength dependence of n_1 using Eq. (2.5.2).

EXAMPLE 2.5.2 Material, waveguide, and chromatic dispersion

Consider a single-mode optical fiber with a core of $\text{SiO}_2\text{-}13.5\%\text{GeO}_2$ (actual silica content is 86.5%) for which the material and waveguide dispersion coefficients are shown in Figure 2.22. Suppose that the fiber is excited from a $1.5\text{-}\mu\text{m}$ laser source with a linewidth $\Delta\lambda_{1/2}$ of 2 nm. What is the dispersion per km of fiber if the core diameter $2a$ is $8 \mu\text{m}$? What should be the core diameter for zero chromatic dispersion at $\lambda = 1.50 \mu\text{m}$?

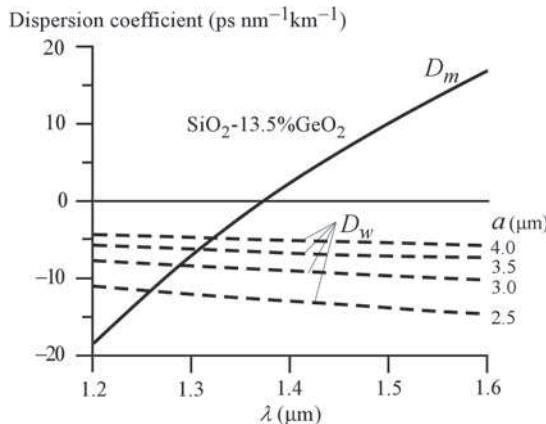


FIGURE 2.22 Material and waveguide dispersion coefficients D_m and D_w in an optical fiber with a core $\text{SiO}_2\text{-}13.5\%\text{GeO}_2$ for $a = 2.5\text{--}4 \mu\text{m}$.

Solution

In Figure 2.22, at $\lambda = 1.50 \mu\text{m}$, $D_m = +10 \text{ ps nm}^{-1} \text{ km}^{-1}$ and, with $a = 4 \mu\text{m}$, $D_w = -5 \text{ ps nm}^{-1} \text{ km}^{-1}$ so that chromatic dispersion is

$$D_{ch} = D_m + D_w = 10 - 5 = 5 \text{ ps nm}^{-1} \text{ km}^{-1}$$

Total dispersion or chromatic dispersion per km of fiber length is then

$$\Delta\tau_{1/2}/L = |D_{ch}| \Delta\lambda_{1/2} = (5 \text{ ps nm}^{-1} \text{ km}^{-1})(2 \text{ nm}) = 10 \text{ ps km}^{-1}$$

Dispersion will be zero at $1.50 \mu\text{m}$ when $D_w = -D_m$ or when $D_w = -10 \text{ ps km}^{-1} \text{ nm}^{-1}$. Examination of D_w vs. λ for $a = 2.5\text{--}4 \mu\text{m}$ shows that a should be about $3 \mu\text{m}$. It should be emphasized that although $D_m + D_w = 0$ at $1.50 \mu\text{m}$, this is only at one wavelength, whereas the input radiation is over a range $\Delta\lambda$ of wavelengths so that in practice chromatic dispersion is never actually zero. It will be minimum at around $1.50 \mu\text{m}$ when, as in the present case, the core radius $a \approx 3 \mu\text{m}$.

EXAMPLE 2.5.3 Chromatic dispersion at different wavelengths

Many fiber manufacturers quote the wavelength dependence of the chromatic dispersion coefficient as

$$D_{ch} = \frac{S_0\lambda}{4} \left[1 - \left(\frac{\lambda_0}{\lambda} \right)^4 \right] \quad (2.5.13)$$

Chromatic dispersion

in which λ_0 is the wavelength at which $D_{ch} = 0$ and S_0 is the slope ($dD/d\lambda$) of the dispersion curve at λ_0 . Consider a standard SMF that has $\lambda_0 = 1310$ nm, and $S_0 = 0.090$ ps nm $^{-2}$ km $^{-1}$. Calculate D_{ch} at 1550 nm and compare with the value in Figure 2.20.

Solution

We can substitute the values into Eq. (2.5.13)

$$D_{ch} = \frac{S_0\lambda}{4} \left[1 - \left(\frac{\lambda_0}{\lambda} \right)^4 \right] = \frac{(0.09 \text{ ps nm}^{-2} \text{ km}^{-1})(1550 \text{ nm})}{4} \left[1 - \left(\frac{1310 \text{ nm}}{1550 \text{ nm}} \right)^4 \right] = 17.1 \text{ ps nm}^{-1} \text{ km}^{-1}$$

This value is not far out from D_{ch} at 1550 nm in Figure 2.20.

EXAMPLE 2.5.4 Waveguide dispersion

One particular single-mode optical fiber has a core diameter of 8.2 μm , $n_2 = 1.4485$, $\Delta = 0.323\%$. What is the waveguide dispersion coefficient at 1310 nm?

Solution

We only need to apply Eq. (2.5.6) or Eq. (2.5.7), that is

$$D_w (\text{ps nm}^{-1} \text{ km}^{-1}) \approx -\frac{83.76 \lambda (\mu\text{m})}{[a(\mu\text{m})]^2 n_2} = -\frac{83.76 (1.31 \mu\text{m})}{[4.1 (\mu\text{m})^2] (1.4485)} = -4.50 \text{ ps nm}^{-1} \text{ km}^{-1}$$

which is approximately the value shown in Figure 2.22.

2.6 DISPERSION MODIFIED FIBERS AND COMPENSATION

A. Dispersion Modified Fibers

Many fibers are named in terms of their dispersion characteristics, because dispersion ultimately controls the bandwidth for information transmission and the way in which the fiber can be used in communications. The standard silica single-mode fiber has its chromatic dispersion zero at around 1300 nm as shown in Figure 2.23. A fiber that has this dispersion curve shifted to have the zero in the middle of the useful 1500–1600 nm (approximate) range of communication channels is called the **zero dispersion-shifted fiber**. The **nonzero dispersion-shifted fiber** has the “zero” dispersion point outside this range. **Reduced slope fiber** has been designed to have the dispersion slope, that is D_{ch} vs. wavelength slope, S , reduced in the same region for use in such applications as wavelength division multiplexing (WDM), which is discussed later. The term **nonzero dispersion fiber** generally defines a fiber that has a nonzero dispersion in its range of operating wavelengths. The 1460–1625 nm region is of special interest in telecommunications due to low fiber losses in this range. It contains three of the communications bands labeled as S (1460–1525 nm), C (1525–1565 nm), and L (1565–1625 nm). Figure 2.23 shows the dispersion characteristics of the above fibers.

The basic idea of controlling the chromatic dispersion, that is D_{ch} vs. λ behavior, is straightforward and involves modifying the waveguide dispersion coefficient D_w . The waveguide dispersion characteristics (D_w vs. λ) can be adjusted by changing the waveguide geometry. Waveguide dispersion arises from the group velocity v_g dependence on the wavelength λ . As the wavelength increases, the field penetrates more into the cladding, which changes the proportion of light energy carried by the core and the cladding and hence changes v_g . We can thus alter the waveguide geometry, that is, *modify the refractive index profile*, and thereby control D_w to yield

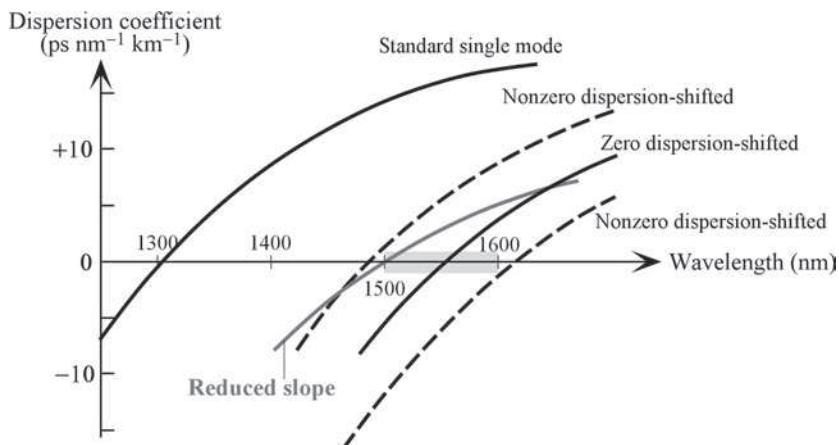


FIGURE 2.23 Various fibers named after their dispersion characteristics. The range 1500–1600 nm is only approximate and depends on the particular application of the fiber.

a total chromatic dispersion that is close to the desired behavior, often either zero at a certain wavelength or has the right slope over a range of wavelengths.

The **dispersion-flattened fiber** is an optical fiber that has a refractive index profile that results in chromatic dispersion being small and flat over a wavelength range, that is, *flattened* between the wavelength λ_1 and λ_2 as shown in Figure 2.24. The refractive index profile of such a fiber looks like a W in which the cladding is a thin layer with a depressed refractive index; the fiber is called **doubly clad**. The simple step-index fiber is singly clad. Greater control on waveguide dispersion can be obtained by using multiply clad fibers. Such fibers are more difficult to manufacture but can exhibit excellent chromatic dispersion $1\text{--}3 \text{ ps nm}^{-1} \text{ km}^{-1}$ over $1.3\text{--}1.6 \mu\text{m}$. Low dispersion over a wavelength range, of course, allows wavelength multiplexing, for example, using a number of different wavelengths as communication channels.

It should be apparent that an important characteristic of a fiber is not only the magnitude of the dispersion but also its slope with respect to the wavelength, which is denoted by $S = dD/d\lambda$ and measured in $\text{ps nm}^{-2} \text{ km}^{-1}$. Fiber specifications normally specify not

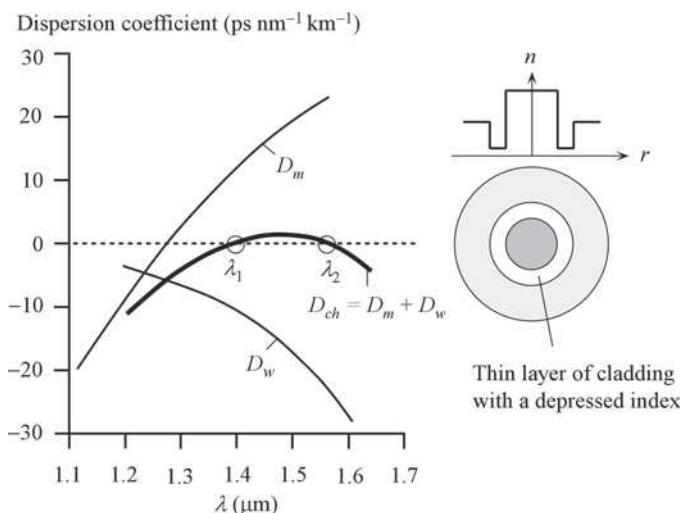


FIGURE 2.24 One type of dispersion-flattened fiber. The material dispersion coefficient (D_m) for the core material and waveguide dispersion coefficient (D_w) for the doubly clad fiber result in a flattened small chromatic dispersion between λ_1 and λ_2 .

only typical dispersion values at various wavelengths of interest but also the dispersion slope. International Telecommunications Union (ITU) has defined certain fiber types for various uses, such as telecommunications using wavelength division multiplexing, by specifying their dispersion characteristics, as well as numerous other properties, including the attenuation along the fiber, mode field diameter, *etc.* For example G.655 is a nonzero dispersion-shifted fiber that has its λ_0 outside the 1500–1600 nm range, a dispersion magnitude less than $6 \text{ ps nm}^{-1} \text{ km}$, PMD less than $0.5 \text{ ps km}^{-1/2}$ and other specifications such as an MFD between 8 μm and 11 μm and an attenuation along the fiber that is less than 0.35 dB km^{-1} . Table 2.3 gives three examples of standard fibers defined by ITU for three single-mode fibers: a standard SMF, and nonzero dispersion-shifted fibers for WDM (wavelength division multiplexing), and DWDM (dense WDM) applications, respectively. A few selected commercial SMF properties are also given. The waveguide geometries and refractive index profiles can be quite sophisticated to generate the required fiber characteristics.

B. Dispersion Compensation

One technique to control and keep the overall chromatic dispersion low is to use a **dispersion compensation**. The idea is to have positive chromatic dispersion build up along a transmission fiber and then use a second fiber, called a **compensating fiber**, with *opposite* (negative) chromatic

TABLE 2.3 Selected single-mode fibers

Fiber	$D_{ch}(\text{ps nm}^{-1} \text{ km}^{-1})$	$S_0(\text{ps nm}^{-2} \text{ km}^{-1})$	$D_{PMD}(\text{ps km}^{-1/2})$	Some attributes
Standard single mode, ITU-T G.652	17 (1550 nm)	≤ 0.093	<0.5 (cabled)	$D_{ch} = 0$ at $\lambda_0 \approx 1312 \text{ nm}$. MFD = 8.6–9.5 μm at 1310 nm. $\lambda_c \leq 1260 \text{ nm}$.
Nonzero dispersion-shifted fiber, ITU-T G.655	0.1–6 (1530 nm)	<0.05 at 1550 nm	<0.5 (cabled)	For 1500–1600 nm range. WDM application. MFD = 8–11 μm .
Nonzero dispersion-shifted fiber, ITU-T G.656	2–14	<0.045 at 1550 nm	<0.20 (cabled)	For 1460–1625 nm range. DWDM application. MFD = 7–11 μm (at 1550 nm). Positive D_{ch} . $\lambda_c < 1310 \text{ nm}$
Corning SMF28e ⁺ (Standard SMF)	18 (1550 nm)	0.088	<0.1	Satisfies G.652. $\lambda_0 \approx 1317 \text{ nm}$, MFD = 9.2 μm (at 1310 nm), 10.4 μm (at 1550 nm); $\lambda_c \leq 1260 \text{ nm}$.
OFS TrueWave RS Fiber	2.6–8.9	0.045	0.02	Satisfies G.655. Optimized for 1530–1625 nm. MFD = 8.4 μm (at 1550 nm); $\lambda_c \leq 1260 \text{ nm}$.
OFS REACH Fiber	5.5–8.9	0.045	0.02	Higher performance than G.655 specification. Satisfies G.656. For DWDM from 1460 to 1625 nm. $\lambda_0 \leq 1405 \text{ nm}$. MFD = 8.6 μm (at 1550 nm)

Note: ITU-T is the International Telecommunications Union with the suffix T representing the Telecommunication Standardization Sector in ITU. G.652, G.655, and G.656 are their standards for three single-mode fibers: a standard SMF, nonzero dispersion-shifted fibers for WDM (wavelength division multiplexing), and DWDM (dense WDM) applications, respectively. A few selected commercial SMF properties are also given. λ_0 is the wavelength at which $D_{ch} = 0$.

dispersion characteristics to bring down the overall dispersion. The compensating fiber has dispersion characteristics that nearly cancel the dispersion in the transmission fiber. The transmission fiber for use is produced with some chromatic dispersion, for example $+10 \text{ ps nm}^{-1} \text{ km}^{-1}$, and dispersion is allowed to build up to say $\Delta\tau_1$ over a certain length L_1 . Then dispersion $\Delta\tau_1$ is “canceled” by using a second fiber, the compensating fiber, that has negative dispersion characteristics with respect to the first fiber, that is, modes that travel fast in the first fiber are slow in the second fiber as illustrated in Figure 2.25. If D_1 and D_2 are the *chromatic dispersion coefficients* of the transmission and compensating fibers with lengths L_1 and L_2 , respectively, then the overall or net dispersion per unit spectral linewidth is $D_1L_1 + D_2L_2$. We can write this net dispersion in terms of an **effective or net chromatic dispersion coefficient** D_{net} , that is,

$$D_{\text{net}}L = D_1L_1 + D_2L_2 \quad (2.6.1)$$

Dispersion compensation

in which L is the total length ($L_1 + L_2$).¹² The compensating fiber has the right D_2L_2 such that the overall dispersion $D_{\text{net}}L$ is reduced below a certain desirable level. Since D_1 and D_2 have opposite signs, D_{net} can be made small as shown in Figure 2.26.

The dispersion D_1 in the transmission fiber is normally positive, and D_2 in the compensating fiber has to be negative to compensate for the dispersion in the transmission fiber, which means that waveguide dispersion in the compensating fiber has to be made dominant and appropriately shaped in terms of D_2 vs. wavelength behavior by suitably engineering the fiber index profile.

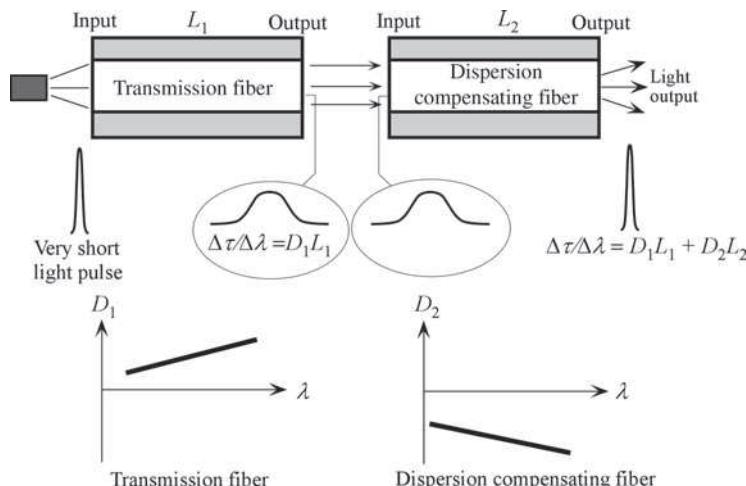


FIGURE 2.25 The chromatic dispersion D_1L_1 in the transmission fiber over its length L_1 is compensated by using a second fiber, a compensating fiber, which has the opposite chromatic dispersion. The length of the compensating fiber is L_2 and the dispersion in it is D_2L_2 . The overall dispersion $\Delta\tau/\Delta\lambda = D_1L_1 + D_2L_2$.

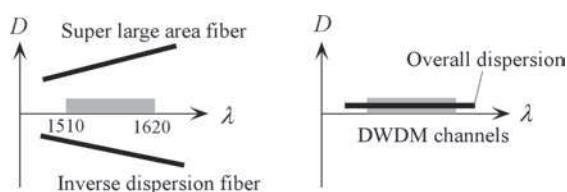


FIGURE 2.26 Dispersion D vs. wavelength characteristics involved in dispersion compensation. Inverse dispersion fiber enables the dispersion to be reduced and maintained flat over the communications wavelengths.

¹² An astute reader would have recognized D_{net} as the average or effective chromatic dispersion coefficient for the whole combined fiber (transmission + compensating fibers together).

Dispersion compensation is also used in generating a flat dispersion over the communication wavelengths as illustrated in Figure 2.26. For example, for ultralong-haul communications the so-called UltraWave fiber that is designed for dense wavelength division multiplexed (DWDM) communications has a large cross-sectional area to enable more power to be fed into the fiber—it is a super large area (SLA) fiber with positive dispersion. The **inverse dispersion fiber** (IDF) has a *negative dispersion slope* that mirrors the dispersion slope in the SLA fiber so that the combination of SLA and IDF, the dispersion managed set, has a small and flat dispersion over the DWDM wavelengths.

The introduction of DWDM has led to certain requirements not only on the magnitude of dispersion but also on its dependence on the wavelength (dispersion slope). Consequently, dispersion management in fibers is one of the most active areas of current research and development. However, while compensating fiber enables dispersion to be controlled, one should note that the compensating fiber has higher attenuation, due to the more heavily doped core region, and needs to be of shorter length. Compensation depends on the temperature. It is more susceptible to nonlinear effects. Different cross-section(s) between the transmission and the compensating fiber can lead to splicing or coupling losses. They are generally used at the receiver end. Manufacturers provide the transmission fiber spliced to a compensating fiber for a well-defined D vs. λ behavior. The **inverse dispersion slope fiber** (mentioned above) has the *inverse dispersion* of the transmission fiber, and cancels the detrimental effect of dispersion across the spectrum of communication wavelengths to allow wavelength division multiplexing to be used.

EXAMPLE 2.6.1 Dispersion compensation

Calculate the total dispersion and the overall net dispersion coefficient when a 400 km transmission fiber with $D_{ch} = +10 \text{ ps nm}^{-1} \text{ km}^{-1}$ is spliced to a compensating fiber that is 30 km long and has $D_{ch} = -100 \text{ ps nm}^{-1} \text{ km}^{-1}$. Assume that the input light spectral width is 1 nm.

Solution

Using Eq. (2.6.1) with $\Delta\lambda = 1 \text{ nm}$, we can find the total dispersion

$$\begin{aligned}\Delta\tau &= (D_1 L_1 + D_2 L_2) \Delta\lambda \\ &= [(+10 \text{ ps nm}^{-1} \text{ km}^{-1})(400 \text{ km}) + (-100 \text{ ps nm}^{-1} \text{ km}^{-1})(30 \text{ km})](1 \text{ nm}) \\ &= 1000 \text{ ps} \quad \text{for } 430 \text{ km}\end{aligned}$$

The net or effective dispersion coefficient can be found from $\Delta\tau = D_{net} L \Delta\lambda$

$$D_{net} = \Delta\tau / (L \Delta\lambda) = (1000 \text{ ps}) / [(430 \text{ km})(1 \text{ nm})] = 2.3 \text{ ps nm}^{-1} \text{ km}^{-1}$$

2.7 BIT RATE, DISPERSION, AND ELECTRICAL AND OPTICAL BANDWIDTH

A. Bit Rate and Dispersion

In digital communications, signals are generally sent as light pulses along an optical fiber. Information is first converted to an electrical signal that is in the form of pulses as illustrated in Figure 2.27. The pulses represent bits of information that are in digital form. For simplicity we have taken the pulses to be very short but generally there is a well-defined pulse duration.

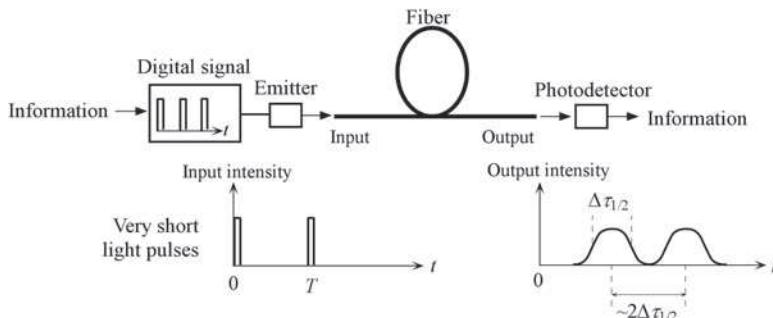


FIGURE 2.27 An optical fiber link for transmitting digital information and the effect of dispersion in the fiber on the output pulses.

The electrical signal drives a light emitter such as a laser diode whose light output is appropriately coupled into a fiber for transmission to the destination. The light output at the destination end of the fiber is coupled to a photodetector that converts the light signal back to an electrical signal. The information is then decoded from this electrical signal. Digital communications engineers are interested in the maximum rate at which the digital data can be transmitted along the fiber. This rate is called the **bit rate capacity** B (bits per second) of the fiber and is directly related to the dispersion characteristics.¹³

Suppose that we feed a light pulse of very short duration into the fiber. The output pulse will be delayed by the transit time τ it takes for the light pulse to travel down the fiber. Due to various dispersion mechanisms there will be a spread $\Delta\tau$ in the arrival times of different guided waves—for example, different fundamental modes for different source wavelengths. This dispersion is typically measured between half-power (or intensity) points and is called **full width at half power** (FWHP), or **full width at half maximum** (FWHM). To clarify the definition of $\Delta\tau$ we can use $\Delta\tau_{1/2}$ to represent the extent of dispersion based on FWHM. Intuitively, as apparent from Figure 2.27, clear distinguishability between two consecutive output pulses, that is no **intersymbol interference**, requires that they be time-separated from peak to peak by at least $2\Delta\tau_{1/2}$. Thus, we can only feed in pulses at the input, at best, at every $2\Delta\tau_{1/2}$ seconds, which then defines the period (T) of the input pulses. Thus the **maximum bit rate**, that is the maximum date rate,¹⁴ B , at which pulses can be transmitted is very roughly $1/(2\Delta\tau_{1/2})$.

$$B \approx \frac{0.5}{\Delta\tau_{1/2}} \text{ (bits s}^{-1}\text{)} \quad (2.7.1)$$

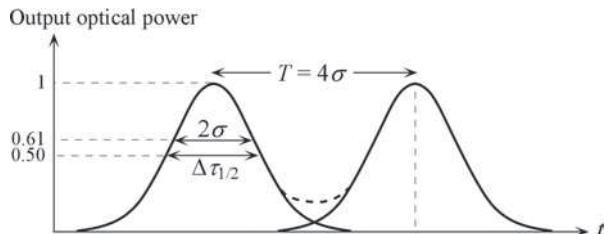
Intuitive
RTZ bit
rate and
dispersion

The maximum bit rate B in Eq. (2.7.1) assumes a pulse representing the binary information 1 that must return to zero for a duration before the next binary information. Two consecutive pulses for two consecutive 1's must have a zero in between as in the output pulses shown in Figures 2.27 and 2.28. This bit rate is called the **return-to-zero (RTZ) bit rate or data rate**. On the other hand, it is also possible to send the two consecutive binary 1 pulses without having to return to zero at the end of each 1-pulse. That is, two 1-pulses are immediately next to each other.

¹³The discussion in this section is more intuitive than rigorous; and we are using delta-pulses at the input to simplify the arguments. The input pulses are normally rectangular pulses of finite duration. Mathematical derivations may be found in more advanced textbooks on optical communications. Eqs. (2.7.1) and (2.7.2), however, remain valid.

¹⁴The data rate or the bit rate (B) is the average number of symbols or characters or a block of characters (bytes) being transferred through a communication system. It is measured in bits per second (symbols per unit second) or in bytes per unit second. For example, 1 megabit per second would be 1 Mb s^{-1} , which is 125 kilobytes per second, 125 kB s^{-1} .

FIGURE 2.28 A Gaussian output light pulse and some tolerable intersymbol interference between two consecutive output light pulses (y-axis in relative units). At time $t = \sigma$ as measured from the pulse center, the relative magnitude is $e^{-1/2} = 0.607$ and full-width root mean square (rms) spread is $\Delta\tau_{\text{rms}} = 2\sigma$ (the RTZ case).



The two pulses in Figure 2.28 can be brought closer until the repetition period $T = 2\sigma$ and the signal is nearly uniform over the length of these two consecutive 1's. Such a maximum data rate is called **nonreturn to zero (NRZ) bit rate**. The NRZ bit rate is twice the RTZ bit rate. The maximum bit rate and dispersion discussions below refer to the RTZ data rate.

In a more rigorous analysis we need to know the temporal shape of the output signal and the criterion for discerning the information; for example, what is the extent of allowed overlap in the output light pulses? There is a general approximate relationship between the bit rate and the *root-mean-square (rms) dispersion* (or the mean standard deviation from the mean¹⁵) σ , which is shown in Figure 2.28. Notice that the *full-width rms time spread* $\Delta\tau_{\text{rms}}$ of the pulse between its rms points is 2σ . For a Gaussian output light pulse, σ and $\Delta\tau_{1/2}$ are related by $\sigma = 0.425\Delta\tau_{1/2}$. The relationship is different for other light pulse shapes. The maximum bit rate B in terms of σ requires that two consecutive output pulses be separated at least by 4σ between their peaks as shown in Figure 2.28. Thus

Maximum
RTZ bit
rate and
dispersion

$$B \approx \frac{0.25}{\sigma} \quad (2.7.2)$$

For a Gaussian pulse, $\sigma = 0.425\Delta\tau_{1/2}$ so that $B = 0.59/\Delta\tau_{1/2}$, and the bit rate is $\sim 18\%$ greater than the intuitive estimate in Eq. (2.7.1). In some cases, the Gaussian pulses are taken to be separated by at least 5σ , which makes the numerator in Eq. (2.7.2) equal to 0.20 instead of 0.25.¹⁶ In the case of Eq. (2.7.2), about 95% of the pulse energy remains within the identifiable pulse and not in the overlap region (Figure 2.28).

In general, the input pulses are not infinitely short and have a certain time-width that must also be considered in the discernibility of the output information along with the method of information decoding. All these factors typically modify the numerical factor in Eq. (2.7.2).¹⁷ Dispersion increases with fiber length L and also with the range of source wavelengths $\Delta\lambda_{1/2}$, measured between the half-intensity points of the source spectrum (intensity vs. wavelength). This means that the bit rate decreases with increasing L and $\Delta\lambda_{1/2}$. It is therefore customary to specify the product of the bit rate B with the fiber length L at the operating wavelength for a given emitter (LED, laser diode, etc.). Suppose that the rms spread (or deviation) of wavelengths in the light output spectrum of the emitter is σ_λ . For a Gaussian output spectrum, for example, $\sigma_\lambda = 0.425\Delta\lambda_{1/2}$.

¹⁵The shape $g(t)$ of a perfectly Gaussian output light pulse centered at $t = 0$ is $g(t) = [1/(2\pi\sigma^2)^{1/2}] \exp[-t^2/(2\sigma^2)]$ where σ is the root mean square deviation; $\Delta\tau_{\text{rms}} = 2\sigma$. When $t = \sigma$, $g(\sigma) = g(0)e^{-1}$. In practice, Gaussian is only an approximation to the actual light output pulse. (See Appendix A.)

¹⁶Indeed, Eq. (2.7.1) is approximately $0.20/\sigma$ and hence equivalent to taking the separation of two consecutive output pulses as being 5σ . The use of 4σ and 5σ in finding B is mixed in the literature but the difference is not significant given the extent of approximations involved.

¹⁷The maximum bit rate depends on the input pulse shape, fiber dispersion characteristics and hence the output pulse shape, and, not least, the modulation scheme of encoding the information. The treatment is covered in optical communications courses and obviously beyond the scope of an elementary book.

If D_{ch} is the chromatic dispersion coefficient, then the rms dispersion of the output light pulse is $LD_{ch}\sigma_\lambda$. Then, the BL product, called the bit rate \times distance product, is given by

$$BL \approx \frac{0.25L}{\sigma} = \frac{0.25}{|D_{ch}|\sigma_\lambda} \quad (2.7.3)$$

Maximum
bitrate \times
distance

It is clear that BL is a characteristic of the fiber, through D_{ch} , and also of the range of source wavelengths. In specifications, the fiber length is taken as 1 km. For example, BL for a step-index single-mode fiber operating at a wavelength of 1.3 μm and excited by a laser diode source is several Gb s^{-1} km. Dividing this by the actual length of the fiber gives the maximum operating bit rate for that length.

When both chromatic (or intramodal) and intermodal dispersion are present, as in a graded index fiber (discussed later), we have to combine the two effects taking into account that their origins are different. Both material and waveguide dispersion arise from a range of input wavelength (*i.e.*, $\Delta\lambda$) and the net effect is simply the linear addition of the two dispersion coefficients, $D_{ch} = D_w + D_m$. The same is *not* true for combining intermodal and intramodal dispersion simply because their origins are different. The *overall dispersion* in terms of an rms dispersion σ can be found from individual rms dispersions by

$$\sigma^2 = \sigma_{\text{intermodal}}^2 + \sigma_{\text{intramodal}}^2 \quad (2.7.4)$$

Total rms
dispersion

or, in terms of FWHM dispersion widths

$$(\Delta\tau_{1/2})^2 = (\Delta\tau_{1/2})_{\text{intermodal}}^2 + (\Delta\tau_{1/2})_{\text{intramodal}}^2 \quad (2.7.5)$$

Total
FWHM
dispersion

and σ can then be used in Eq. (2.7.2) to approximately find B . Equation (2.7.4) is generally valid for finding the resultant rms deviation from individual rms deviations whenever two independent processes are superimposed.

Equations (2.7.2) and (2.7.3) seem innocuously simple due to the fact that they have been written in terms of the rms dispersion. To determine B from $\Delta\tau_{1/2}$ we need to know the pulse shape. It is common to assume Gaussian pulse shapes as in Figure 2.28, and for an ideal **Gaussian pulse**, $\sigma = 0.425\Delta\tau_{1/2}$ and $B = 0.25/\sigma = 0.59/\Delta\tau_{1/2}$. For a rectangular pulse with a full-width ΔT , the FWHM, $\Delta\tau_{1/2}$, is the same as ΔT . Further, $\sigma = 0.29\Delta\tau_{1/2} = 0.29\Delta T$, which is quite different than the Gaussian pulse case. The maximum bit rate in Eq. (2.7.2) for Gaussian pulses does not make intuitive sense for a rectangular pulse. The rectangular pulses must be separated by at least ΔT so that $B < 1/\Delta T$. Table 2.4 summarizes the relationships between $\Delta\tau_{1/2}$, σ , and B for different optical pulse shapes.

B. Optical and Electrical Bandwidth

The emitter in the simple optical fiber link in Figure 2.27 can also be driven, or modulated, by an analog signal that varies continuously with time. We can, for example, drive the emitter using a sinusoidal signal as shown in Figure 2.29. The input light intensity into the fiber then becomes modulated to be a sinusoidal with time at the same frequency f as the modulating signal. The light output intensity at the fiber destination should also be sinusoidal with only a shift in phase due to the time it takes for waves to travel along the fiber. We can determine the transfer characteristics of the fiber by feeding in sinusoidal light intensity signals, which have the same intensity but different modulation frequencies f . Ideally, the light output should have the same intensity for the various modulation frequencies. Figure 2.29 shows the observed optical transfer characteristic of the fiber, which is defined as the output light power per unit input light power (P_o/P_i), as a function of

TABLE 2.4 Relationships between dispersion parameters, maximum bit rates, and bandwidths

Dispersed pulse shape	Pulse shape and FWHM width, $\Delta\tau_{1/2}$	$\Delta\tau_{1/2}$ FWHM width	B (RZ)	f_{op}
Gaussian with rms deviation σ		$\Delta\tau_{1/2} = 2.353\sigma$ $\sigma = 0.425\Delta\tau_{1/2}$	$0.25/\sigma$	$0.75B = 0.19/\sigma$
Triangular pulse with full-width ΔT		$\Delta\tau_{1/2} = 0.5\Delta T = (6^{1/2})\sigma$ $\sigma = \Delta\tau_{1/2}/2.45 = 0.408\Delta\tau_{1/2}$	$0.25/\sigma$	$0.99B = 0.247/\sigma$
Rectangular with full-width ΔT		$\sigma = 0.289\Delta T = 0.289\Delta\tau_{1/2}$ $\Delta\tau_{1/2} = \Delta T = (2)(3^{1/2})\sigma$	$<1/\Delta T$	$0.69B = 0.17/\sigma$

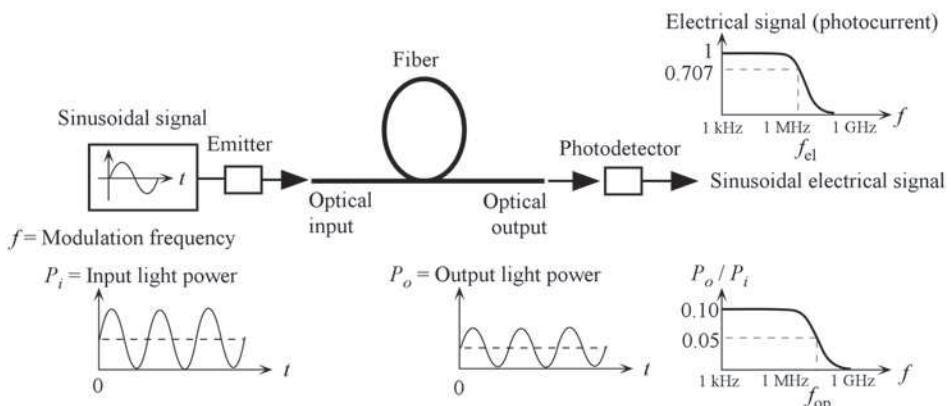
(Source: Data from J. Gowar, *Optical Communication Systems*, 2nd Edition (Prentice Hall, Pearson Education, 1993) Chapter 1.)

Note: RZ = Return-to-zero pulses.

modulation frequency f . The response is flat and then falls with frequency. The reason is that as the frequency becomes too high, the pulse separation period becomes too short, so that dispersion effects smear out the light at the output. The frequency f_{op} at which the output intensity is 50% below the flat region defines the **optical bandwidth** f_{op} of the fiber and hence the useful frequency range in which modulated optical signals can be transferred along the fiber. Intuitively, the optical cutoff frequency f_{op} should correspond roughly to the bit rate, that is, we would expect $f_{op} \approx B$. This is not entirely true because B can tolerate some pulse overlap depending on the shape of the output pulse and the discernibility criterion. If the fiber dispersion characteristics are Gaussian, then

$$f_{op} \approx 0.75B \approx \frac{0.19}{\sigma} \quad (2.7.6)$$

Optical bandwidth for Gaussian dispersion

**FIGURE 2.29** An optical fiber link for transmitting analog signals and the effect of dispersion in the fiber on the bandwidth, f_{op} . P_o/P_i has been arbitrarily set to 0.1 to allow for some attenuation in the fiber.

in which σ is the total rms dispersion through the fiber. The optical bandwidth and fiber length product $f_{\text{op}}L$ is then approximately 25% smaller than the BL product.

It is important to realize that the electrical signal from the photodetector (photocurrent or voltage) does not exhibit the same bandwidth. This is because bandwidth f_{el} for electrical signals is measured where the signal is 70.7% of its low frequency value as indicated in Figure 2.29. The electrical signal (photocurrent) from the photodetector is proportional to the fiber-output light power and when this is 50% below the flat region at f_{op} , so is the electrical signal. This f_{op} is therefore greater than f_{el} . It is, of course, f_{el} that is of interest in optical receiver system design. The relationship between f_{el} and f_{op} depends on the dispersion through the fiber. For Gaussian dispersion, $f_{\text{el}} \approx 0.71f_{\text{op}}$. The use of the term *bandwidth* alone normally implies an optical bandwidth, and is commonly denoted BW. Table 2.4 summarizes the relationships between various dispersion parameters, maximum bit rates, and bandwidths for Gaussian, triangular, and rectangular dispersed pulse shapes. Remember that digital information transmission using a nonreturn-to-zero (NRZ) scheme has twice the bit rate of the RZ scheme.

EXAMPLE 2.7.1 Bit rate and dispersion for a single-mode fiber

Consider a single-mode optical fiber for use in medium-haul communications at 1310 nm. It has a dispersion coefficient $4 \text{ ps nm}^{-1} \text{ km}^{-1}$. Calculate the bit rate distance product (BL) and the optical and electrical bandwidths for a 10-km fiber if a laser diode source with an FWHP linewidth $\Delta\tau_{1/2}$ of 3 nm is used.

Solution

For FWHP dispersion,

$$\Delta\tau_{1/2}/L = |D_{ch}| \Delta\lambda_{1/2} = (4 \text{ ps nm}^{-1} \text{ km}^{-1})(3 \text{ nm}) = 12 \text{ ps km}^{-1}$$

Assuming a Gaussian light pulse shape, the RTZ bit rate \times distance product (BL) is

$$BL = 0.59L/\Delta\tau_{1/2} = 0.59/(12 \text{ ps km}^{-1}) = 49.2 \text{ Gbs}^{-1} \text{ km}$$

The optical bandwidth distance product $f_{\text{op}}L = 0.75BL = 36.9 \text{ GHz km}$.

The maximum bit rate, optical and electrical bandwidths for a 10 km-fiber is

$$B = (49.2 \text{ Gbs}^{-1} \text{ km})/(10 \text{ km}) = 4.9 \text{ Gb s}^{-1}$$

$$f_{\text{op}} = 0.75B = 0.75(49.2 \text{ Gbs}^{-1} \text{ km})/(10 \text{ km}) = 3.7 \text{ GHz}$$

and

$$f_{\text{el}} = 0.71f_{\text{op}} = 2.6 \text{ GHz}$$

For comparison, the 100- μm core MMF in Example 2.3.5, operating at 850 nm, has full-width dispersion, $\Delta\tau/L = 11.3 \text{ ns km}^{-1}$. Its maximum bit rate (see Table 2.4) would be

$$BL \approx L/\Delta\tau = 1/(11.3 \text{ ns km}^{-1}) = 87 \text{ Mb s}^{-1} \text{ km}$$

2.8 THE GRADED INDEX (GRIN) OPTICAL FIBER

A. Basic Properties of GRIN Fibers

The main drawback of the single-mode step-index fiber is the relatively small numerical aperture (NA) and hence the difficulty in the amount of light that can be coupled into it. Increasing the NA means only increasing the V -number, which must be less than 2.405. Multimode fibers

have a relatively larger NA and hence accept more light from broader angles. The greater NA and wider core diameter of the multimode fiber allows not only easier light coupling but also more optical power to be launched into the fiber. Multimode fibers suffer from intermodal dispersion. In terms of visualizing rays this means that the rays representing the modes travel along different path lengths and hence arrive at different times at the end of the fiber as shown in Figure 2.30 (a).

In the **graded index fiber** the refractive index is not constant within the core but decreases from n_1 at the center, as a power law, to n_2 at the cladding as shown in Figure 2.30 (b). The refractive index profile across the core is close to being parabolic. Such a refractive index profile is capable of minimizing intermodal dispersion to a virtually innocuous level. All the rays (e.g., marked 1, 2, and 3) in the graded index fiber arrive at the same time as illustrated in Figure 2.30 (b). The intuitive reason for this is that the velocity along the ray path, c/n , is not constant and increases as the ray is farther away from the center. A ray such as 2 that has a longer path than ray 1 then experiences a faster velocity during a part of its journey to enable it to catch up with ray 1. Similarly, ray 3 experiences a faster velocity than 2 during part of its propagation to catch up with ray 2, and so on.

Suppose that we consider the graded index fiber to consist of many thin concentric layers each of constant refractive index, $n_a > n_b > n_c$, etc., as illustrated in Figure 2.31 (a) above the fiber axis. We have essentially stratified the medium. If θ_a is the incidence angle of ray 1 in layer a on the interface $a-b$, then the refracted ray in b will have an angle θ_b such that $n_a \sin \theta_a = n_b \sin \theta_b$. Similarly, ray 1 would be refracted going from b to c such that $n_b \sin \theta_b = n_c \sin \theta_c$ and so on, until the incidence angle reaches that for TIR at the $f-g$ interface as shown for ray 1 in Figure 2.31(a). Ray 2 starts with a wider θ_a and hence it reaches TIR at $c-d$ below ray 1. If we engineer the n -variation just right, then the rays 1 and 2 leaving from O on the fiber axis will arrive at the same time at O' further along the axis, which is essentially what happens in a graded index fiber.

In a graded index fiber, the index changes continuously, which is analogous to having a ray travel from layer to layer almost immediately so that there is an immediate sequence of refractions as shown in Figure 2.31 (b). The ray path then becomes *bent continuously*. As the ray propagates, it obeys Snell's law, $n \sin \theta = \text{constant}$, which means that θ must continuously increase, while n continuously decreases along r , until TIR is reached and the ray is reflected. The ray then continues on a continuously bent path to O' . A proper analysis shows that ray trajectories in a graded index

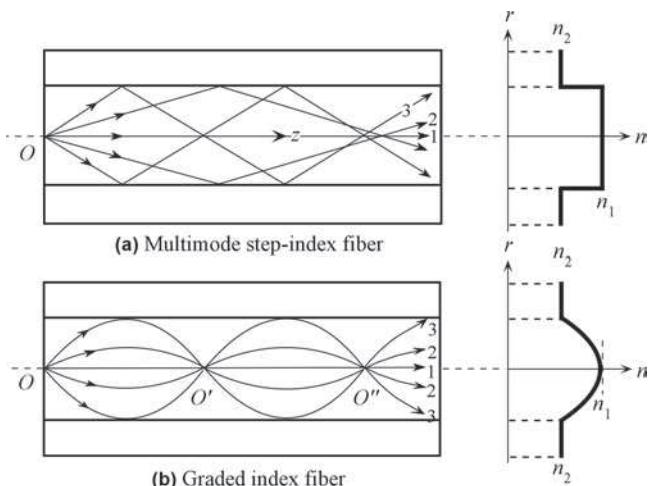


FIGURE 2.30 Multimode fibers.
 (a) Multimode step-index fiber. Ray paths are different so that rays arrive at different times. (b) Graded index fiber. Ray paths are different but so are the velocities along the paths so that all the rays arrive at the same time.

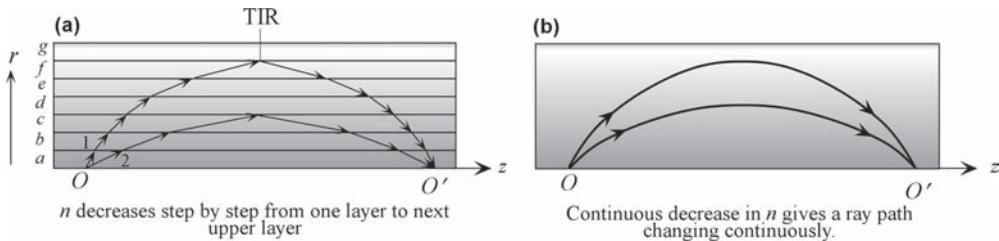


FIGURE 2.31 (a) We can visualize a graded index fiber by imagining a stratified medium with the layers of refractive indices $n_a > n_b > n_c \dots$ as shown for the portion of the fiber above the fiber axis (z -axis). Consider two close rays 1 and 2 launched from O on the fiber axis at the same time but with slightly different launching angles. Ray 1 suffers total internal reflection at the $f-g$ interface. Ray 2 becomes refracted at the $c-d$ interface exactly below that for ray 1. Both rays arrive at the same time at O' if we have a perfectly optimized graded index fiber. (b) Graded index medium in which the ray trajectory is continuously bent.

fiber are sinusoidal-type continuously bent curves as in Figure 2.31(b).¹⁸ All the modes have their maxima at the same location on the z -axis and at about the same time. These ideas apply essentially to meridional rays, those rays that cross the core axis. In addition, there are also continuously bent skew rays, which are **helical rays** that result from rays entering the fiber core off the axis. When we examine intermodal dispersion, we have to also consider these helical rays. When we consider the propagation of all the modes through the graded index fiber, intermodal dispersion is not totally absent though it is reduced enormously in magnitude from the multimode step-index fiber.

The refractive index profile can generally be described by a power law with an index γ called the **profile index** (or the **coefficient of index grading**) so that¹⁹

$$n = n_1[1 - 2\Delta(r/a)^\gamma]^{1/2}; \quad r < a, \text{ core} \quad (2.8.1a)$$

$$n = n_2; \quad r = a, \text{ cladding} \quad (2.8.1b)$$

Graded index profile definition

which looks like the schematic refractive index profile shown in Figure 2.30 (b). Also note that $n_2 = n_1[1 - 2\Delta]^{1/2}$, so that the definition of Δ remains the same as in a step-index fiber. The V-number definition is retained as before in terms of n_1 and n_2 , that is, $V = ka(n_1^2 - n_2^2)^{1/2}$, where k is the free-space propagation constant.

The intermodal dispersion becomes minimum when the fiber index profile (*i.e.*, γ) is optimized when it has a certain value γ_o , *i.e.*, $\gamma = \gamma_o$. To find γ_o we need to consider EM wave propagation in the graded index fiber, and find the right γ ($\gamma = \gamma_o$) that shortens the time interval ΔT between the earliest and latest arrival (delay) times of propagating modes. From ΔT we can find the rms dispersion σ (and hence the optical bandwidth) if we also make an assumption on the dispersed light pulse shape. Further, we also need to consider the wavelength dependence of Δ , that is, $d\Delta/d\lambda$, which is profile dispersion. With the assumption that the fiber is weakly guiding, that is, $\Delta \ll 1$, the optimized profile index is²⁰

$$\gamma_o \approx 2 + \delta - \Delta \frac{(4 + \delta)(3 + \delta)}{5 + 2\delta} \approx 2 \quad (2.8.2)$$

Optimal profile index

¹⁸See B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, 2nd Edition (John Wiley & Sons, 2007), Ch. 1.

¹⁹Some books use α , q , or g instead of γ . Since α is used for attenuation in this chapter, and α and a are visually close in small print, γ is used.

²⁰This was derived by Robert Olshansky and Donald Keck, *Appl. Opt.*, 15, 482, 1976.

in which δ , the profile dispersion parameter of the graded index fiber, is defined by

Profile dispersion parameter

$$\delta = -\left(\frac{n_1 \lambda}{N_{g1} \Delta}\right) \frac{d\Delta}{d\lambda} \quad (2.8.3)$$

The profile index γ_o in Eq. (2.8.2) is called the **optimal profile index**. Since δ is small, γ_o is close to 2 (*parabolic*). Since δ depends on the wavelength, so does γ . A graded index fiber that has been optimized at one wavelength would not be optimized at another wavelength.

With this optimal profile index, the rms dispersion $\sigma_{\text{intermode}}$ in the output light pulse per unit length is usually given by²¹

Dispersion in graded index fiber

$$\frac{\sigma_{\text{intermode}}}{L} \approx \frac{n_1}{20\sqrt{3}c} \Delta^2 \quad (2.8.4)$$

There have been other theoretical treatments of the graded index fiber with the result that the full delay time spread ΔT is $L n_1 \Delta^2 / (8c)$. If we assume a triangular pulse shape, the resulting rms dispersion is not far out from that in Eq. (2.8.4).

It is straightforward to show that the above optimized dispersion is easily a factor of 10^3 smaller than the dispersion in a corresponding step-index multimode fiber (see Example 2.8.1). However, a cautionary note is in order. As the profile index γ deviates from the optimal value γ_o , the dispersion worsens very rapidly. A 2.5% change in γ can increase $\sigma_{\text{intermode}}$ by a factor of 10.

The definition of the numerical aperture (NA) in Eq. (2.4.1) for the step-index fiber is generally also retained for graded index fiber, *i.e.*,

Numerical aperture for GRIN fibers

$$\text{NA} = \text{NA}(r) = [n(r)^2 - n_2^2]^{1/2} \quad (2.8.5)$$

but now NA depends on the radial position (r) since $n(r)$ decreases with r . The NA of a graded index fiber is maximum and the same value as the step-index fiber only at the center of the core; and then decreases with r . It is customary to average $\text{NA}(r)^2$ to obtain an **effective NA** for graded index fiber. For $\gamma \approx 2$, the effective NA is approximately given by

Effective numerical aperture for GRIN fibers

$$\text{NA}_{\text{GRIN}} \approx (1/2^{1/2})(n_1^2 - n_2^2)^{1/2} \quad (2.8.6)$$

which is a factor of $2^{1/2}$ or 1.414 less than the NA for a step-index fiber with the same n_1 and n_2 . Even if a graded index fiber has the same NA at the core center as the step-index fiber, it nonetheless accepts less light than the corresponding step-index fiber.

The number of modes M in a graded index fiber is given by

Number of modes in graded index fiber

$$M \approx \left(\frac{\gamma}{\gamma + 2}\right) \frac{V^2}{2} \quad (2.8.7)$$

Note that, as expected, M reduces to $V^2/2$ for $\gamma = \infty$, *i.e.*, for the step-index MMF. Further, for the graded index fiber with the optimal index ($\gamma = 2$) $M = V^2/4$; there are half as many modes as in the multimode step-index fiber.

If we examine Figure 2.31 (b), we see that for an optimized graded index fiber, all rays emanating from a point O on the fiber axis are bent and then eventually brought to the same point O' on the axis. Clearly, the core, just like a lens, bends and focuses the rays to a point O' .

²¹ See Olshansky, *Rev. Mod. Phys.*, 51, 341, 1979 or J. Senior, *Optical Fiber Communications*, 3rd Edition (Prentice Hall, Pearson Education, 2009), Ch. 3.

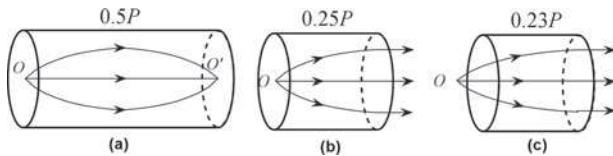
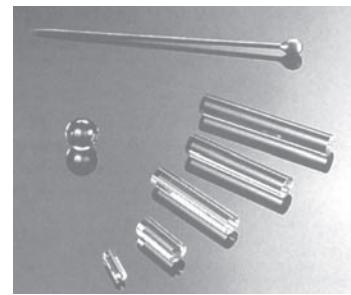


FIGURE 2.32 Graded index rod lenses of different pitches.

(a) Point O is on the rod face center and the lens focuses the rays onto O' on to the center of the opposite face. (b) The rays from O on the rod face center are collimated out. (c) O is slightly away from the rod face and the rays are collimated out.



GRIN rod lenses and a spherical lens (a ball lens) used in coupling light into fibers. (Courtesy of CVI Melles Griot.)

on the axis. A **graded index rod lens** is a glass rod whose refractive index changes parabolically from its central axis where the index is maximum. It is like a very thick, short graded index fiber whose diameter is perhaps 0.5–5 mm. Such GRIN rod lenses of different lengths can be used to focus or collimate light rays in fiber optics, as illustrated in Figure 2.32. The principle of operation can be understood by considering ray trajectories in the graded index medium shown in Figure 2.31 (b) in which the ray trajectories are sinusoidal paths. One pitch (P) is a full one period variation in the ray trajectory along the rod axis. Figure 2.32 (a), (b), and (c) show half-pitch ($0.5P$), quarter-pitch ($0.25P$), and $0.23P$ GRIN rod lenses. The point O in (a) and (b) is on the rod face center whereas in (c) it is slightly away from the rod face. Such GRIN rod lenses, along with spherical lenses, are used to couple light from emitters and detectors into optical fibers, from fibers into photodetectors, or couple one fiber to another.

B. Telecommunications

Graded index fibers are widely used in optical data communications, especially in short- and medium-haul communications, usually over distances up to about 1 km. Local area networks (LANs) invariably use graded index fibers. Manufacturers produce carefully profiled graded index fibers to maintain low dispersion. They are operated mainly at two wavelengths, either 850 nm or 1310 nm, but the 850 nm is much more common today in LANs due to reasonably priced laser emitters as narrow spectral width sources. At present, vertical cavity surface emitting lasers (VCSELs)²² are used most commonly, which have a spectral width less than 1 nm. The 10 Gb s^{-1} GRIN MMF fiber systems are available and 40 Gb s^{-1} and higher are expected soon. Table 2.5 summarizes typical graded index fiber characteristics and standards for LAN applications at 850 nm. For example, the OM4 standard would allow 10G networks with bit rates of 10 Gb s^{-1} to reach distances of 550 m at an operating wavelength of 850 nm.

The overall dispersion σ in the graded index fiber limits the maximum bit rate (data rate) and the optical bandwidth of the fiber. The overall dispersion σ includes both intermodal dispersion $\sigma_{\text{intermode}}$ and intramodal dispersion $\sigma_{\text{intramode}}$ so that

$$\sigma^2 = \sigma_{\text{intermode}}^2 + \sigma_{\text{intramode}}^2 \quad (2.8.8)$$

Total dispersion in a graded index fiber

²²The characteristics of VCSELs are described in Chapter 4. In this section, we are only interested in their spectral width, $<1 \text{ nm}$, and maximum modulation rates, 10 Gb s^{-1} .

TABLE 2.5 Graded index multimode fibers

MMF (d/D)	Compliance standard	Source	Typical D_{ch} ($\text{ps nm}^{-1} \text{km}^{-1}$)	Bandwidth (MHz · km)	NA	$\alpha(\text{dB km}^{-1})$	Reach in 10G and 40G networks
50/125	OM4	VCSEL	-100	4700 (EMB) 3500 (OFLBW)	0.200	<3	550 m (10G) 150 m (40G)
50/125	OM3	VCSEL	-100	2000 (EMB) 500 (OFLBW)	0.200	<3	300 m (10G)
62.5/125	OM1	LED	-117	200 (OFLBW)	0.275	<3	33 m (10G)

Note: d = core diameter (μm), D = cladding diameter (μm). Typical properties at 850 nm. VCSEL is a vertical cavity surface emitting laser. α is attenuation along the fiber. OM1, OM3, and OM4 are fiber standards for LAN data links (Ethernet). α are reported typical attenuation values. 10G and 40G networks represent data rates of 10 Gbs^{-1} and 40 Gbs^{-1} and correspond to 10 GbE (Gigabit Ethernet) and 40 GbE systems.

The intramodal dispersion is due to *chromatic dispersion*, which is essentially material dispersion, and depends on the spectral width of the light source. If the fiber is operated at the wavelength of zero chromatic dispersion, usually around 1300 nm, and if the spectral width is kept small, then the bit rate is essentially limited by intermodal dispersion. Thus, the graded index profile, and its quality control during manufacturing, plays an important role in determining the final bandwidth of the fiber. Most MMFs used in telecommunications are graded index fibers. When the fiber is used at 850 nm, one must include the chromatic (material) dispersion in the bandwidth calculation, which requires the source spectral width; VCSELs have a spectral width that is typically less than 1 nm.

Laser diode excitation allows a wider bandwidth to be attained. However, the laser beam, unlike an LED beam, is narrow with a small spot size, and does not cover the whole core of the MMF. The fiber bandwidth then depends on what narrow portion of the core the laser beam hits, a small region of the core area, or the whole area. **Effective modal bandwidth** (EMB) is used to specify the bandwidth of MMFs that are excited from a small spot size laser source. EMB is also known as the **laser bandwidth**. The **overfilled launch bandwidth** (OFLBW) refers to the bandwidth when an LED is used to excite the whole core. Since LEDs cannot usually be modulated at speeds far exceeding 600 Mb s^{-1} even if the MMF has a wide bandwidth, the actual bottleneck may well be the LED bandwidth rather than the length of the fiber. VCSELs can be easily modulated upto 10 Gbs^{-1} .

EXAMPLE 2.8.1 Dispersion in a graded index fiber and bit rate

Full width of dispersion in an optimized graded index fiber

Consider an optimized graded index fiber. The theoretical temporal time spread in the output light pulse due to the delay times of the first and last modes arriving at the output has been evaluated and quoted in the literature as²³

$$\Delta T_{\text{graded}} \approx \frac{Ln_1 \Delta^2}{8c} \quad (2.8.9)$$

Compare the above full-width dispersion with that for a step-index MMF, assuming a Δ of 0.01.

²³See, for example, J. Gowar, *Optical Communications Systems*, 2nd Edition (Prentice Hall, Pearson Education, 1993), Ch. 9.

What would you expect for the rms dispersion for a rectangular and triangular pulse shapes, and how do these compare with Eq. (2.8.4)?

Solution

The full pulse spread ΔT_{step} for a step-index MMF is $Ln_1\Delta/c$ so that the ratio of the two dispersions is

$$\Delta T_{\text{graded}}/\Delta T_{\text{step}} = (Ln_1\Delta^2/8c)/(Ln_1\Delta/c) = \Delta/8 = 1.25 \times 10^{-3}$$

which shows that the graded index fiber should decrease the dispersion by a factor of 800. In practice the reduction is not as much as this. To obtain the rms dispersion, we need to make an assumption on the shape of the output light pulse given that the input light pulse is a delta function. For rectangular and triangular pulses, the dispersions σ_r and σ_t , respectively, are

$$\sigma_r = \frac{\Delta T_{\text{graded}}}{(2)(3)^{1/2}} \approx \frac{Ln_1\Delta^2}{27.7c} \quad \text{and} \quad \sigma_t = \frac{\Delta T_{\text{graded}}}{(2)(6^{1/2})} \approx \frac{Ln_1\Delta^2}{39.2c}$$

On the other hand, Eq. (2.8.4) is

$$\sigma = \frac{Ln_1\Delta^2}{20\sqrt{3}c} = \frac{Ln_1\Delta^2}{34.6c}$$

In fact, if we assume a rectangular pulse shape, Eq. (2.8.4) corresponds to a full spread ΔT that has a factor of 10, instead of 8 in the denominator of Eq. (2.8.9). While theoretical predictions differ in the exact numerical factor in the denominator, in practice, the actual dispersion is limited by deviations from the optimal index profile and various other factors.

EXAMPLE 2.8.2 Dispersion in a graded index fiber and bit rate

Consider a graded index fiber whose core has a diameter of 50 μm and a refractive index of $n_1 = 1.4750$, $\Delta = 0.010$. (The corresponding NA at the core center is about 0.20.) The fiber is used in LANs at 850 nm with a vertical cavity surface emitting laser that has a very narrow linewidth that is about 0.4 nm (FWHM). Assume that the chromatic dispersion at 850 nm is $-100 \text{ ps nm}^{-1} \text{ km}^{-1}$ as shown in Table 2.5. Assume the fiber has been optimized at 850 nm, and find the minimum rms dispersion. How many modes are there? What would be the upper limit on its bandwidth? What would be the bandwidth in practice?

Solution

Given Δ and n_1 , we can find n_2 from $\Delta = 0.01 = (n_1 - n_2)/n_1 = (1.4750 - n_2)/1.4750$. Thus, $n_2 = 1.4603$. The V-number is then

$$V = (2\pi a/\lambda)(n_1^2 - n_2^2)^{1/2} = [(2\pi)(25 \mu\text{m})/(0.850 \mu\text{m})(1.4750^2 - 1.4603^2)]^{1/2} = 38.39$$

For the number of modes we can simply take $\gamma = 2$ and use

$$M = (V^2/4) = (38.39^2/4) = 368 \text{ modes}$$

The lowest intermodal dispersion for a profile optimized graded index fiber for a 1 km of fiber, $L = 1 \text{ km}$, is

$$\begin{aligned} \frac{\sigma_{\text{intermode}}}{L} &\approx \frac{n_1}{20\sqrt{3}c} \Delta^2 = \frac{1.4750}{20\sqrt{3}(3 \times 10^8)} (0.010)^2 \\ &= 14.19 \times 10^{-15} \text{ s m}^{-1} \quad \text{or} \quad 14.19 \text{ ps km}^{-1} \end{aligned}$$

Thus, assuming a triangular output light pulse and the relationship between σ and $\Delta\tau_{1/2}$ given in Table 2.4, the intermodal spread $\Delta\tau_{\text{intermode}}$ (FWHM) in the group delay over 1 km is

$$\Delta\tau_{\text{intermode}} = (6^{1/2})\sigma_{\text{intermode}} = (2.45)(14.19 \text{ ps}) = 34.8 \text{ ps}$$

We also need the chromatic dispersion at the operating wavelength over 1 km, which makes up the intramodal dispersion $\Delta\tau_{\text{intramode}}$ (FWHM)

$$\Delta\tau_{\text{intramode}} = L|D_{ch}|\Delta\lambda_{1/2} = (1 \text{ km})(-100 \text{ ps nm}^{-1} \text{ km}^{-1})(0.40 \text{ nm}) = 40.0 \text{ ps}$$

Note that the chromatic dispersion is essentially material dispersion. The overall dispersion $\Delta\tau$ (FWHM) from Eq. (2.8.8) is

$$\Delta\tau^2 = \Delta\tau_{\text{intermode}}^2 + \Delta\tau_{\text{intramode}}^2 = (34.8)^2 + (40.0)^2$$

so that

$$\Delta\tau = 53.0 \text{ ps}$$

The corresponding maximum bit (data) rate B from Table 2.4 (assuming a triangular pulse) is

$$B = \frac{0.25}{\sigma} = \frac{0.25}{0.408\Delta\tau} = \frac{0.61}{\Delta\tau} = \frac{0.61}{53.0 \times 10^{-12} \text{ s}} = 11.5 \text{ Gb s}^{-1}$$

so that the BL product is $11.5 \text{ Gb s}^{-1} \text{ km}$. The optical bandwidth (f_{op}) would be $0.99B$, which is 11.4 GHz. This is the upper limit since we assumed that the graded index fiber is perfectly optimized with $\sigma_{\text{intermode}}$ being minimum. Small deviations around the optimum γ cause large increases in $\sigma_{\text{intermode}}$, which would sharply reduce the bandwidth.

If this were a multimode step-index fiber with the same n_1 and n_2 , then the full dispersion (total spread) would roughly be

$$\frac{\Delta\tau}{L} \approx \frac{n_1 - n_2}{c} = \frac{n_1\Delta}{c} = \frac{(1.4750)(0.01)}{3 \times 10^8} = 4.92 \times 10^{-11} \text{ s m}^{-1} \quad \text{or} \quad 49.2 \text{ ns km}^{-1}$$

To calculate the BL we assume a rectangular pulse shape as in Table 2.4,

$$BL \approx \frac{L}{\Delta\tau} = \frac{1}{49.2 \times 10^{-9} \text{ s km}^{-1}} = 20 \text{ Mb s}^{-1} \text{ km}$$

which is nearly ~ 575 times smaller.

Note: LANs now use graded index MMFs, and the step-index MMFs are used mainly in low-speed instrumentation.

2.9 ATTENUATION IN OPTICAL FIBERS

A. Attenuation Coefficient and Optical Power Levels

In general, when light propagates through a material it becomes *attenuated* in the direction of propagation as illustrated in Figure 2.33. The electric field and hence the intensity decrease in the direction of propagation, usually in an exponential manner as shown in Figure 2.33.

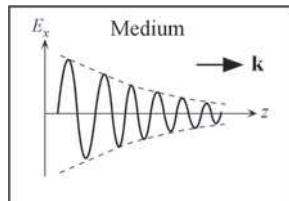


FIGURE 2.33 The attenuation of light in a medium.

We distinguish between *absorption* and *scattering*, both of which give rise to a loss of intensity in the normal direction of propagation (z -direction). In addition, extrinsic factors such as *fiber bending* can also lead to light attenuation.

As light propagates through an optical fiber it becomes attenuated by a number of processes that depend on the wavelength of light. Suppose that the input optical power into a fiber of length L is P_{in} and the output optical power at the end is P_{out} and power anywhere in the fiber at a distance z from the input is P . The optical power **attenuation coefficient** α is defined as the fractional decrease in the optical power per unit distance, that is,

$$\alpha = -\frac{1}{P} \frac{dP}{dz} \quad (2.9.1)$$

Definition of
attenuation
coefficient

We can integrate this over the length L of the fiber to relate α to P_{out} and P_{in} by

$$\alpha = \frac{1}{L} \ln\left(\frac{P_{\text{in}}}{P_{\text{out}}}\right) \quad (2.9.2)$$

Attenuation
coefficient

We can just as well define attenuation in terms of light intensity but we have used power to follow convention since attenuation tests on optical fibers measure the optical power. If we know α then we can always find P_{out} from P_{in} through

$$P_{\text{out}} = P_{\text{in}} \exp(-\alpha L) \quad (2.9.3)$$

Exponential
power
decay

The units for the attenuation coefficient α in exponential decays with distance is simply m^{-1} . However, in general, optical power attenuation in an optical fiber is expressed in terms of decibels per unit length of fiber, typically as dB km^{-1} . The attenuation of the signal in decibels per unit length is defined in terms of the logarithm to base 10 by

$$\alpha_{\text{dB}} = \frac{1}{L} 10 \log\left(\frac{P_{\text{in}}}{P_{\text{out}}}\right) \quad (2.9.4)$$

Attenuation
coefficient
in dB m^{-1}

Substituting for $P_{\text{in}}/P_{\text{out}}$ from above we obtain

$$\alpha_{\text{dB}} = \frac{10}{\ln(10)} \alpha = 4.34\alpha \quad (2.9.5)$$

Conversion
to decibels

The attenuation of the electric field, of course, also follows an exponential decay but its attenuation coefficient is $\alpha/2$ (see Chapter 1).

The attenuation α_{dB} in Eq. (2.9.4) represents relative power values and does not say anything about absolute power values. Often, we would like to indicate the power level involved in a fiber link. Since the attenuation is in terms of dB (per km), we would like to somehow maintain the dB unit. In fiber optic communications, the power level P_{dBm} is measured in dBm and is defined in terms of the actual power level P_{mW} in mW by using a 1-mW reference power level, that is,

$$P_{\text{dBm}} = 10 \log\left(\frac{P_{\text{mW}}}{1 \text{ mW}}\right) \quad (2.9.6)$$

Power
level dBm

Clearly, this definition is similar to that for α_{dB} in Eq. (2.9.4) though different quantities are defined. Thus a power level of 1 mW is 0 dBm. Positive dBm values represent power levels *greater* than 1 mW and negative values *less* than 1 mW. A laser emitter with a power of 50 mW has a power level of 17 dBm.

EXAMPLE 2.9.1 Attenuation along an optical fiber

The optical power launched into a single-mode optical fiber from a laser diode is approximately 1 mW. The photodetector at the output requires a minimum power of 1 nW to provide a clear signal (above noise). The fiber operates at $1.31\text{ }\mu\text{m}$ and has an effective attenuation coefficient of 0.40 dB km^{-1} . What is the maximum ideal length of fiber that can be used without inserting a repeater (to regenerate the signal)?

Solution

We can use

$$\alpha_{\text{dB}} = \frac{1}{L} 10 \log \left(\frac{P_{\text{in}}}{P_{\text{out}}} \right)$$

so that

$$L = \frac{1}{\alpha_{\text{dB}}} 10 \log \left(\frac{P_{\text{in}}}{P_{\text{out}}} \right) = \frac{1}{0.4} 10 \log \left(\frac{10^{-3}}{10^{-9}} \right) = 150\text{ km}$$

There will be additional losses, such as fiber-bending losses which arise from the bending of the fiber or splice losses (two fibers fused together to make a connection between the two). These losses will reduce this length to below this limit. A simple calculation shows that if we increase the attenuation to 0.6 dB km^{-1} , L becomes 100 km. For long-distance communications, the signal has to be amplified, using an optical amplifier, after a distance of a few hundred kilometers, and eventually regenerated by using a repeater.

B. Intrinsic Attenuation in Optical Fibers

Figure 2.34 shows the *standard* attenuation coefficient, as dB per km, of a typical silica glass-based optical fiber as a function of wavelength. The sharp increase in the attenuation at wavelengths beyond $1.6\text{ }\mu\text{m}$ in the *infrared* region is due to energy absorption by “lattice vibrations” of the constituent ions of the glass material. Fundamentally, energy absorption in this region corresponds to the stretching of the Si–O bonds in ionic polarization induced by the EM wave. Absorption increases with wavelength as we approach the resonance wavelength of the Si–O bond, which is around $9.2\text{ }\mu\text{m}$. In the case of Ge–O glasses, this is further away, around $11.0\text{ }\mu\text{m}$. There is another intrinsic material absorption in the region below 500 nm, not shown in the figure, which is due to photons exciting electrons from the valence band to the conduction band of the glass.

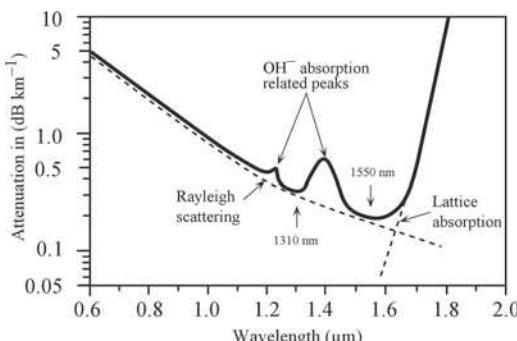


FIGURE 2.34 Attenuation vs. wavelength for a standard silica-based fiber.

There is a marked attenuation peak centered at 1.38 μm , and a barely discernible minor peak at about 1.24 μm . These attenuation regions arise from the presence of hydroxyl ions (OH^-) as impurities in the glass structure inasmuch as it is difficult to remove all traces of hydroxyl (water) products during fiber production. Further, hydrogen atoms can easily diffuse into the glass structure at high temperatures during production, which leads to the formation of hydrogen bonds in the silica structure and OH^- ions. Absorbed energy is mainly by the stretching vibrations of the OH^- bonds within the silica structure, which has a fundamental resonance in the infrared region (2.72 μm) but overtones or harmonics at lower wavelengths (or higher frequencies). The first overtone, twice the fundamental frequency, at around 1.38 μm is the most significant as can be seen in the figure. The second overtone is around 0.94 μm and in high-quality fibers this is negligible. A combination of first overtone of the OH^- vibration (at 1.38 μm) and the fundamental vibrational frequency of SiO_2 (at 9.2 μm) gives rise to a minor loss peak at around 1.24 μm .²⁴

There are two important windows in the α vs. λ behavior in which the attenuation exhibits minima. The window at around 1.3 μm is the region between two neighboring OH^- absorption peaks. This window is widely used in optical communications at 1310 nm. The window at around 1.55 μm is between the first harmonic absorption of OH^- and the infrared lattice absorption tail and represents the lowest attenuation. It can be seen that it is important to keep the hydroxyl content in the fiber within tolerable levels.

There is a background attenuation process that decreases with wavelength and is due to the Rayleigh scattering of light by the local variations in the refractive index. Glass has a noncrystalline structure, which means that there is no long-range order to the arrangement of the atoms but only a short-range order, typically a few bond lengths. We know that the structure of a liquid is noncrystalline, that is, disordered. The glass structure is as if the structure of the melt has been suddenly frozen. The disorder in the liquid becomes frozen in the glass structure. We can only define the number of bonds a given atom in the structure will have. Random variations in the bond angle from atom to atom lead to a disordered structure. There is therefore a random local variation in the density over several bond lengths, that is, from one small region to the next, that leads to fluctuations in the refractive index around some average value for the whole sample. These random fluctuations in the refractive index give rise to the light scattering and hence light attenuation along the fiber. In addition to fluctuations in the density, there can also be changes in the chemical composition over several bond lengths, which introduce more random changes in the refractive index. The compositional fluctuations are small in pure silica glasses but not negligible in multicomponent glasses. It should be apparent that since a degree of structural randomness is an intrinsic property of the glass structure, this scattering process is unavoidable and represents the lowest attenuation possible through a glass medium. As one may surmise, attenuation by scattering in a medium is minimum for light propagating through a “perfect” crystal. In this case the only scattering mechanisms will be due to thermodynamic defects (e.g., vacancies) and the random thermal vibrations of the lattice atoms.

There is another strong fundamental absorption region (not shown) in the short (UV) wavelength range at wavelengths shorter than 600 nm. This absorption is due to the excitation of electrons from the valence band to the conduction band, and is very strong with a tail that extends into longer wavelengths. Its effect over the wavelengths of interest in telecommunications, from 800 nm to 1600 nm, is negligible.

²⁴If f_o is the fundamental frequency (c/λ_o) of vibrations associated with a particular bond such as OH, the first overtone will be at $2f_o$. However, because of the nature of the bond (so-called unharmonicity), it is actually at a slightly lower frequency than $2f_o$.

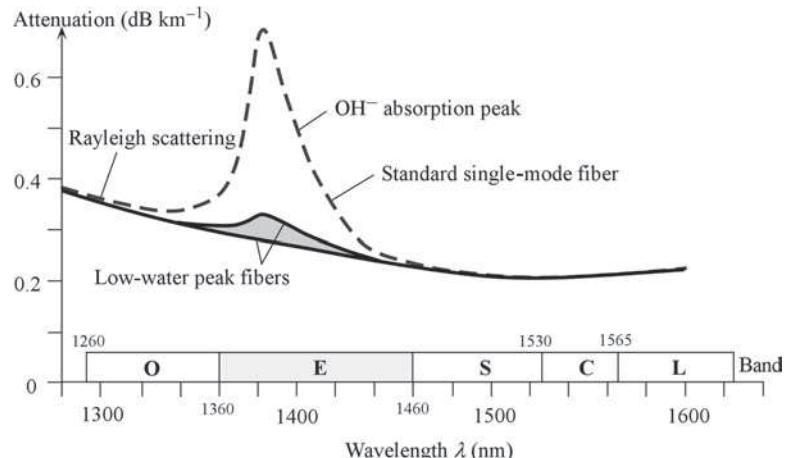


FIGURE 2.35 Typical attenuation vs. wavelength for a standard single-mode optical fiber and also in a low water peak fiber. These fibers are called *all-wave fiber*.

By proper design, the attenuation window at around $1.5 \mu\text{m}$ may be lowered to approach the Rayleigh scattering limit. The OH^- peak at around $1.38 \mu\text{m}$ has prevented the use of the E-band in optical communications around this wavelength. However, recent developments in fiber manufacturing have almost totally eliminated the water peak as apparent in the α vs. λ characteristic of a newly developed low water peak (LWP) fiber shown in Figure 2.35. LWP fibers are commercially available with the attenuation being near the scattering limit from 1290 nm to 1625 nm . The ITU-T G.652D standard defines what we expect from commercial LWP fibers.

C. Intrinsic Attenuation Equations

The tail of the *fundamental infrared absorption* can be represented by a simple exponential

Fundamental lattice absorption

$$\alpha_{\text{FIR}} = A \exp\left(-\frac{B}{\lambda}\right) \quad (2.9.7)$$

where for silica fibers $A = 7.81 \times 10^{11} \text{ dB km}^{-1}$, $B = 48.5 \mu\text{m}$. Although A may seem unreasonably large, Eq. (2.9.7) predicts the right values in the tail region of lattice absorption. These values depend on the composition of the core.

Rayleigh scattering represents the lowest attenuation one can achieve using a glass structure. As mentioned above, the Rayleigh scattering process decreases with wavelength, and it is inversely proportional to λ^4 . The attenuation α_R of a silica fiber due to Rayleigh scattering can be written as

Rayleigh scattering in silica

$$\alpha_R = \frac{A_R}{\lambda^4} \quad (2.9.8)$$

in which λ is the free-space wavelength and A_R is a coefficient that depends on the composition of the core. It also depends on the thermal history involved in manufacturing the fiber. Many books quote $A_R \approx 0.90 \text{ dB km}^{-1}$ for estimating the Rayleigh scattering loss. As shown in Example 2.9.2, it serves as a good rule of thumb, given that it contains no information on the core composition or the refractive index difference. For pure silica (undoped) preforms from which fibers are drawn, A_R is smaller and about 0.63 dB km^{-1} . This attenuation increases as dopants are introduced.

TABLE 2.6 Approximate attenuation coefficients for silica-based fibers for use in Eqs. (2.9.7) and (2.9.8)

Glass	$A_R(\text{dB km}^{-1})$	Comment
Silica fiber	0.90	“Rule of thumb”
$\text{SiO}_2\text{-GeO}_2$ core step-index fiber	$0.63 + 2.06 \times \text{NA}$	NA depends on $(n_1^2 - n_2^2)^{1/2}$ and hence on the doping difference.
$\text{SiO}_2\text{-GeO}_2$ core graded index fiber	$0.63 + 1.75 \times \text{NA}$	
Silica, SiO_2	0.63	Measured on preforms. Depends on annealing. $A_R(\text{Silica}) = 0.59 \text{ dB km}^{-1}$ for annealed.
65% SiO_2 35% GeO_2	0.75	On a preform. $A_R/A_R(\text{silica}) = 1.19$
$(\text{SiO}_2)_{1-x}(\text{GeO}_2)_x$	$A_R(\text{silica}) \times (1 + 0.62x)$	$x = [\text{GeO}_2] = \text{Concentration as a fraction (10\% GeO}_2, x = 0.1).$ For preform.

(Source: Data mainly from K. Tsujikawa *et al.*, *Electron. Letts.*, 30, 351, 1994; *Opt. Fib. Technol.*, 11, 319, 2005; H. Hughes, *Telecommunications Cables*, John Wiley & Sons, 1999, and references therein.)

Note: Square brackets represent concentration as a fraction. NA = numerical aperture. $A_R = 0.59$ used as reference for pure silica, and represents $A_R(\text{silica})$.

Table 2.6 summarizes some typical values for the Rayleigh scattering coefficient, A_R in Eq. (2.9.8). A_R increases as silica is doped, for example, with GeO_2 (germania), to modify its refractive index, and the increase depends on the dopant concentration. Further, given one type of dopant, for example, GeO_2 , the refractive index change is approximately proportional to the concentration of dopants so that A_R can also be expressed in terms of the numerical aperture (NA). The lowest attenuation at 1.55 μm for a silica fiber has been reported to be about 0.15 dB km^{-1} .

EXAMPLE 2.9.2 Rayleigh scattering equations

Consider a single-mode step-index fiber, which has a numerical aperture of 0.14. Predict the expected attenuation at 1.55 μm , and compare your calculation with the reported (measured) value of 0.19–0.20 dB km^{-1} for this fiber. Repeat the calculations at 1.31 μm , and compare your values with the reported 0.33–0.35 dB km^{-1} values.

Solution

First, we should check the fundamental infrared absorption at 1550 nm. Using Eq. (2.9.7)

$$\alpha_{\text{FIR}} = A \exp(-B/\lambda) = 7.8 \times 10^{11} \exp[-(48.5)/(1.55)] = 0.020 \text{ dB km}^{-1}$$

which is very small.

Next, for the Rayleigh scattering at 1550 nm, the simplest equation with $A_R = 0.90 \text{ dB km}^{-1} \mu\text{m}^4$, gives

$$\alpha_R = A_R/\lambda^4 = (0.90 \text{ dB km}^{-1} \mu\text{m}^4)/(1.55 \mu\text{m})^4 = 0.156 \text{ dB km}^{-1}$$

This equation is basically a rule of thumb. The total attenuation is then

$$\alpha_R + \alpha_{\text{FIR}} = 0.156 + 0.02 = 0.176 \text{ dB km}^{-1}$$

The above “rule of thumb” predicted value is very close to that reported for the Corning SMG28e⁺ SMF.

The current fiber has NA = 0.14. Thus

$$A_R = 0.63 + 2.06 \times \text{NA} = 0.63 + 2.06 \times 0.14 = 0.918 \text{ dB km}^{-1} \mu\text{m}^4$$

i.e.,

$$\alpha_R = A_R/\lambda^4 = (0.918 \text{ dB km}^{-1} \mu\text{m}^4)/(1.55 \mu\text{m})^4 = 0.159 \text{ dB km}^{-1}$$

which gives a total attenuation of $0.159 + 0.020$ or 0.179 dB km^{-1} .

We can repeat the above calculations at $\lambda = 1.31 \mu\text{m}$. However, we do not need to add α_{FIR} .

$$\alpha_R = A_R/\lambda^4 = (0.90 \text{ dB km}^{-1} \mu\text{m}^4)/(1.31 \mu\text{m})^4 = 0.306 \text{ dB km}^{-1}$$

and using the NA-based value for A_R from above

$$\alpha_R = A_R/\lambda^4 = (0.918 \text{ dB km}^{-1} \mu\text{m}^4)/(1.31 \mu\text{m})^4 = 0.312 \text{ dB km}^{-1}$$

Both close to the measured value.

D. Bending losses

External factors, such as bending, can also lead to attenuation in the optical fiber. **Bending losses** are those losses that arise whenever a fiber is bent as illustrated in Figure 2.36. Bending losses are classified into two broad categories: microbending and macrobending losses, which are illustrated in Figure 2.36. **Microbending loss** occurs whenever the radius of curvature of the bend is sharp, that is, when the bend radius is comparable to the diameter of the fiber as shown in Figure 2.36 (a). Typically microbending losses are significant when the radius of curvature of the bend is less than $0.1\text{--}1 \text{ mm}$. They can arise from careless or poor cabling of the fiber or even from flaws in manufacturing that result in variations in the fiber geometry over small distances. **Macrobending losses** are those losses that arise when the bend is much larger than the fiber size, typically much greater than 1 mm , as illustrated in Figure 2.36 (b). They typically occur when the fiber is bent during the installation of a fiber optic link such as turning the fiber around a corner. There is no simple precise and sharp boundary line between microbending and macrobending loss definitions. Both losses essentially result from changes in the waveguide geometry and properties as the fiber is subjected to external forces that bend the fiber. Typically, macrobending loss crosses over into microbending loss when the radius of curvature becomes less than a few millimeters. Below we simply use the term *bending loss* to describe those significant losses arising from the bending of the fiber.

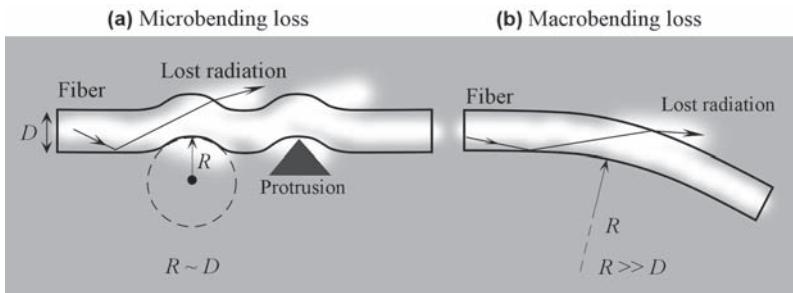


FIGURE 2.36 Definitions of (a) microbending and (b) macrobending loss and the definition of the radius of curvature, R . (A schematic illustration only.) The propagating mode in the fiber is shown as white-shaded area. Some radiation is lost in the region where the fiber is bent. D is the fiber diameter, including the cladding.

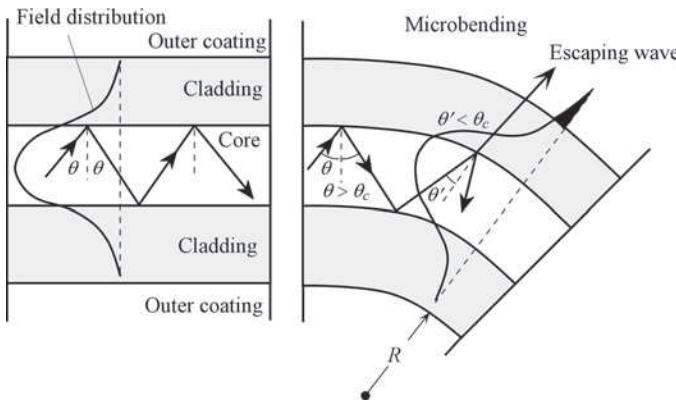


FIGURE 2.37 Sharp bends change the local waveguide geometry that can lead to waves escaping. The zigzagging ray suddenly finds itself with an incidence angle smaller than θ' that gives rise to either a transmitted wave, or to a greater cladding penetration; the field reaches the outside medium and some light energy is lost.

When a fiber is bent, the guide geometry and the refractive index profile change locally, which leads to some of the light energy radiating away from the guiding direction. A sharp bend, as illustrated intuitively in Figure 2.37, will change the local waveguide geometry in such a way that a zigzagging ray suddenly finds itself with an incidence angle θ' , narrower than its normal angle θ ($\theta' < \theta$), which gives rise to either a transmitted wave (a refracted wave into the cladding) or to a greater cladding penetration. If $\theta' < \theta_c$, the critical angle (which would have been modified by the bent), then there will be no total internal reflection and substantial light power will be radiated into the cladding where it will be lost, by, for example, reaching the outer coating (polymer coating, etc.) where losses are significant. Attenuation increases sharply with the extent of bending; as θ' gets narrow and TIR is lost, substantially more energy is transferred into the cladding. Further, highest modes propagate with incidence angles θ close to θ_c , which means that these modes are most severely affected. Multimode fibers therefore suffer more from bending losses than single-mode fibers.

While Figure 2.37 provides an intuitive explanation, it is important to understand bending losses in terms of the mode field diameter of the fiber and what happens to the evanescent wave in the cladding upon bending. When a fiber is bent sharply, the propagating wavefront along the straight fiber cannot bend around and continue as a wavefront because a portion of it beyond some critical radial distance r_c , as shown in Figure 2.38, has to travel a *longer* distance and to

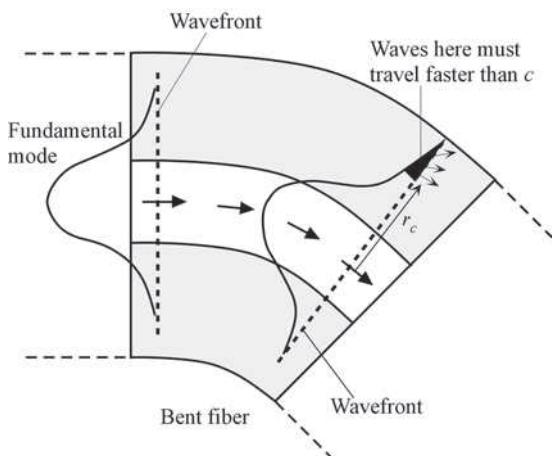


FIGURE 2.38 When a fiber is bent sharply, the propagating wavefront along the straight fiber cannot bend around and continue as a wavefront because a portion of it (black shaded) beyond the critical radial distance r_c must travel faster than the speed of light in vacuum. This portion is lost in the cladding, that is, radiated away.

stay in phase it must travel faster than the speed of light in vacuum, c . Since this is impossible, this portion is lost in the cladding, *i.e.*, radiated away from the propagating mode, which reduces the power in it.

The bending loss α_B increases rapidly with increasing bend sharpness, that is, with decreasing radius of bend curvature, R . Figure 2.39 (a) shows how the bending loss α_B depends on the radius of curvature R for three types of single-mode fibers: a standard SMF and two trench fibers. Trench fibers have the refractive index profile as shown in Figure 2.39 (b). There is a refractive index trench placed in the cladding where the refractive index n_3 is lower than the cladding index (n_2). Note also that the outer cladding, after the trench, has a different refractive index, n_4 . The confinement of the light wave in trench fibers is such that they are relatively insensitive to bending compared to standard fibers. As apparent, α_B increases exponentially with decreasing R as shown in Figure 2.39 (a), which is a semilogarithmic plot, that is,

Bending loss

$$\alpha_B = A \exp(-R/R_c) \quad (2.9.9)$$

where A and R_c are constants that depend on the fiber properties. Note that α_B is normally reported in dB/turn of fiber for a given radius of curvature as in Figure 2.39 (a).

For single-mode fibers, a quantity called a **MAC-value**, or a **MAC-number**, N_{MAC} , has been used to characterize the bending loss. N_{MAC} is defined by

MAC number

$$N_{\text{MAC}} = \frac{\text{Mode field diameter}}{\text{Cutoff wavelength}} = \frac{\text{MFD}}{\lambda_c} \quad (2.9.10)$$

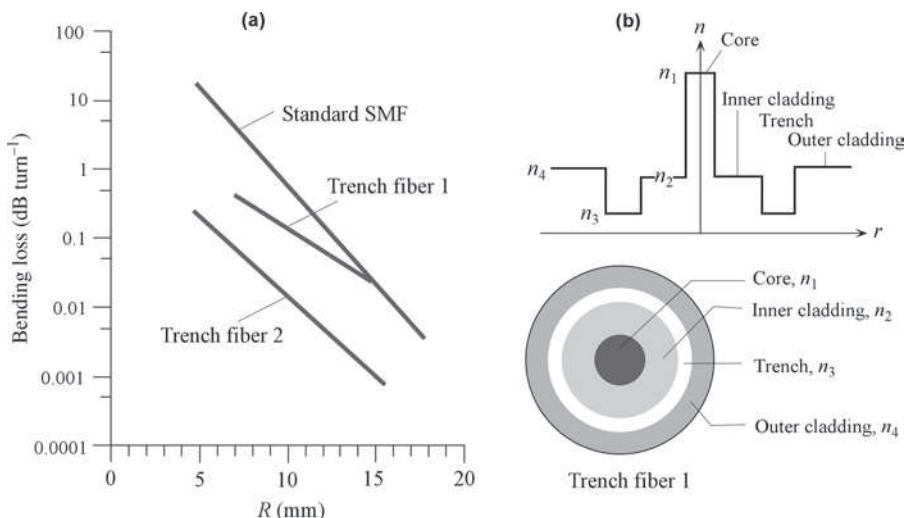


FIGURE 2.39 (a) Bending loss in dB per turn of fiber for three types of fibers: standard single mode and two trench fibers, around 1.55–1.65 μm . (b) The index profile for the trench fiber 1 in (a), and a schematic view of the fiber cross-section. Experimental data used to generate the plots have been combined from various sources. (Source: Standard fiber, M.-J. Li *et al.*, *J. Light Wave Technol.*, 27, 376, 2009; Trench fiber 1, K. Himeno *et al.*, *J. Light Wave Technol.*, 23, 3494, 2005; Trench fiber 2, L.-A. de Montmorillon *et al.*, “Bend-Optimized G.652D Compatible Trench-Assisted Single-Mode Fibers,” *Proceedings of the 55th IWCS/Focus*, pp. 342–347, November, 2006.)

The above definition is based on observations that α_B generally increases with increasing MFD and decreasing cutoff wavelength. The bending loss increases exponentially with the MAC-number. It is not a universal function, and different types of fibers, with very different refractive index profiles, can exhibit different α_B vs. N_{MAC} behavior. N_{MAC} values are typically in the range 6–8.

It is clear from Figure 2.38 that bending losses involve the escape of power from the propagating mode when the waves beyond the critical distance r_c in the cladding cannot keep up with the overall propagation of the mode. The extent of penetration into the cladding is therefore important. It should, in principle, be possible to shape the refractive index profile in the “cladding” to better contain the radiation to the core, and hence avoid losses. Indeed, bending losses can be reduced by designing a suitable cladding that constrains the power in the propagating mode along the core even when the fiber is bent. The simple step-index fiber cannot constrain the power to stay within the core when the fiber is sufficiently bent. The trench fibers shown in Figure 2.39 (b) are able to reduce the bending loss by ensuring that the fiber has to be bent much more sharply than the standard fiber for the field to reach r_c . Thus, the optical power is better confined within the core region.

Bending loss is of particular interest in MMFs. Since MMFs are used in short-haul networks such as links within buildings, they invariably become bent in installation. Technicians would like to bend a fiber just like a copper or a coaxial cable in laying down connections. Bending losses of a fiber are typically measured by wounding the fiber for N turns on a drum or a mandrel of certain radius R , and then measuring the attenuation of the fiber. The standard OM1 MMF for use at 850 nm and 1310 nm optical links has been specified to have a maximum bending (macrobending) loss of 0.5 dB when the fiber is wound 100 turns with a radius 75 mm, *i.e.*, a bending loss of $0.005 \text{ dB turn}^{-1}$ for a bend radius of 75 mm. Bend-insensitive fibers have been designed to have lower bend losses. For example, some fiber manufacturers specify the allowed bend radius for a given level of attenuation at a certain wavelength (*e.g.*, 1310 nm).

EXAMPLE 2.9.3 Bending loss for SMF

Experiments on a standard SMF operating around 1550 nm have shown that the bending loss is $0.124 \text{ dB turn}^{-1}$ when the bend radius is 12.5 mm and $15.0 \text{ dB turn}^{-1}$ when the bend radius is 5.0 mm. What is the loss at a bend radius of 10 mm?

Solution

We apply Eq. (2.9.9), $\alpha_B = A \exp(-R/R_c)$, for the two set of values given,

$$0.124 \text{ dB turn}^{-1} = A \exp[-(12.5 \text{ mm})/R_c] \quad \text{and} \quad 15.0 \text{ dB turn}^{-1} = A \exp[-(5.0 \text{ mm})/R_c]$$

We have two equations, and two unknowns, A and R_c . Dividing the first by the second and separating out R_c we find

$$R_c = (12.5 \text{ mm} - 5.0 \text{ mm})/\ln(15.0/0.124) = 1.56 \text{ mm}$$

and

$$A = (15.0 \text{ dB turn}^{-1}) \exp[(5.0 \text{ mm})/(1.56 \text{ mm})] = 370 \text{ dB turn}^{-1}$$

Thus, at $R = 10 \text{ mm}$, $\alpha_B = A \exp(-R/R_c) = (370) \exp[-(10 \text{ mm})/(1.56 \text{ mm})] = 0.61 \text{ dB turn}^{-1}$.

The experimental value is also $0.61 \text{ dB turn}^{-1}$ to within two decimal places.

2.10 FIBER MANUFACTURE

A. Fiber Drawing

There are a number of processes for producing optical fibers for various applications. We will consider the **outside vapor deposition** (OVD) technique, which is one of the widely used processes and produces fibers with low loss. It is also known as the outside vapor phase oxidation process.

The first step is to prepare a **preform**, which is a glass rod that has the right refractive index profile across its cross-section and the right glass properties (*e.g.*, negligible amounts of impurities). This rod is typically 10–30 mm in diameter and about 1–2 m in length. The optical fiber is drawn from this preform using special fiber-drawing equipment that is schematically illustrated in Figure 2.40.

The preform rod is slowly fed into a hot furnace that has a hot zone around 1900–2000°C, where the glass flows like a viscous melt (resembling honey). As the rod reaches the hot zone and its end begins to flow, its tip is pulled, with just the right tension, to come out as a fiber and is spooled on a rotating take-up drum. The diameter of the fiber must be strictly controlled to achieve the required waveguide characteristics. An optical thickness monitor gauge provides information on the changes of the fiber diameter that is used (in an automatic feedback control system) to adjust the speed of the fiber-winding mechanism and the speed of the preform feeder to maintain a constant fiber diameter, typically better than 0.1%. In some cases, the preform is hollow, that is, it has a thin central hole along the rod axis. The hollow simply collapses during the drawing and does not affect the final drawn fiber.

It is essential that, as soon as the fiber is drawn, it is coated with a polymeric layer (*e.g.*, urethane acrylate) to mechanically and chemically protect the fiber surface. When a bare fiber glass surface is exposed to ambient conditions, it quickly develops various microcracks on

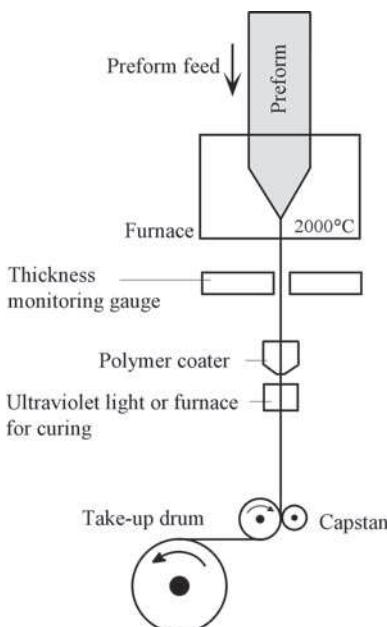


FIGURE 2.40 Schematic illustration of a fiber drawing tower.

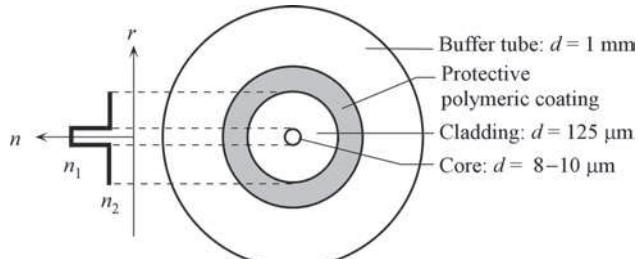


FIGURE 2.41 The cross-section of a typical single-mode fiber with a tight buffer tube (d = diameter).

the surface and these dramatically reduce the mechanical strength (fracture strength) of the fiber. The applied polymeric coating is initially a viscous liquid and needs to be cured (hardened), which is done as the coated fiber passes through a curing oven, or with ultraviolet lamps if it is UV hardenable. Sometimes two layers of polymeric coating are applied. Cladding is typically 125 μm for many types of fibers, and the overall diameter with the polymeric coating can be 250–500 μm . A schematic diagram of the cross-section of a typical single-mode optical fiber is shown in Figure 2.41. In this example, there is a thick polymeric buffer tube, or a buffer jacket, surrounding the fiber and its coating to cushion the fiber against mechanical pressure and microbending (sharp bending). Some fibers are buffered by having the fiber loose within a buffer tube. The tube may then contain a filling compound to increase the buffering ability. Single and multiple fibers are invariably used in cable form and the structure of the cable depends on the application (e.g., long-haul communications), the number of fibers carried, and the cable environment (e.g., underground, underwater, overhead).

B. Outside Vapor Deposition

Outside vapor deposition (OVD) is one of the vapor deposition techniques used to produce the rod preform used in fiber drawing. It was mainly pioneered by Donald Keck, Bob Maurer, and Peter Schultz at Corning during the late 1970s. The OVD process is illustrated in Figure 2.42 and has two stages. The first **laydown** stage involves using a ceramic rod such as alumina (or even a graphite rod) as a target rod, also known as the *bait rod*, as shown Figure 2.42 (a). This acts as a mandrel and is rotated. The required glass material for the preform with the right composition is grown on the outside surface of this target rod by depositing glass soot particles. The deposition is achieved by burning various gases in an oxyhydrogen burner (torch) flame where glass soot is produced as reaction products.

Suppose that we need a preform with a core that has germania (GeO_2) in silica glass so that the core has a higher refractive index. The required gases, SiCl_4 (silicon tetrachloride), GeCl_4 (germanium tetrachloride), and fuel in the form of oxygen (O_2) and methane (CH_4) or hydrogen (H_2)

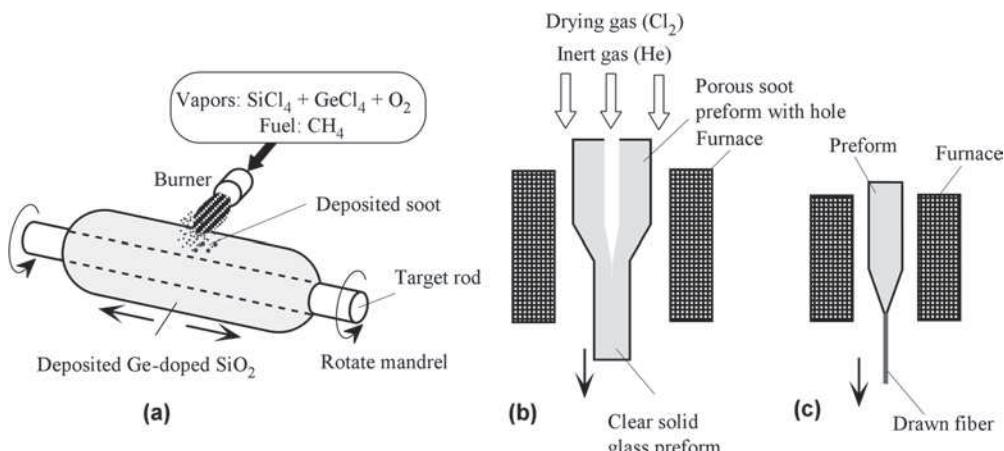
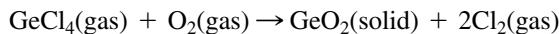
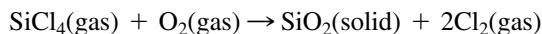


FIGURE 2.42 Schematic illustration of OVD and the preform preparation for fiber drawing. (a) Reaction of gases in the burner flame produces glass soot that deposits on to the outside surface of the mandrel. (b) The mandrel is removed and the hollow porous soot preform is consolidated; the soot particles are sintered, fused together, to form a clear glass rod. (c) The consolidated glass rod is used as a preform in fiber drawing.

Donald Keck, Bob Maurer, and Peter Schultz (left to right) at Corning shortly after announcing the first low loss optical fibers made in 1970. Keck, Maurer, and Schultz developed the outside vapor deposition (OVD) method for the fabrication of preforms that are used in drawing fibers with low losses. Their OVD was based on Franklin Hyde's vapor deposition process earlier at Corning in 1930s. OVD is still used today at Corning in manufacturing low loss fibers. (*Courtesy of Corning.*)



are burnt in a burner flame over the target rod surface as shown in Figure 2.42 (a). The important reactions of the gases in the flame are



These reactions produce fine glass particles, silica and germania, called “soot” that deposit on the outside surface of the target rod and form a porous glass layer as the burner travels along the mandrel. The glass preform is built layer by layer by slowly running the burner up and down along the length of the rotating mandrel either by moving the mandrel, or moving the burner (the same result). First, the layers for the core region are deposited and then the gas composition is adjusted to deposit the layers for the cladding region. Typically, there may be about 200 layers in the final preform. The composition and hence the refractive index of each layer can of course be controlled by adjusting the relative amounts of SiCl_4 and GeCl_4 gases fed into the burner for the chemical reaction. Indeed, any desired refractive index profile, in principle, can be obtained by controlling the layer compositions.

Peter Schultz making a germania-doped multimode fiber preform using the outside vapor deposition (OVD) process around 1972 at Corning. Soot is deposited layer by layer on a thin bait rod rotating and translating in front of the flame hydrolysis burner. The first fibers made by the OVD at that time had an attenuation of 4 dB km^{-1} , which was among the lowest, and below what Charles Kao thought was needed for optical fiber communications, 20 dB km^{-1} . (*Courtesy of Peter Schultz.*)



EXAMPLE 2.10.1 Fiber drawing

In a certain fiber production process a preform of length 110 cm and diameter 20 mm is used to draw a fiber. Suppose that the fiber drawing rate is 5 m s^{-1} . What is the maximum length of fiber that can be drawn from this preform if the last 10 cm of the preform is not drawn and the fiber diameter is $125 \mu\text{m}$? How long does it take to draw the fiber?

Solution

We assume that a length L_p of the preform is drawn as fiber. Since the density is the same and mass is conserved, the volume must also be conserved. If d_f and d_p are the fiber and preform diameters, and L_f and L_p are the lengths of the fiber and drawn preform (110 cm – 10 cm), respectively, then,

$$L_f d_f^2 = L_p d_p^2$$

i.e.,

$$L_f = \frac{(1.1 - 0.1 \text{ m})(20 \times 10^{-3} \text{ m})^2}{(125 \times 10^{-6} \text{ m})^2} = 25,600 \text{ m} \quad \text{or} \quad 25.6 \text{ km}$$

Since the rate is 5 m s^{-1} , the time it takes to draw the fiber in minutes is

$$\text{Time (hrs)} = \frac{\text{Length (km)}}{\text{Rate (km hr}^{-1}\text{)}} = \frac{25,600 \text{ m}}{(5 \text{ m s}^{-1})(60 \times 60 \text{ s hr}^{-1})} = 1.4 \text{ hrs}$$

Typical drawing rates are in the range $5\text{--}20 \text{ m s}^{-1}$ so that 1.4 hrs would be on the long-side.

Additional Topics

2.11 WAVELENGTH DIVISION MULTIPLEXING: WDM

A simplified block diagram of a **wavelength division multiplexed** optical communication system is shown in Figure 2.43. N optical information channels are generated by modulating N light waves, each with a different wavelength from λ_1 to λ_N . The N optical channels are then multiplexed into a single fiber for transmission. Optical amplifiers are used at various points along the transmission medium to amplify the attenuated signal. If a particular information channel, for example, the third channel at λ_3 , needs to be dropped and another information channel (at the same wavelength) needs to be added, then this can be done by using a **drop-add multiplexer**. The new channel is labeled as λ'_3 . At the destination, the optical signal has to be separated into its constituent N channels by a **wavelength demultiplexer**. Each channel is then fed into a receiver which, like the transmitter, can be quite complicated in terms of electronics.

An examination of WDM system in Figure 2.43 illustrates a number of important basic functions: (a) generation of different wavelengths of light each with a narrow spectrum to avoid any overlap in wavelengths; (b) modulation of light without wavelength distortion, that is, without *chirping* (variation in the frequency of light due to modulation); (c) efficient coupling of different wavelengths into a single transmission medium; (d) optical amplification of all the wavelengths by an amount that compensates for attenuation in the transmission medium, which

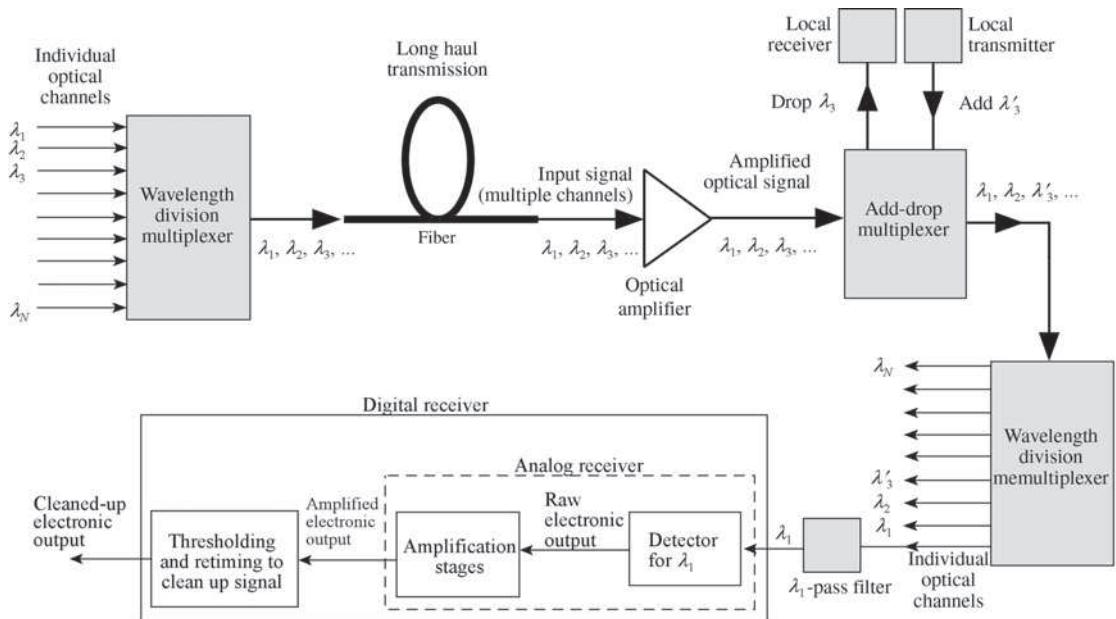


FIGURE 2.43 Wavelength division multiplexing, an example of an optoelectronic system.

depends on the wavelength; (e) dropping and adding channels when necessary during transmission; (f) demultiplexing the wavelengths into individual channels at the receiving end; (g) detecting the signal in each channel. To achieve an acceptable bandwidth, we need to dispersion-manage the fiber (use dispersion compensation fibers), and to reduce cross-talk and unwanted signals, we have to use optical filters to block or pass the required wavelengths. We need various optical components to connect the devices together and implement the whole system.

Channel density is an important quantity in designing WDM systems. It is usually measured in terms of frequency spacing between two neighboring optical channels. If the frequency spacing $\Delta\nu$ is less than 200 GHz then WDM is called **dense wavelength division multiplexing and denoted** as DWDM. At present, DWDM stands typically at 100 GHz separated channels, which is equivalent to a wavelength separation of 0.8 nm. DWDM imposes stringent requirements on lasers and modulators used for generating the optical signals. It is not possible to tolerate even slight shifts in the optical signal frequency when channels are spaced closely. As the channel spacing becomes narrower as in DWDM, one also encounters various other problems not previously present. For example, any nonlinearity in a component carrying the channels can produce intermodulation between the channels—an undesirable effect. Thus, the total optical power must be kept below the onset of nonlinearity in the fiber and optical amplifiers within the WDM system. (Nonlinear effects are discussed in Section 2.12.)

The WDM example in Figure 2.43 is one of many examples on optoelectronic systems. The complexity of such systems depends not only on whether individual optoelectronic devices can efficiently implement the required function, but also on the availability of various optical components that are needed to properly interconnect the devices. Optoelectronics today is an interdisciplinary activity. Advances at the device level are enabling systems that would have been thought impossible a couple of decades ago (DWDM being an excellent example).

2.12 NONLINEAR EFFECTS IN OPTICAL FIBERS AND DWDM

The refractive index n of the optical fiber material is not only a constant material property at a given wavelength but also depends on the intensity of light. As we know from Chapter 1, n depends on the relative permittivity through $n = \epsilon_r^{1/2}$. For **linear materials**, or at sufficiently low optical fields, ϵ_r is constant. If we consider just one channel at a frequency ν_1 , the induced polarization P in the medium is proportional to the field E , that is, $P = \epsilon_o(\epsilon_r - 1)E$. This represents a linear response. At high fields, ϵ_r becomes field dependent and the polarization P depends not only on E but also on E^2 and E^3 , and the material is said to be **nonlinear**; the medium responds nonlinearly. Put differently,

$$P = a_1 E + a_2 E^2 + a_3 E^3 \quad (2.12.1)$$

Nonlinear behavior

where a_1 , a_2 , and a_3 are the coefficients in the expansion of P in terms of E . We ignored the higher-order terms. The *effect*, the polarization P , that is, the response, is a function of the *cause* E . For isotropic materials such as glasses, a_2 must be zero, as explained in Chapter 6, but a_3 , though small, is finite.²⁵ Thus

$$P = a_1 E + a_3 E^3 \quad (2.12.2)$$

Nonlinear behavior for glasses

If we now have three channels with fields E_1 , E_2 , and E_3 at different frequencies ν_1 , ν_2 , and ν_3 , respectively, then the polarization P would also have a nonlinear term proportional to $a_3(E_1 + E_2 + E_3)^3$ as apparent from Eq. (2.12.2). The term $a_3(E_1 + E_2 + E_3)^3$ would have various cross product terms of the type $E_1 E_2 E_3$, etc. The channel E_2 at ν_2 can therefore modulate the channel E_1 at ν_1 and E_3 at ν_3 , which is called **cross-modulation**. It arises from the nonlinearity of the medium.

Further, the $E_1 E_2 E_3$ term is proportional to $\cos(2\pi\nu_1 t) \times \cos(2\pi\nu_2 t) \times \cos(2\pi\nu_3 t)$. The product of cosines can be written in terms of *sums* of cosines (from straightforward trigonometry) with frequencies that correspond to various sum and difference frequencies, i.e., $\nu_1 + \nu_2 + \nu_3$ and $\nu_1 + \nu_2 - \nu_3$. This type of mixing that occurs at high optical powers leads to a phenomenon called four-photon mixing.

In **four-photon (wave) mixing**, the nonlinearity of the glass medium results in three signal frequencies, ν_1 , ν_2 , and ν_3 , as above mixing together to generate a fourth frequency ν_4 that acts as a noise source on another channel in the form of cross channel interference. Consider three optical signals at ν_1 , $\nu_2 = \nu_1 + 100$ GHz, and $\nu_3 = \nu_1 + 200$ GHz. The nonlinearity in the optical fiber mixes signals to generate a new signal at $\nu_4 = \nu_1 + \nu_2 - \nu_3 = \nu_1 - 100$ GHz, on top of the channel below ν_1 . Four-photon mixing thus adds noise into the communications system.

The refractive index n depends on the relative permittivity ϵ_r , which depends on the polarization to the field ratio, so that

$$n^2 = \epsilon_r = \frac{P}{\epsilon_o E} + 1 = \left(\frac{a_1}{\epsilon_o} + 1 \right) + \left(\frac{a_3}{\epsilon_o} \right) E^2$$

We can now readily obtain n by square-rooting the right-hand side. Since the $a_3 E^2$ term is small, and E^2 represents the light intensity, a simple expansion gives n as

$$n = n_o + n_2 I \quad (2.12.3)$$

Nonlinear refractive index

²⁵It can easily be seen that a_2 must be zero for glasses that are isotropic—the same properties in all directions. When we change the direction of the field and use a negative field, $-E$, in Eq. (2.12.1), we should get the same magnitude for P but in the reverse direction, that is, it should be $-P$. This is only possible if $a_2 = 0$.

where n_o is the constant refractive index at low intensities and n_2 is a coefficient that gauges the extent of the nonlinear effect and is called the **nonlinear refractive index**.

The phase ϕ of a propagating wave at a point z depends on $n(2\pi/\lambda)z$ where λ is the free-space wavelength. The change Δn in n with the light intensity I changes the phase by $\Delta\phi$, that is,

$$\Delta\phi = \frac{2\pi z \Delta n}{\lambda} = \frac{2\pi z n_2 I}{\lambda} \quad (2.12.4)$$

Self-phase modulation

The dependence of the phase of a wave on its light intensity, as expressed in Eq. (2.12.4), represents **self-phase modulation** (SPM). Phase changes result in frequency changes since, by definition, the frequency $\omega = d\phi/dt$. The shift in the frequency $\Delta\omega$, due to $\Delta\phi$ is

$$\Delta\omega = \frac{d\Delta\phi}{dt} = \frac{2\pi z n_2}{\lambda} \frac{dI}{dt} \quad (2.12.5)$$

Frequency chirp

The initial rising intensity of the pulse increases the frequency and falling intensity in the pulse tail decreases the frequency as shown in Figure 2.44. The frequency becomes **chirped**, that is, it varies along the pulse. The frequency spectrum of the pulse becomes broader, an undesirable effect in DWDM.

Further, changes in n with the light intensity cause changes in the group velocity with intensity. The waves at the peak intensity travel more slowly. Thus, there is an additional pulse dispersion mechanism arising from self-phase modulation, called **SPM dispersion**.

Cross-phase modulation (XPM) occurs when two light pulses (say from different channels) are present at the same time, so-called overlapping pulses. The high intensity in one light pulse modifies the index n and hence affects the phase of the other pulse. One pulse distorts the other when both are at the same place and leads to “cross talk,” or interchannel interference.

While SPM appears as an undesirable phenomenon, it can also have a number of useful applications. For example, it can be used to generate short pulses, all optical switches, wavelength conversion, demultiplexing among other beneficial applications.

There are limits to how much optical power that can be efficiently transmitted in a fiber. The transmitted optical power through a fiber does not increase linearly with the input power when the latter is sufficiently high to cause **stimulated Brillouin scattering** (SBS). SBS occurs when at sufficiently high optical powers, the interaction of a propagating EM wave with acoustic vibrations in the glass leads to some of the optical power being reflected back.

From basic physics, we know that atoms in the solid exhibit thermal vibrations. These atomic vibrations give rise to traveling waves in the bulk, which are normally described by phonons. Collective vibrations of the atoms in a solid give rise to lattice waves²⁶ inside the

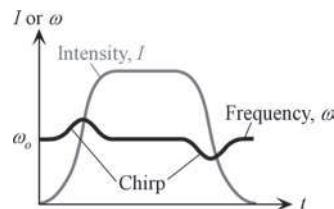


FIGURE 2.44 Schematic illustration of the effect of self-phase modulation on the frequency and the resulting chirps.

²⁶Even if the solid is not crystalline, the term *lattice wave* nonetheless is used to describe waves in the glass that arise from atomic vibrations.

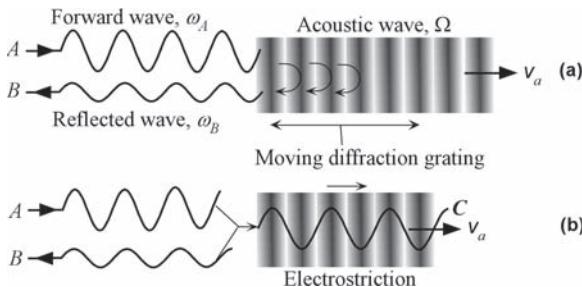


FIGURE 2.45 (a) Scattering of a forward-traveling EM wave A by an acoustic wave results in a reflected, backscattered, wave B that has a slightly different frequency, by Ω . (b) The forward and reflected waves, A and B , respectively, interfere to give rise to a standing wave that propagates at the acoustic velocity v_a . As a result of electrostriction, an acoustic wave is generated that reinforces the original acoustic wave—stimulates scattering.

solid; a sound wave is an example of a lattice wave that has a long wavelength. Acoustic lattice waves in a solid involve periodic strain variations along the direction of propagation. Changes in the strain result in changes in the refractive index through a phenomenon called the **photoelastic effect**. (The refractive index depends on strain.) Thus, there is a periodic variation in the refractive index which moves with an acoustic velocity v_a as illustrated in Figure 2.45 (a). This moving diffraction grating reflects back some of the forward propagating EM wave A to give rise to a *back scattered* wave B as shown in Figure 2.45 (a). The frequency ω_B of the back scattered wave B is Doppler shifted from that of the forward wave ω_A by the frequency Ω of the acoustic wave, *i.e.*, $\omega_B = \omega_A - \Omega$.

The forward wave A and the back scattered wave B interfere and give rise to a standing wave C . This standing wave C represents a periodic variation in the field that moves with a velocity v_a along the direction of the original acoustic wave as shown in Figure 2.45 (b). The field variation in C produces a periodic displacement of the atoms in the medium, through a phenomenon called **electrostriction**. (The application of an electric field causes a substance to change shape, that is, experience strain.) Therefore, a periodic variation in strain develops, which moves at the acoustic velocity v_a . The moving strain variation is really an acoustic wave, which reinforces the original acoustic wave and stimulates more back scattering. Thus, it is clear that a condition can be easily reached in which Brillouin scattering stimulates further scattering: **stimulated Brillouin scattering (SBS)**.

The SBS effect increases as the input light power increases, and the spectral width of the input light becomes narrower. The onset of SBS depends not only on the fiber type and core diameter, but also on the spectral width $\Delta\lambda$ of the laser output spectrum that is coupled into the fiber. SBS is enhanced as the laser spectral width $\Delta\lambda$ is narrowed or the duration of light pulse is lengthened. For example, for a directly modulated laser diode emitting at 1550 nm into a single-mode fiber, the onset of SBS is expected to occur at power levels greater than 20–30 mW. In DWDM systems with externally modulated lasers, that is, narrower $\Delta\lambda$, the onset of SBS can be as low as ~ 10 mW. SBS is an important limiting factor in transmitting high power signals in WDM systems.

2.13 BRAGG FIBERS

Bragg fibers have a core region surrounded by a cladding that is made up of concentric layers of high low refractive index dielectric media as shown in Figure 2.46 (a). The core can be a low refractive index solid material or simply hollow. In the latter case, we have a **hollow Bragg fiber**. The high (n_1) and low (n_2) alternating refractive index profile in Figure 2.46 (b) constitutes a Bragg grating which acts as a dielectric mirror. There is a band of wavelengths, forming a **stop band**, that are not allowed to propagate into the Bragg grating. We can also view the periodic variation in n as forming a *photonic crystal cladding in one-dimension*, along the radial direction,

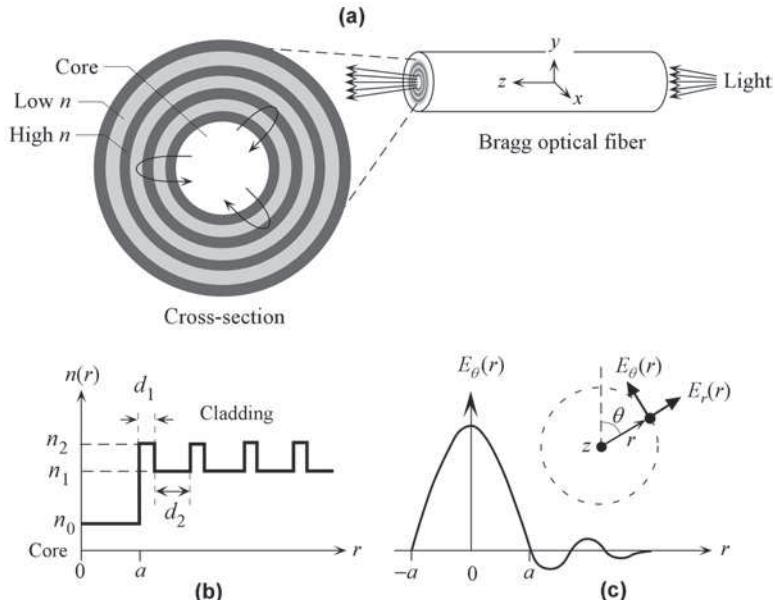


FIGURE 2.46 (a) A Bragg fiber and its cross-section. The Bragg grating in the cladding can be viewed to reflect the waves back into the core over its stop band. (b) The refractive index variation (exaggerated). (c) Typical field distribution for the circumferential field E_θ .

with a stop band, that is, a photonic bandgap. Consequently light is bound within the core of the guide for wavelengths within this stop band. Light can only propagate along z .

It is apparent that the waveguiding principle in the present case is not based on total internal reflection but the containment of light in the core by a concentric dielectric mirror. Optimal optical confinement in the core is obtained when the cladding is such that the reflections from the cladding are maximized. We know from Chapter 1 that the reflectance of a dielectric stack mirror (Bragg reflector) is maximum when the thickness of each layer is a quarter of wavelength. If d_1 and d_2 are thicknesses of layers n_1 and n_2 , respectively, and λ is the free-space wavelength, then $d_1 = \lambda/n_1$ and $d_2 = \lambda/n_2$. Consequently, we need $d_1n_1 = d_2n_2$. As illustrated in Figure 2.46 (c), some of the field obviously penetrates into the cladding but its magnitude decays after several layers.

One of the advantages of Bragg fibers is that they can be made to have a hollow core. A hollow core would allow higher optical powers to be used without the adverse effects of glass nonlinearities. In principle, the attenuation can be made very small but there would still be some scattering from the small irregularities in the smoothness of the inner cladding interface along the fiber. Further, the material dispersion in the hollow core is smaller than that in a glass core (there is still some field penetrating into the cladding), and allows dispersion-flattened Bragg fibers to be designed. Bragg fibers, like photonic crystal fibers below, are currently specialty fibers.

2.14 PHOTONIC CRYSTAL FIBERS—HOLEY FIBERS

Photonic crystal fibers (PCFs) are fibers that have been prepared to have a periodic array of holes in the cladding region while the core region is either solid or hollow as shown in Figures 2.47 (a) to (c). The holes run like thin hollow capillary tubes along the length of the fiber. The situation is quite different than the Bragg fiber because the cladding is now a two-dimensional (2D) photonic crystal not 1D. The principles involved in the guiding of light along the core of a PCF are different

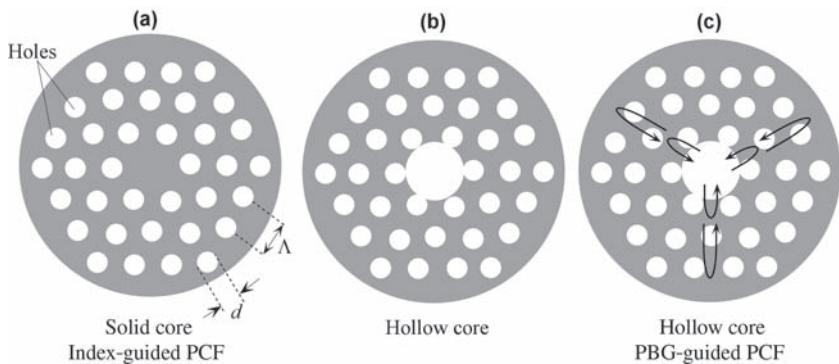


FIGURE 2.47 (a) A solid core PCF. Light is index guided. The cladding has a hexagonal array of holes. d is the hole diameter and Λ is the array pitch, spacing between the holes; (b) and (c) a hollow core PCF. Light is photonic bandgap (PBG) guided.

for the solid and hollow core PCFs. The PCFs shown in Figure 2.47 have a hexagonal lattice of holes in the cladding; other lattice types are also possible.

The guiding of light along the solid core PCF is straightforward to understand. Both the core and cladding use the same material, usually silica, but the air holes in the cladding result in an effective refractive index that is lower than the solid core region. The cladding has a lower effective refractive index than the core, and the whole structure then mimics a step-index fiber. Total internal reflection then allows the light to be propagated just as in a step-index standard silica fiber. Light is **index guided**. There are, however, distinct advantages. The solid can be pure silica, rather than germania-doped silica, and hence exhibits lower scattering loss. There is one distinct and unexpected advantage to having a photonic crystal as a cladding. It turns out that single-mode propagation can occur over a very large range of wavelengths, almost as if the fiber is **endlessly single mode** (ESM). The reason is that the PC in the cladding acts as a filter in the transverse direction, which allows the higher modes to escape (leak out) but not the fundamental mode. Figure 2.48 (a) and (b) illustrate the principle of ESM operation

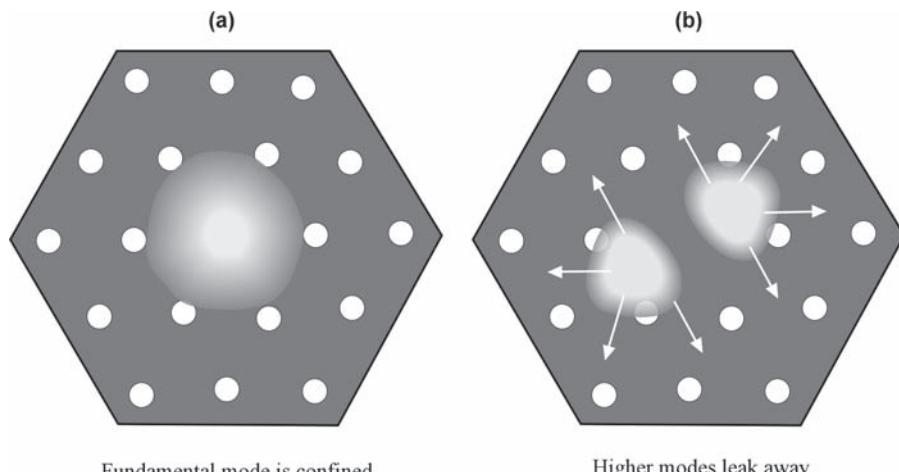


FIGURE 2.48 (a) The fundamental mode is confined. (b) Higher modes have more nodes and can leak away through the space.

by comparing the behavior of the fundamental and a higher mode. A high mode of radiation has many nodes (zeros) in the field, that is, it is made of many small volumes of optical power. These small volumes can “leak” through the spaces between the holes and be lost. The fundamental mode is stuck in the core because of its large volume of optical power; it is unable to “squeeze” through the holes. The PC in the cladding in Figure 2.47 (a) has a hexagonal lattice of holes in which the hole diameter is d and the spacing between the holes (pitch) is Λ , which represents the periodicity of the lattice. If the ratio d/Λ is kept below 0.43, then the light propagation is ESM. Other periodic arrangements of holes are, of course, also possible, and provide more flexibility in PCF design.

As a result of ESM operation, the core of a PCF can be made quite large without losing the single-mode operation. The refractive index difference can also be made large by having holes in the cladding. Consequently, PCFs can have high numerical apertures and large core areas; thus, more light can be launched into a PCF. Further, the manipulation of the shape and size of the hole and the type of lattice (and hence the periodicity, that is, the lattice pitch) leads to a much greater control of chromatic dispersion.

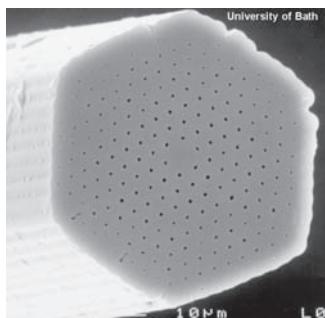
Hollow-core PCFs operate on a different principle. First, notice that the core has a lower refractive index (it is hollow) so that we cannot rely on total internal reflection (TIR) to explain light propagation. Light in the transverse direction is reflected back into the core over frequencies within the stop band, that is, photonic bandgap (PBG), of the photonic crystal in the cladding as illustrated in Figure 2.47 (c). When light is launched from one end of the fiber, it cannot enter the cladding (its frequency is within the stop band) and is therefore confined within the hollow core. Light is **photonic bandgap guided**. As recalled by Philip Russell, “My idea, then, was to trap light in a hollow core by means of a 2D photonic crystal of microscopic air capillaries running along the entire length of a glass fiber. Appropriately designed, this array would support a PBG for incidence from air, preventing the escape of light from a hollow core into the photonic-crystal cladding and avoiding the need for TIR.”²⁷ Thus, the periodic arrangement of the air holes in the cladding creates a photonic bandgap in the transverse direction that confines the light to the hollow core.

The hollow core provides certain distinct advantages. First, obviously, material dispersion is absent. Second, the attenuation in principle should be potentially very small since there is no Rayleigh scattering in the core. However, at present, scattering from irregularities in the air–cladding interface, that is, surface roughness, seems to limit the attenuation. A third advantage is that high powers of light can be launched without having nonlinear effects, such as stimulated Brillouin scattering, limits the propagation. There are other distinct advantages that are related to strong nonlinear effects, which are associated with the photonic crystal cladding.

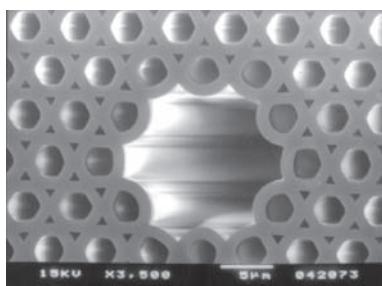
One can view both the solid and hollow core PCFs as a perfect photonic crystal in which a defect has been introduced along the center. The defect disrupts the periodicity in the core and runs along the fiber. We know from Chapter 1 that such defects in PCs can trap light, that is, confine light at the defect. Thus, light becomes confined in the defect region along core, and can only propagate along the core (see Figure 1.53).

PCFs can be manufactured by stacking together many capillary (thin hollow) tubes and the required number of thin solid tubes to represent the cross-section of the final fiber. These tubes are heated and fused together to make a preform. The preform is subsequently fed into a drawing furnace to draw out a PCF.

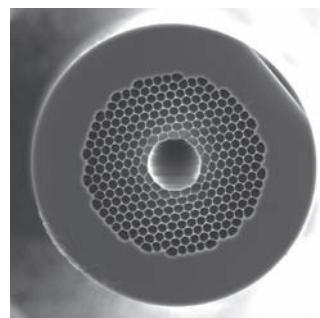
²⁷P. St. J. Russell, *J. Light Wave Technol.*, 24, 4729, 2006.



The first solid core photonic crystal fiber prepared by Philip Russell and coworkers at the University of Bath in 1996; an endlessly single-mode fiber.
(Courtesy of Philip Russell.)



One of the first hollow core photonic crystal fibers, guiding light by the photonic bandgap effect (1998).
(Courtesy of Philip Russell.)



A commercially available hollow core photonic crystal fiber from Blaze Photonics.
(Courtesy of Philip Russell.)



Philip Russell (center), currently at the Max Planck Institute for the Science of Light in Erlangen, Germany, led the team of two postdoctoral research fellows, Jonathan Knight (left) and Tim Birks (right) (both currently professors at the University of Bath in England), that carried out the initial pioneering research on photonic crystal fibers in the 1990s. (See reviews by P. St. J. Russell, *Science*, 299, 358, 2003; J. C. Knight, *Nature*, 424, 847, 2003. Courtesy of Philip Russell.)

2.15 FIBER BRAGG GRATINGS AND SENSORS

Fiber Bragg grating (FBG) is a fiber in which there is a region along its length that has been given a periodic variation in the refractive index of the core as illustrated in Figure 2.49. The periodic variation in the refractive index $n(z)$ along the core (along z in the figure) results in a Bragg diffraction grating, and since this is embedded within the fiber, the Bragg grating is called a **fiber Bragg grating**. The rectangular periodic variation in the core index shown in Figure 2.49 is highly idealized; the variation in the refractive index $n(z)$ is usually less sharp, often *sinusoidal*, and depends on the way in which the grating was inscribed into the core. For simplicity, an idealized variation is assumed in which, as shown in Figure 2.49, $n(z)$ changes as a step from n_1 to n'_1 , which is $n_1 + 2\Delta n$, and then back to n_1 over the period Λ . Δn is therefore the amplitude of the index change about the mean $(n_1 + n'_1)/2$. Typically the fiber is a single-mode fiber with a core diameter that is 5–10 μm and a total fiber diameter of 125 μm .

The periodic variation in n in the grating acts like a dielectric mirror in Chapter 1. Partial Fresnel reflections from the changes in the refractive index add in phase and give rise to a back reflected (back diffracted) wave only at a certain wavelength λ_B , called the **Bragg wavelength**, that satisfies

$$q\lambda_B = 2\bar{n}\Lambda \quad (2.15.1)$$

where q is an integer, 1, 2, ..., \bar{n} is the *effective* or *average refractive index* experienced by the propagating mode, also called the **modal index**, and Λ is the periodicity of the grating or the

Bragg wavelength,
FBG

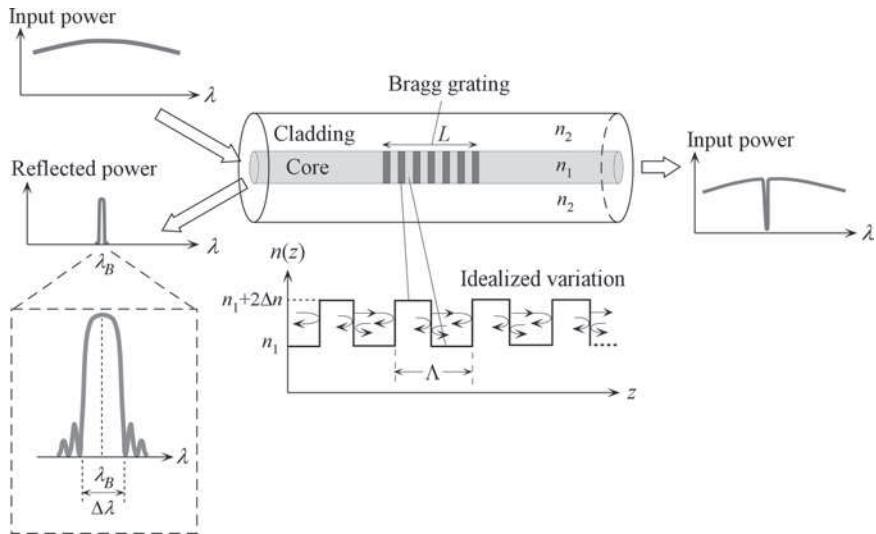


FIGURE 2.49 Fiber Bragg grating has a Bragg grating written in the core of a single-mode fiber over a certain length of the fiber. The Bragg grating reflects any light that has the Bragg wavelength λ_B , which depends on the refractive index and the periodicity. The transmitted spectrum has the Bragg wavelength missing.

grating pitch. (\bar{n} is quite close to n_1 , the core index.) The integer q , as we know from Chapter 1, is called the diffraction order. There are many Bragg wavelengths but usually we are only interested in the fundamental, corresponding to $q = 1$. We will use Eq. (2.15.1) with $q = 1$. Further, as explained below, in practice, the variation of $n(z)$ is almost sinusoidal rather than step-like changes, in which case q can only be 1.

If we were to couple light with a broad spectrum into the FBG, as shown in Figure 2.49, we will observe that there is a reflected light at λ_B with a certain bandwidth $\Delta\lambda$, both of which depend on the characteristics of the Bragg grating. The transmitted spectrum would have λ_B missing. One can also use a tunable laser whose wavelength can be varied at the input. We would observe reflected light at λ_B , which is missing from the transmission spectrum.

It should be immediately apparent from Eq. (2.15.1) that changes in the temperature and strain will modify Λ and \bar{n} , and hence shift the Bragg wavelength. Thus, FBGs can be used as temperature and strain sensors; we expand this application below.

The reflectance R of a FBG can be calculated by considering all the partial reflections in the grating, not only in the backward direction but also in the forward direction, and multiple interferences. Further, the idealized rectangular variation in Figure 2.49 should be replaced, at least as an approximation, by a sinusoidal of the type $\Delta n \sin(2\pi z/\Lambda)$. The final result for the maximum reflectance R at λ_B ²⁸

Reflectance
of a FBG
at λ_B

$$R = \tanh^2(\kappa L); \quad \kappa = \pi \left(\frac{\Delta n}{\lambda} \right) \quad (2.15.2)$$

²⁸The equation, as one might have surmised, is somewhat simplified. It also assumes that the propagating mode fully overlaps the Bragg grating region. Further, the n -variation is only approximately sinusoidal, and hence one must consider the higher harmonics as well to represent the index variation. The interested student can find an extensive treatise in Raman Kashyap, *Fiber Bragg Gratings*, 2nd Edition (Academic Press, Elsevier, 2010).

where L is the length of the grating as shown in Figure 2.49, and κ is a characteristic FBG parameter, called the **coupling coefficient** that depends only on the amplitude Δn of the index variation and the operating wavelength λ . (The EM waves inside the grating are coupled due to multiple reflections.) In Eq. (2.15.2), κL values of $1/2$, 1 , and 2 result in R values that are 21% , 58% , and 93% , respectively. The quantity κL serves to measure the strength of the grating. Gratings with $\kappa L < 1$ are **weak**, whereas those with $\kappa L > 1$ are **strong**. The bandwidth, or the spectral width, $\Delta\lambda$ of the Bragg reflected light is most easily defined between the zeros of the side lobes as indicated in Figure 2.49. $\Delta\lambda$ depends on whether the gratings is weak or strong, that is,

$$\Delta\lambda_{\text{weak}} = \frac{\lambda_B^2}{\bar{n}L} \quad \text{or} \quad \Delta\lambda_{\text{strong}} = \frac{4\kappa\lambda_B^2}{\pi\bar{n}} \quad (2.15.3)$$

Bandwidth of an FBG

While for weak gratings, the spectral width is inversely proportional to the grating length, for strong gratings, it is proportional to the coupling coefficient, and independent of L .

It can be seen that FBGs have the ability to filter the wavelength λ_B from the spectrum of input light and reflect light at λ_B . They are therefore used in telecommunications for adding or dropping a channel at a wavelength λ_B in WDM. Further, since they act as a wavelength selective mirror, they are used in *fiber lasers*. Inasmuch as the reflection of light results from dielectric changes that are distributed over a length of the fiber, the FBG is sometimes called a **distributed Bragg reflector**.

Various techniques have been used to fabricate FBGs. In one common technique the photosensitivity of the Ge-doped silica core fiber is used. The refractive index of Ge-doped silica increases upon exposure to UV light. An interference pattern (an interference fringe) is generated from a UV laser source. The core of the fiber is placed in the interference pattern. High light intensity regions go through a photoinduced refractive index change (n increases). Thus, a diffraction grating is generated in the core. Although the change Δn is normally small, $10^{-6} - 10^{-3}$, the length L can be long, yielding good reflectance at λ_B .

Some of the most important applications of FBGs is in sensors. First, consider a small increase in the temperature, δT . The fiber will expand and hence Λ will increase. The refractive index of silica will also increase. Thus there will be a shift (increase) in λ_B . This shift can be calibrated to measure the temperature. From Eq. (2.15.1), the change in λ_B is

$$\delta\lambda_B = 2\Lambda\delta\bar{n} + 2\bar{n}\delta\Lambda \quad (2.15.4)$$

and dividing by λ_B , and using the fact that the linear expansion coefficient $\alpha_L = (1/\Lambda)(\delta\Lambda/\delta T)$ and the temperature coefficient of the refractive index (TCRI) $\alpha_{\text{TCRI}} = (1/\bar{n})(\delta\bar{n}/\delta T)$, we have

$$\frac{\delta\lambda_B}{\lambda_B} = (\alpha_{\text{TCRI}} + \alpha_L)\delta T \quad (2.15.5)$$

FBG temperature sensor

For silica, $\alpha_L \approx 0.5 \times 10^{-6} \text{ K}^{-1}$, and $\alpha_{\text{TCRI}} \approx 7 \times 10^{-6} \text{ K}^{-1}$ at 1550 nm, so we get $(\delta\lambda_B/\lambda_B) \approx 7.5 \times 10^{-6} \text{ K}^{-1}$. At 1550 nm, this is a change of 0.011 nm or 11 pm for a 1°C rise. The shift of 0.011 nm is small and we would need a high-end optical spectrum analyzer to measure it. There are however other optical techniques (based on using Fabry–Perot cavities) to measure such small changes.

Suppose we stretch the fiber and induce a strain ε in it. By definition, $\delta\Lambda/\Lambda = \varepsilon$. The refractive index will also be affected because of the **photoelastic effect**, that is, the displacement

of atoms in the glass due to an induced strain ε causes a change δn in the refractive index. The photoelastic effect can be written in a simple manner as

Photoelastic effect

$$\delta n = -\frac{1}{2} n^3 p_e \varepsilon \quad (2.15.6)$$

where p_e is the photoelastic or elasto-optic coefficient of the medium.²⁹ (The negative sign implies that extending the fiber, positive ε , decreases n .) If we differentiate Eq. (2.15.1) with respect to Λ and then divide by λ_B , and use Eq. (2.15.6), we can easily show that

FBG strain sensor

$$\frac{\delta\lambda_B}{\lambda_B} = \left(1 - \frac{1}{2} n^2 p_e\right) \varepsilon \quad (2.15.7)$$

The photoelastic coefficient p_e of silica glass is approximately 0.22, which means that for a strain of 10^{-4} , and operation (λ_B) at 1550 nm, $\delta\lambda_B = 0.12$ nm. If the fiber diameter is 125 μm , this strain would represent a force (load) of roughly 0.1 N on the fiber. FBG strain sensors are already in use in the measurement of strain, load and pressure in monitoring the structural health of various structures, especially in earth quake prone regions, such as buildings, bridges, dams, tunnels, ships, aircrafts, deep sea oil rigs, pipelines, large power generators, wind turbines, etc.

FBG sensors are immune to EM interference and changes in the light intensity. They are light, flexible, and easily embedded in structures. One distinct advantage is that they can be multiplexed. By having Bragg gratings with different λ_B (different Λ) on different locations on a fiber one can have a number of sensors distributed along the fiber as shown in Figure 2.50. One can therefore construct a sensor array to measure the temperature or strain at different locations on the fiber. There would be a series of Bragg wavelengths, $\lambda_{B1}, \lambda_{B2}, \dots$, in the reflected light, each associated with a different sensor on the fiber. The reflected light is fed into an optical spectrum analyzer or a polychromator to monitor the changes in the Bragg wavelengths. One can therefore easily monitor the shifts in λ_B and hence measure the temperature or strain at various locations almost simultaneously.

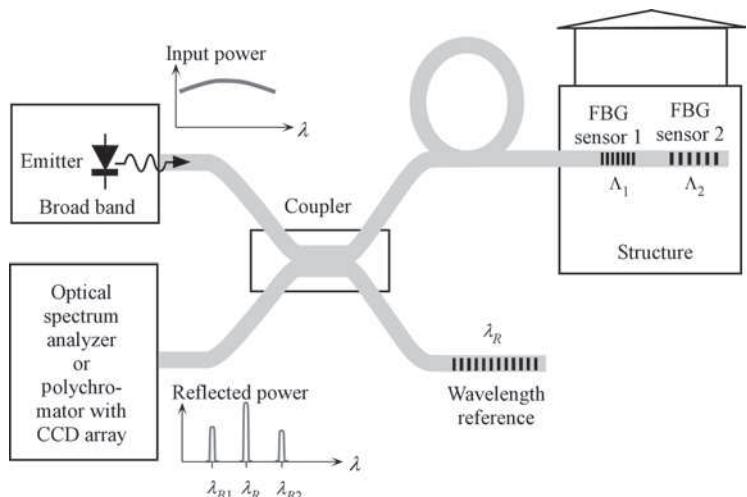


FIGURE 2.50 A highly simplified schematic diagram of a multiplexed Bragg grating-based sensing system. The FBG sensors are distributed on a single fiber that is embedded in the structure in which strains are to be monitored. The coupler allows optical power to be coupled from one fiber to the other.

²⁹This is actually an effective photoelastic coefficient equivalent to $p_{12} - \nu(p_{11} + p_{12})$, where n is the Poisson ratio, and p_{11} and p_{12} are the true coefficients of the photoelastic coefficient tensor that relates n and strain ε .

A polychromator is an instrument in which a diffraction grating separates out different wavelengths (different λ_{Bs}) of light, which are then made incident on a CCD (charge coupled device) optical image sensor array (a pixelated photodetector). As the wavelength λ_{B1} shifts, the corresponding light beam becomes displaced, and moves along the pixels of the CCD sensor, which are read out from the CCD into a computer for data analysis and interpretation (done by software). Note that there is usually another FBG that generates a reference wavelength λ_R as shown in Figure 2.50.

EXAMPLE 2.15.1 Fiber Bragg grating at 1550 nm

A silica fiber-based FBG is required to operate at 1550 nm. What should be the periodicity of the grating Λ ? If the amplitude of index variation Δn is 10^{-4} and total length of the FBG is 10 mm, what is the reflectance at the Bragg wavelength and the spectral width of the reflected light? Assume that the effective refractive index \bar{n} is 1.450.

Solution

We can use $\lambda_B = 2\bar{n}\Lambda$, that is, $1550 \text{ nm} = 2(1.450)(\Lambda)$ so that the grating periodicity is

$$\Lambda = 534.5 \text{ nm}$$

The coupling coefficient κ is given in Eq. (2.15.2)

$$\kappa = \pi\Delta n/\lambda = \pi(10^{-4})/(1.550 \times 10^{-9} \text{ m}) = 202.7 \text{ m}^{-1}$$

Thus, $\kappa L = 2.027$, so the FBG is a strong grating. The reflectance is

$$R = \tanh^2(\kappa L) = \tanh^2(2.027) = 0.933 \quad \text{or} \quad 93.3\%$$

The spectral width is given by

$$\Delta\lambda_{\text{strong}} = \frac{4\kappa\lambda_B^2}{\pi\bar{n}} = \frac{4(202.7)(1.55 \times 10^{-6})^2}{\pi(1.450)} = 0.428 \text{ nm}$$

The reflectance and spectral width values are approximate inasmuch as Eqs. (2.15.2) and (2.15.3) contain a number of assumptions.

Questions and Problems

2.1 Symmetric dielectric slab waveguide

Consider two rays such as 1 and 2' interfering at point P in Figure 2.4. Both are moving with the same incidence angle but have different κ_m wave vectors just before point P . In addition, there is a phase difference between the two due to the different paths taken to reach point P . We can represent the two waves as $E_1(y, z, t) = E_0 \cos(\omega t - \kappa_m y - \beta_m z + \delta)$ and $E_2(y, z, t) = E_0 \cos(\omega t + \kappa_m y - \beta_m z)$ where the $\kappa_m y$ terms have opposite signs indicating that the waves are traveling in opposite directions. δ has been used to indicate that the waves have a phase difference and travel different optical paths to reach point P . We also know that $\kappa_m = k_1 \cos \theta_m$ and $\beta_m = k_1 \sin \theta_m$, and obviously have the waveguide condition already incorporated into them through θ_m . Show that the superposition of E_1 and E_2 at P is given by

$$E(y, z, t) = 2E_0 \cos\left(\kappa_m y - \frac{1}{2}\delta\right) \cos\left(\omega t - \beta_m z + \frac{1}{2}\delta\right)$$

What do the two cosine terms represent?

The planar waveguide is symmetric, which means that the intensity, E^2 , must be either maximum (even m) or minimum (odd m) at the center of the guide. Choose suitable δ values and plot the relative magnitude of

the electric field across the guide for $m = 0, 1$, and 2 for the following symmetric dielectric planar guide: $n_1 = 1.4550$, $n_2 = 1.4400$, $a = 10\text{ }\mu\text{m}$, $\lambda = 1.5\text{ }\mu\text{m}$ (free space), the first three modes have $\theta_1 = 88.84^\circ$, $\theta_2 = 87.67^\circ$, $\theta_3 = 86.51^\circ$. Scale the field values so that the maximum field is unity for $m = 0$ at the center of the guide. [Alternatively, you can choose δ so that intensity (E^2) is the same at the boundaries at $y = a$ and $y = -a$; it would give the same distribution.]

- 2.2 Standing waves inside the core of a symmetric slab waveguide** Consider a symmetric planar dielectric waveguide. Upward and downward traveling waves inside the core of the planar waveguide set up a standing wave along y . The standing wave can only exist if the wave can be replicated after it has traveled along the y -direction over one round trip. Put differently, a wave starting at A in Figure 2.51 and traveling toward the upper face will travel along y , be reflected at B , travel down, become reflected again at A , and then it would be traveling in the same direction as it started. At this point, it must have an identical phase to its starting phase so that it can replicate and not destroy itself. Given that the wave vector along y is κ_m , derive the waveguide condition.

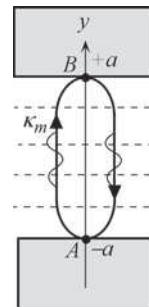


FIGURE 2.51 Upward and downward traveling waves along y set up a standing wave. The condition for setting up a standing wave is that the wave must be identical, and able to replicate itself, after one round trip along y .

2.3 Dielectric slab waveguide

- (a) Consider the two parallel rays 1 and 2 in Figure 2.52. Show that when they meet at C at a distance y above the guide center, the phase difference is

$$\Phi_m = k_1 2(a - y) \cos \theta_m - \phi_m$$

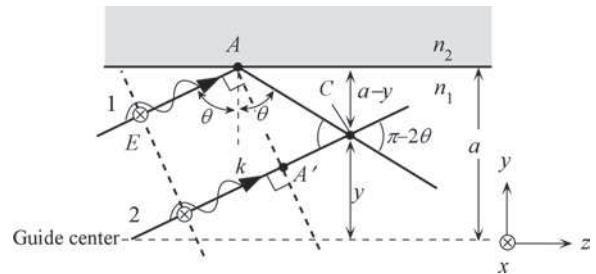


FIGURE 2.52 Rays 1 and 2 are initially in phase as they belong to the same wavefront. Ray 1 experiences total internal reflection at A . 1 and 2 interfere at C . There is a phase difference between the two waves. (For simplicity θ_m is shown as θ .)

- (b) Using the waveguide condition, show that

$$\Phi_m = \Phi_m(y) = m\pi - \frac{y}{a}(m\pi + \phi_m)$$

- (c) The two waves interfering at C can be most simply and conveniently represented as

$$E(y) = A \cos(\omega t) + A \cos[\omega t + \Phi_m(y)]$$

Hence find the amplitude of the field variation along y , across the guide. What is your conclusion?

- 2.4 Slab waveguide** Consider an asymmetric planar waveguide with $n_1 = 2.35$ and $n_2 = 2.25$. The width of the waveguide is $2.5\text{ }\mu\text{m}$, operating at 1300 nm . Calculate the number of TE and TM modes guided by this waveguide. Find out the range of wavelengths for which this waveguide will remain single-mode.

- 2.5 TE and TM Modes in dielectric slab waveguide** Consider a planar dielectric guide with a core thickness of 15 μm , $n_1 = 1.45$, $n_2 = 1.43$, and light wavelength of 1.30 μm . Given the waveguide condition, and the expressions for phase changes ϕ and ϕ' in TIR for the TE and TM modes, respectively,

$$\tan\left(\frac{1}{2}\phi_m\right) = \frac{\left[\sin^2\theta_m - \left(\frac{n_2}{n_1}\right)^2\right]^{1/2}}{\cos\theta_m} \quad \text{and} \quad \tan\left(\frac{1}{2}\phi'_m\right) = \frac{\left[\sin^2\theta_m - \left(\frac{n_2}{n_1}\right)^2\right]^{1/2}}{\left(\frac{n_2}{n_1}\right)^2 \cos\theta_m}$$

using a graphical solution find the angle θ for the fundamental TE and TM modes and compare their propagation constants along the guide.

- 2.6 Group velocity** We can calculate the group velocity of a given mode as a function of frequency using a convenient math software package. It is assumed that the math software package can carry out symbolic algebra such as partial differentiation (the author used LiveMath by MathMonkeys, though others can also be used). The propagation constant of a given mode is $\beta = k_1 \sin\theta$, where β and θ imply β_m and θ_m . The objective is to express β and ω in terms of θ . Since $k_1 = n_1\omega/c$, the waveguide condition is

$$\tan\left(a\frac{\beta}{\sin\theta} \cos\theta - m\frac{\pi}{2}\right) = \frac{\left[\sin^2\theta - (n_2/n_1)^2\right]^{1/2}}{\cos\theta}$$

so that

$$\beta \approx \frac{\tan\theta}{a} \left[\arctan\left(\sec\theta\sqrt{\sin^2\theta - (n_2/n_1)^2}\right) + m(\pi/2) \right] = F_m(\theta) \quad (\text{P2.1})$$

where $F_m(\theta)$ is a function of θ at a given m . The frequency ω is given by

$$\omega = \frac{c\beta}{n_1 \sin\theta} = \frac{c}{n_1 \sin\theta} F_m(\theta) \quad (\text{P2.2})$$

Both β and ω are now a function of θ in Eqs. (P2.1) and (P2.2). Then the group velocity is found by differentiating Eqs. (P2.1) and (P2.2) with respect to θ , that is,

$$v_g = \frac{d\omega}{d\beta} = \left[\frac{d\omega}{d\theta} \right] \times \left[\frac{d\theta}{d\beta} \right] = \frac{c}{n_1} \left[\frac{F'_m(\theta)}{\sin\theta} - \frac{\cos\theta}{\sin^2\theta} F_m(\theta) \right] \times \left[\frac{1}{F'_m(\theta)} \right]$$

that is,

$$v_g = \frac{c}{n_1 \sin\theta} \left[1 - \cot\theta \frac{F_m(\theta)}{F'_m(\theta)} \right] \quad (\text{P2.3})$$

Group
velocity,
planar
waveguide

where $F'_m = dF_m/d\theta$ is found by differentiating the second term of Eq. (P2.1). For a given m value, Eqs. (P2.2) and (P2.3) can be plotted parametrically, that is, for each θ value we can calculate ω and v_g and plot v_g vs. ω . Figure 2.11 shows an example for a guide with the characteristics as given in the figure caption. Using a convenient math software package, or by other means, obtain a ω vs. β_m graph for this waveguide and obtain v_g . If ω vs. β_m is a straight line, point out its significance. Plot v_g vs. ω . Discuss intermodal dispersion.

- 2.7 Dielectric slab waveguide** Consider a dielectric slab waveguide that has a thin GaAs layer of thickness 0.25 μm between two AlGaAs layers. The refractive index of GaAs is 3.6 and that of the AlGaAs layers is 3.4. What is the cutoff wavelength beyond which only a single mode can propagate in the waveguide, assuming that the refractive index does not vary greatly with the wavelength? If a radiation of wavelength 860 nm (corresponding to bandgap radiation) is propagating in the GaAs layer, what is the penetration of the evanescent wave into the AlGaAs layers? What is the mode field width (MFW) of this radiation? Point out the effect of change of radiation wavelength (λ) on the MFW

- 2.8 Dielectric slab waveguide** Consider a slab dielectric waveguide that has a core thickness ($2a$) of 20 μm , $n_1 = 3.00$, $n_2 = 1.50$. Solution of the waveguide condition in Eq. (2.1.9) gives the mode angles θ_0 and θ_1 for the TE₀ and TE₁ modes for selected wavelengths as summarized in Table 2.7. For each wavelength calculate ω and β_m and then plot ω vs. β_m . On the same plot show the lines with slopes c/n_1 and c/n_2 . Compare your plot with the dispersion diagram in Figure 2.10.

TABLE 2.7 The solution of the waveguide condition for $a = 10 \mu\text{m}$, $n_1 = 3.00$, $n_2 = 1.50$ gives the incidence angles θ_0 and θ_1 for modes 0 and 1 at the wavelengths shown

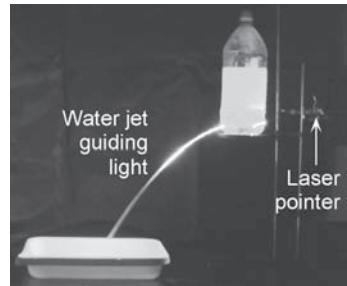
$\lambda (\mu\text{m})$	15	20	25	30	40	45	50	70	100	150	200
θ_0°	77.8	74.52	71.5	68.7	63.9	61.7	59.74	53.2	46.4	39.9	36.45
θ_1°	65.2	58.15	51.6	45.5	35.5	32.02	30.17	—	—	—	—

2.9 Dielectric slab waveguide Consider a planar dielectric waveguide with a core thickness of $10 \mu\text{m}$, $n_1 = 1.4446$, $n_2 = 1.4440$. Calculate the V-number, the mode angle θ_m for $m = 0$ (use a graphical solution, if necessary), penetration depth, and mode field width, MFW = $2a + 2\delta$, for light wavelengths of $1.0 \mu\text{m}$ and $1.5 \mu\text{m}$. What is your conclusion? Compare your MFW calculation with $2w_o = 2a(V + 1)/V$. The mode angle θ_0 is given as $\theta_0 = 88.85^\circ$ for $\lambda = 1 \mu\text{m}$ and $\theta_0 = 88.72^\circ$ for $\lambda = 1.5 \mu\text{m}$ for the fundamental mode $m = 0$.

2.10 A multimode fiber Consider a multimode fiber with a core diameter of $60 \mu\text{m}$, core refractive index of 1.47, and a cladding refractive index of 1.45, both at 870 nm . Consider operating this fiber at $\lambda = 870 \text{ nm}$. (a) Calculate the numerical aperture. (b) Find out the normalized core-cladding index difference. (c) Calculate the V-number for the fiber and estimate the number of guided modes. (d) Calculate the wavelength beyond which the fiber becomes single-mode. (e) Calculate the modal dispersion $\Delta\tau$ and hence estimate the bit rate \times distance product.

2.11 A water jet guiding light One of the early demonstrations of the way in which light can be guided along a higher refractive index medium by total internal reflection involved illuminating the starting point of a water jet as it comes out from a water tank. The refractive index of water is 1.330. Consider a water jet of diameter 3 mm that is illuminated by green light of wavelength 560 nm. What is the V-number, numerical aperture, total acceptance angle of the jet? How many modes are there? What is the cutoff wavelength? The diameter of the jet increases (slowly) as the jet flows away from the original spout. However, the light is still guided. Why?

Light guided along a thin water jet. A small hole is made in a plastic soda drink bottle full of water to generate a thin water jet. When the hole is illuminated with a laser beam (from a green laser pointer), the light is guided by total internal reflections along the jet to the tray. Water with air bubbles (produced by shaking the bottle) was used to increase the visibility of light. Air bubbles scatter light and make the guided light visible. First such demonstration has been attributed to Jean-Daniel Colladon, a Swiss scientist, who demonstrated a water jet guiding light in 1841.



2.12 Single-mode fiber Consider a fiber with an $\text{SiO}_2\text{-}13.5\%\text{GeO}_2$ (*i.e.* 86.5% SiO_2 -13.5% GeO_2 %) core of diameter of $6 \mu\text{m}$ and refractive index of 1.47 and a cladding refractive index of 1.46, both refractive indices at 1300 nm , where the fiber is to be operated using a laser source with a half maximum width of 2 nm . (a) Calculate the V-number for this fiber. (b) What is the maximum allowed diameter of the core that maintains operations in single-mode? (c) Calculate the wavelength below which the fiber becomes multimode. (d) Calculate the numerical aperture. (e) Calculate the maximum acceptance angle. (f) Obtain the material dispersion and waveguide dispersion and hence estimate the bit rate \times distance product ($B \times L$) of the fiber.

2.13 Single-mode fiber Consider a step-index fiber with a core of diameter of $8 \mu\text{m}$ and refractive index of 1.45 at $1.55 \mu\text{m}$ and a normalized refractive index difference of 0.25% where the fiber is to be operated using a laser source with a half-maximum width of 3 nm . At $1.55 \mu\text{m}$, the material and waveguide dispersion coefficients of this fiber are approximately given by $D_m = 12 \text{ ps nm}^{-1} \text{ km}^{-1}$ and $D_w = -6 \text{ ps nm}^{-1} \text{ km}^{-1}$ (a) Calculate the V-number for the fiber. Is this a single-mode fiber? (b) Calculate the wavelength below which the fiber becomes multimode. (c) Calculate the numerical aperture. (d) Calculate the maximum total acceptance angle. (e) Calculate the material, waveguide, and chromatic dispersion per kilometer of fiber. (f) Estimate the bit rate \times distance product ($B \times L$) of this fiber. (g) What is the mode field diameter? (h) Find out optical and electrical bandwidth for a 10 km link length of this fiber.

- 2.14 Normalized propagation constant b** Consider a weakly guiding step-index fiber in which $(n_1 - n_2)/n_1$ is very small. Show that

$$(a) \quad b = \frac{(\beta/k)^2 - n_2^2}{n_1^2 - n_2^2} \approx \frac{(\beta/k) - n_2}{n_1 - n_2}$$

Note: Since Δ is very small, $n_2/n_1 \approx 1$ can be assumed were convenient. The first equation can be rearranged as

$$\beta/k = [n_2^2 + b(n_1^2 - n_2^2)]^{1/2} = n_2^2(1 + x)^{1/2}; \quad x = b(n_1^2 - n_2^2)/n_2^2$$

where x is small. Taylor's expansion in x to the first linear term would then provide a linear relationship between β and b .

- (b) Plot a graph between V value vs. b of the step-index fiber for fundamental and first higher order modes.

- 2.15 Group velocity of the fundamental mode** Reconsider Example 2.3.4, which has a single-mode fiber with core and cladding indices of 1.4480 and 1.4400, core radius of 3 μm , operating at 1.5 μm . Use the equation

$$b \approx \frac{(\beta/k) - n_2}{n_1 - n_2}; \quad \beta = n_2 k [1 + b\Delta]$$

to recalculate the propagation constant β . Change the operating wavelength to 1.3 μm , and then recalculate the new propagation constant β . Then determine the group velocity v_g of the fundamental mode at 1.55 μm , and the group delay τ_g over 2 km of fiber.

- 2.16 A single-mode fiber design** The Sellmeier dispersion equation provides n vs. λ for pure SiO_2 and $\text{SiO}_2\text{-}13.5\text{ mol.\%GeO}_2$ in Table 1.2 in Chapter 1. The refractive index increases linearly with the addition of GeO_2 to SiO_2 from 0 to 13.5 mol.%. A single-mode step-index fiber is required to have the following properties: NA = 0.15, core diameter of 8 μm , and a cladding of pure silica, and operate at 1.3 μm and 1.55 μm . What should the core composition be at these wavelengths?

- 2.17 Material dispersion** If N_g is the group refractive index of the core material of a step fiber, then the propagation time (group delay time) of the fundamental mode is

$$\tau = L/v_g = LN_{g1}/c.$$

Since N_g will depend on the wavelength, show that the material dispersion coefficient D_m is given approximately by

$$D_m = \frac{d\tau}{Ld\lambda} \approx \frac{\lambda}{c} \frac{d^2n}{d\lambda^2}$$

Using the Sellmeier equation and the constants in Table 1.2 in Chapter 1, evaluate the material dispersion at $\lambda = 1.3$ and 1.55 μm for pure silica (SiO_2) and $\text{SiO}_2\text{-}13.5\text{ mol.\%GeO}_2$ glass.

- 2.18 Waveguide dispersion** Waveguide dispersion arises as a result of the dependence of the propagation constant on the V -number, which depends on the wavelength. It is present even when the refractive index is constant—no material dispersion. Let us suppose that n_1 and n_2 are wavelength (or k) independent. Suppose that β is the propagation constant of mode lm and $k = 2\pi/\lambda$ in which λ is the free-space wavelength. Then the normalized propagation constant b and propagation constant β are related by

$$\beta = n_2 k (1 + b\Delta) \quad (\text{P2.4})$$

The group velocity is defined and given by

$$v_g = \frac{d\omega}{d\beta} = c \frac{dk}{d\beta}$$

Show that the propagation time, or the group delay time, τ of the mode is

$$\tau = \frac{L}{v_g} = \frac{Ln_2}{c} + \frac{Ln_2\Delta}{c} \frac{dk}{d\lambda} \quad (\text{P2.5})$$

Given the definition of V

$$V = ka[n_1^2 - n_2^2]^{1/2} \approx kan_2(2\Delta)^{1/2} \quad (\text{P2.6})$$

and

$$\frac{d(Vb)}{dV} = \frac{d}{dV} [bkan_2(2\Delta)^{1/2}] = an_2(2\Delta)^{1/2} \frac{d}{dV}(bk) \quad (\text{P2.7})$$

Show that

$$\frac{d\tau}{d\lambda} = -\frac{Ln_2\Delta}{c\lambda} V \frac{d^2(Vb)}{dV^2} \quad (\text{P2.8})$$

and that the waveguide dispersion coefficient is

$$D_w = \frac{d\tau}{Ld\lambda} = -\frac{n_2\Delta}{c\lambda} V \frac{d^2(Vb)}{dV^2} \quad (\text{P2.9})$$

Figure 2.53 shows the dependence of $V[d^2(Vb)/dV^2]$ on the V -number. In the range $1.5 < V < 2.4$,

$$V \frac{d^2(Vb)}{dV^2} \approx \frac{1.984}{V^2}$$

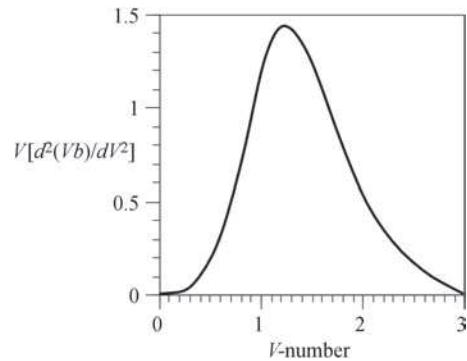


FIGURE 2.53 $d^2(Vb)/dV^2$ vs. V -number for a step-index fiber. (Source: Data extracted from W. A. Gambling *et al.*, *The Radio and Electronics Engineer*, 51, 313, 1981.)

Show that,

$$D_w \approx -\frac{n_2\Delta}{c\lambda} \frac{1.984}{V^2} = -\frac{(n_1 - n_2)}{c\lambda} \frac{1.984}{V^2} \quad (\text{P2.10})$$

which simplifies to

$$D_w \approx -\frac{1.984}{c(2\pi a)^2 2n_2} \lambda \quad (\text{P2.11})$$

that is,

$$D_w(\text{ps nm}^{-1} \text{ km}^{-1}) \approx -\frac{83.76 \lambda(\mu\text{m})}{[a(\mu\text{m})]^2 n_2} \quad (\text{P2.12})$$

Consider a fiber with a core of diameter of $8 \mu\text{m}$ and refractive index 1.468 and a cladding refractive index of 1.464, both refractive indices at 1300 nm. Suppose that a $1.3\text{-}\mu\text{m}$ laser diode with a spectral linewidth of 2 nm is used to provide the input light pulses. Estimate the waveguide dispersion per kilometer of fiber using Eqs. (P2.9) and (P2.11).

- 2.19 Profile dispersion** Total dispersion in a single-mode, step-index fiber is primarily due to material dispersion and waveguide dispersion. However, there is an additional dispersion mechanism called *profile dispersion* that arises from the propagation constant β of the fundamental mode also depending on the refractive index difference Δ . Consider a light source with a range of wavelengths $\delta\lambda$ coupled into a step-index fiber. We can view this as a change $\delta\lambda$ in the input wavelength λ . Suppose that n_1, n_2 , hence Δ depends on the wavelength λ . The propagation time, or the group delay time, τ_g per unit length is

$$\tau_g = 1/v_g = d\beta/d\omega = (1/c)(d\beta/dk) \quad (\text{P2.13})$$

where k is the free-space propagation constant ($2\pi/\lambda$), and we used $d\omega = cdk$. Since β depends on n_1 , Δ , and V , consider τ_g as a function of n_1 , Δ (thus n_2), and V . A change $\delta\lambda$ in λ will change each of these quantities. Using the partial differential chain rule

$$\frac{\partial \tau_g}{\partial \lambda} = \frac{\partial \tau_g}{\partial n_1} \frac{\partial n_1}{\partial \lambda} + \frac{\partial \tau_g}{\partial V} \frac{\partial V}{\partial \lambda} + \frac{\partial \tau_g}{\partial \Delta} \frac{\partial \Delta}{\partial \lambda} \quad (\text{P2.14})$$

The mathematics turns out to be complicated, but the statement in Eq. (P2.14) is equivalent to

$$\begin{aligned} \text{Total dispersion} &= \text{Material dispersion (due to } \frac{\partial n_1}{\partial \lambda} \text{)} \\ &\quad + \text{Waveguide dispersion (due to } \frac{\partial V}{\partial \lambda} \text{)} + \text{Profile dispersion (due to } \frac{\partial \Delta}{\partial \lambda} \text{)} \end{aligned}$$

in which the last term is due to Δ depending on λ ; although small, this is not zero. Even the statement in Eq. (P2.14) above is oversimplified but nonetheless provides an insight into the problem. The total intramode (chromatic) dispersion coefficient D_{ch} is then given by

$$D_{ch} = D_m + D_w + D_p \quad (\text{P2.15})$$

in which D_m , D_w , and D_p are material, waveguide, and profile dispersion coefficients, respectively. The waveguide dispersion is given by Eqs. (P2.11) and (P2.12), and the *profile dispersion coefficient* is (very) approximately,³⁰

$$D_p \approx -\frac{N_{g1}}{c} \left(V \frac{d^2(Vb)}{dV^2} \right) \left(\frac{d\Delta}{d\lambda} \right) \quad (\text{P2.16})$$

Profile dispersion coefficient

in which b is the normalized propagation constant and $Vd^2(Vb)/dV^2 \approx 1.984/V^2$.

Consider a fiber with a core of diameter of 8 μm . The refractive and group indices of the core and cladding at $\lambda = 1.55 \mu\text{m}$ are $n_1 = 1.4500$, $n_2 = 1.4444$, $N_{g1} = 1.4680$, $N_{g2} = 1.4628$, and $d\Delta/d\lambda = 232 \text{ m}^{-1}$. Estimate the waveguide and profile dispersion per km of fiber per nm of input light linewidth at this wavelength. (Note: The values given are approximate and for a fiber with silica cladding and 3.6% germania-doped core.)

- 2.20 Dispersion at zero dispersion coefficient** Since the spread in the group delay $\Delta\tau$ along a fiber depends on the $\Delta\lambda$, the linewidth of the source, we can expand $\Delta\tau$ as a Taylor series in $\Delta\lambda$. Consider the expansion at $\lambda = \lambda_0$ where $D_{ch} = 0$. The first term with $\Delta\lambda$ would have $d\Delta\tau/d\lambda$ as a coefficient, that is D_{ch} , and at λ_0 this will be zero; but not the second term with $(\Delta\lambda)^2$ that has a differential, $d^2\Delta\tau/d\lambda^2$ or $dD_{ch}/d\lambda$. Thus, the dispersion at λ_0 would be controlled by the slope S_0 of D_{ch} vs. λ curve at λ_0 . Show that the chromatic dispersion at λ_0 is

$$\Delta\tau = \frac{L}{2} S_0(\Delta\lambda)^2$$

Chromatic dispersion at λ_0

A single-mode fiber has a zero dispersion at $\lambda_0 = 1310 \text{ nm}$, dispersion slope $S_0 = 0.090 \text{ ps nm}^{-2} \text{ km}$. What is the dispersion for a laser with $\Delta\lambda = 1.5 \text{ nm}$? What would control the dispersion in practice?

- 2.21 Polarization mode dispersion** A fiber manufacturer specifies a maximum value of $0.025 \text{ ps km}^{-1/2}$ for the polarization mode dispersion (PMD) in its single-mode fiber. What would be the dispersion, maximum bit rate, and the optical bandwidth for this fiber over an optical link that is 400 km long if the only dispersion mechanism was PMD?

- 2.22 Polarization mode dispersion** Consider a particular single-mode fiber (ITU-T G.652 compliant) that has a chromatic dispersion of $18 \text{ ps nm}^{-1} \text{ km}^{-1}$. The chromatic dispersion is zero at 1300 nm, and the dispersion slope is $0.092 \text{ ps nm}^{-2} \text{ km}^{-1}$. The PMD coefficient is $0.05 \text{ ps km}^{-1/2}$. Calculate the total dispersion over 180 km if the fiber is operated at 1300 nm and the source is a laser diode with a linewidth $\Delta\lambda = 1 \text{ nm}$. What should be the linewidth of the laser source so that over 180 km the chromatic dispersion is the same as PMD?

- 2.23 Dispersion compensation** Calculate the total dispersion and the overall net dispersion coefficient when a 1000-km transmission fiber with chromatic dispersion $D_{ch} = +20 \text{ ps nm}^{-1} \text{ km}^{-1}$ is spliced to a compensating fiber that is 100 km long and has $D_{ch} = -100 \text{ ps nm}^{-1} \text{ km}^{-1}$. Assume that the input light spectral width is 1 nm.

- 2.24 Cladding diameter** A comparison of two step-index fibers, one SMF and the other MMF, shows that the SMF has a core diameter of 9 μm but a cladding diameter of 125 μm , while the MMF has a core diameter of 100 μm but a cladding diameter of 125 μm . Discuss why the manufacturer has chosen those values.

³⁰J. Gowar, *Optical Communications Systems*, 2nd Edition (Prentice Hall, Pearson Education, 1993), Ch. 8 has the derivation of this equation.

- 2.25 Graded index fiber** Consider an optimal graded index fiber with a core diameter of 20 μm and a refractive index of 1.475 at the center of the core and a cladding refractive index of 1.453. What are its NA at the fiber axis, and its effective NA? Calculate the V value and the number of modes at 1300 nm operation for this fiber. Suppose that the fiber is coupled to a laser diode emitter at 1300 nm and a spectral linewidth (FWHM) of 2 nm. The material dispersion coefficient at this wavelength is about $-5 \text{ ps nm}^{-1} \text{ km}^{-1}$. Calculate the total dispersion and estimate the bit rate \times distance product of the fiber. How does this compare with the performance of a multimode fiber with same core radius, and n_1 and n_2 ? What would the total dispersion and maximum bit rate be if an LED source of spectral width (FWHM) $\Delta\tau_{1/2} \approx 60 \text{ nm}$ is used?
- 2.26 Graded index fiber** Consider a graded index fiber with a core diameter of 62.5 μm and a refractive index of 1.47 at the center of the core and a cladding refractive index of 1.45 at a wavelength 1550 nm. Suppose that we use a laser diode emitter with a spectral FWHM linewidth of 2 nm to transmit along this fiber at a wavelength of 1550 nm. Calculate the total dispersion and estimate the bit rate \times distance product of the fiber. The material dispersion coefficient D_m at 1550 nm is $-20 \text{ ps nm}^{-1} \text{ km}^{-1}$. How does this compare with the performance of a multimode fiber with the same core radius, and n_1 and n_2 ?
- 2.27 Graded index fiber** A standard graded index fiber from a particular fiber manufacturer has a core diameter of 62.5 μm , cladding diameter of 125 μm , a NA of 0.300. The core refractive index n_1 is 1.455. The manufacturer quotes minimum optical bandwidth \times distance values of 400 MHz \cdot km at 1300 nm and 800 MHz \cdot km at 1550 nm. Assume that a laser is to be used with this fiber and the laser linewidth $\Delta\lambda = 2 \text{ nm}$. What are the corresponding dispersion values? What type of dispersion do you think dominates? Is the graded index fiber assumed to have the ideal optimum index profile? (State your assumptions.) What is the optical link distance for operation at 2 Gbs $^{-1}$ at 850 nm and 1550 nm?
- 2.28 Graded index fiber and optimum dispersion** The graded index fiber theory and equations tend to be quite complicated. If γ is the profile index then the rms *intermodal dispersion* is given by³¹

$$\sigma = \frac{Ln_1\Delta}{2c} \left(\frac{\gamma}{\gamma + 1} \right) \left(\frac{\gamma + 2}{3\gamma + 2} \right)^{1/2} \times \left[c_1^2 + \frac{4c_1c_2(\gamma + 1)\Delta}{2\gamma + 1} + \frac{16c_2^2\Delta^2(\gamma + 1)^2}{(5\gamma + 2)(3\gamma + 2)} \right]^{1/2} \quad (\text{P2.17})$$

where c_1 and c_2 are given by

$$c_1 = \frac{\gamma - 2 - \delta}{\gamma + 2}; \quad c_2 = \frac{3\gamma - 2 - 2\delta}{2(\gamma + 2)}; \quad \delta = -2 \left(\frac{n_1}{N_{g1}} \right) \left(\frac{\lambda}{\Delta} \right) \left(\frac{d\Delta}{d\lambda} \right) \quad (\text{P2.18})$$

where δ is a small unitless parameter that represents the change in Δ with λ . The optimum profile coefficient γ_o is

$$\gamma_o = 2 + \delta - \Delta \frac{(4 + \delta)(3 + \delta)}{(5 + 2\delta)} \quad (\text{P2.19})$$

Consider a graded index fiber for use at 850 nm, with $n_1 = 1.475$, $N_{g1} = 1.489$, $\Delta = 0.015$, $d\Delta/d\lambda = -683 \text{ m}^{-1}$. Plot σ in ps km $^{-1}$ vs. γ from $\gamma = 1.8$ to 2.4 and find the minimum. (Consider plotting σ on a logarithmic axis.) Compare the minimum σ and the optimum γ , with the relevant expressions in Section 2.8. Find the percentage change in γ for a 10 \times increase in σ . What is your conclusion?

- 2.29 GRIN rod lenses** Figure 2.32 shows graded index rod lenses. (a) How would you represent Figure 2.32 (a) using two conventional converging lenses. What are O and O' ? (b) How would you represent Figure 2.32 (b) using a conventional converging lens. What is O' ? (c) Sketch ray paths for a GRIN rod with a pitch between 0.25 P and 0.5 P starting from O at the face center. Where is O' ? (d) What use is 0.23 P GRIN rod lens in Figure 2.32 (c)?
- 2.30 Photonic Crystal Fibers (PCFs)** Explain the lightwave guidance mechanism in solid and hollow core PCFs. Why are PCFs called endlessly single-mode fibers?
- 2.31 Attenuation** Average power launched into a 10 km length of optical fiber is 100 μW and the average optical power at the fiber output is 2 μW . (a) Find the signal attenuation in dB assuming there are no splices. (b) Find the signal attenuation per km for this fiber. (c) Find the overall signal attenuation for a 10 km optical link using the same fiber with splices at 1 km intervals, with splice loss of 1 dB.

³¹R. Olshansky and D. Keck, *Appl. Opt.*, 15, 483, 1976.

- 2.32 Cut-back method of attenuation measurement** Cut-back method is a destructive measurement technique for determining the attenuation α of a fiber. The first part of the experiment involves measuring the optical power P_{far} coming out from the fiber at the far end as shown in Figure 2.54. Then, in the second part, keeping everything the same, the fiber is cut close to the launch or the source end. The output power P_{near} is measured at the near end from the shortcut fiber. The attenuation is then given by

$$\alpha = (-10/L) \log (P_{\text{far}}/P_{\text{near}})$$

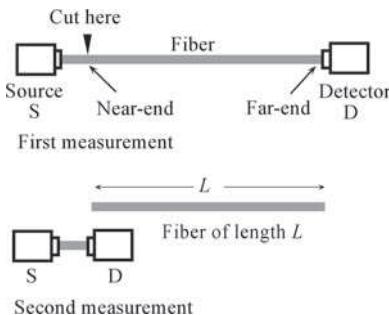


FIGURE 2.54 Illustration of the cut-back method for measuring the fiber attenuation. S is an optical source and D is a photodetector.

in which L is the separation of the measurement points, the length of the cut fiber, and α is in dB per unit length. The output P_{near} from the shortcut fiber in the second measurement is actually the input into the fiber under test in the first experiment. Usually a mode scrambler (mode stripper) is used for multimode fibers before the input. The power output from a particular fiber is measured to be 13 nW. Then, 10 km of fiber is cut out and the power output is measured again and found to be 43 nW. What is the attenuation of the fiber?

2.33 Intrinsic losses

- (a) Consider a silica-based standard single-mode fiber with a NA of 0.15. What is its attenuation at 1550 nm? How does this compare with the attenuation quotes for the Corning SMF@28e⁺, 0.16–0.19 dB km⁻¹ at 1550 nm?
- (b) Consider a step index fiber $n_1 = 1.47$ and $n_2 = 1.46$, core diameter = 50 μm . Corresponding to the maximum value of the incident angle for which light rays will be guided through this fiber, calculate the number of reflections and the loss that would take place in one kilometre length of the fiber assuming 0.01% loss of power at each reflection.

2.34 Scattering losses and fictive temperature Rayleigh scattering process decreases with wavelength, and as mentioned in Chapter 1, it is inversely proportional to λ^4 . The expression for the attenuation α_R in a single component glass such as silica due to Rayleigh scattering is approximately given by two sets of different equations in the literature³²

$$\alpha_R \approx \frac{8\pi^3}{3\lambda^4} n^8 p^2 \beta_T k_B T_f \quad \text{and} \quad \alpha_R \approx \frac{8\pi^3}{3\lambda^4} (n^2 - 1)^2 \beta_T k_B T_f$$

in which λ is the free-space wavelength, n is the refractive index at the wavelength of interest, β_T is the isothermal compressibility (at T_f) of the glass, k_B is the Boltzmann constant, and T_f is a quantity called the *fictive temperature* (or the glass transition temperature) at which the liquid structure during the cooling of the fiber is frozen to become the glass structure. Fiber is drawn at high temperatures. As the fiber cools, eventually the temperature drops sufficiently for the atomic motions to be so sluggish that the structure becomes essentially “frozen-in” and remains like this even at room temperature. Thus, T_f marks the temperature below which the liquid structure is frozen and hence the density fluctuations are also frozen into the glass structure. Use these two equations and calculate the attenuation in dB km⁻¹ due to Rayleigh scattering at $\lambda = 1.55 \mu\text{m}$ given that pure silica (SiO_2) has the following properties: $T_f \approx 1180^\circ\text{C}$; $\beta_T \approx 7 \times 10^{-11} \text{ m}^2 \text{ N}^{-1}$ (at high temperatures); $n = 1.5$ at $1.5 \mu\text{m}$; $p = 0.28$. The lowest reported attenuation around this wavelength is about 0.14 dB km⁻¹. What is your conclusion? (Note that α_B above is in Nepers per meter.)

2.35 Bending loss Bending losses always increase with the mode field diameter. Since the MFD increases for decreasing V , $2w \approx 2 \times 2.6a/V$, smaller V fibers have higher bending losses. How does the bending loss

³²For example, R. Olshansky, *Rev. Mod. Phys.*, 51, 341, 1979.

α vs. radius of curvature R behavior look like on a semilogarithmic plot [as in Figure 2.39 (a)] for two values of the V -number V_1 and V_2 if $V_2 > V_1$? It is found that for a single-mode fiber with a cutoff wavelength $\lambda_c = 1180$ nm, operating at 1300 nm, the microbending loss reaches 1 dB m^{-1} when the radius of curvature of the bend is roughly 6 mm for $\Delta = 0.00825$, 12 mm for $\Delta = 0.00550$, and 35 mm for $\Delta = 0.00275$. Explain these findings.

- 2.36 Bend loss reduced fibers** Consider the bend loss measurements listed in Table 2.8 for four different types of fiber. The trench fibers have a trench placed in the cladding where the refractive index is lowered as shown in Figure 2.39. The nanoengineered fiber is shown in Figure 2.55. There is a ring of region in the cladding in which there are gas-filled nanoscale voids. (They are introduced during fabrication.) A void in the ring has a circular cross-section but has a length along the fiber that can be a few meters. These voids occupy a volume in the ring that is only 1–10%. Plot the bending loss semilogarithmically (α on a log scale and R on a linear scale) and fit the data to $\alpha_{\text{micobend}} = A \exp(-R/R_c)$ and find A and R_c . What is your conclusion? Suppose that we set our maximum acceptable bending loss to 0.1 dB turn^{-1} in installation (the present goal is to bring the bending loss to below 0.1 dB turn^{-1}). What are the allowed radii of curvature for each turn?

TABLE 2.8 Bend radius R in mm, α in dB turn^{-1} . Data over $1.55\text{--}1.65 \mu\text{m}$ range

Standard SMF ^a 1550 nm		Trench Fiber 1 ^b 1650 nm		Trench Fiber 2 ^c 1625 nm		Nanoengineered Fiber ^a 1550 nm	
R (mm)	α (dB turn^{-1})	R (mm)	α (dB turn^{-1})	R (mm)	α (dB turn^{-1})	R (mm)	α (dB turn^{-1})
5.0	15.0	7.50	0.354	5.0	0.178	5.0	0.031
7.0	4.00	10.0	0.135	7.5	0.0619	7.5	0.0081
10.0	0.611	15.0	0.020	10.0	0.0162	10.0	0.0030
12.5	0.124			15.0	0.00092	15.0	0.00018
16.0	0.0105						
17.5	0.0040						

(Source: Data used from a number of sources: ^aM.-J. Li *et al.*, *J. Light Wave Technol.*, 27, 376, 2009; ^bK. Himeno *et al.*, *J. Light Wave Technol.*, 23, 3494, 2005; ^cL.-A. de Montmorillon, *et al.* “Bend-Optimized G.652D Compatible Trench-Assisted Single-Mode Fibers,” *Proceedings of the 55th IWCS/Focus*, pp. 342–347, November, 2006.)

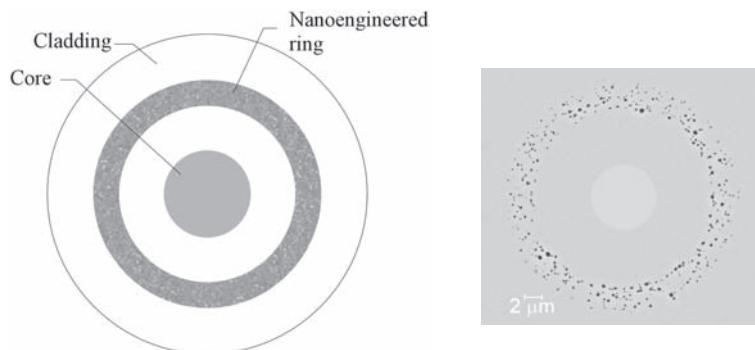


FIGURE 2.55 Left: The basic structure of bend-insensitive fiber with a nanoengineered ring in the cladding. Right: An SEM picture of the cross-section of a nanoengineered fiber with reduced bending losses. (Courtesy of Ming-Jun Li, Corning Inc. For more information see U.S. Patent 8,055,110, 2011.)

- 2.37 Microbending loss** Microbending loss α_B depends on the fiber characteristics and wavelength. We will calculate α_B approximately given various fiber parameters using the single-mode fiber microbending loss equation (D. Marcuse, *J. Opt. Soc. Am.*, 66, 216, 1976)

$$\alpha_B \approx \frac{\pi^{1/2} \kappa^2}{2\gamma^{3/2} V^2 [K_1(\gamma a)]^2} R^{-1/2} \exp\left(-\frac{2\gamma^3}{3\beta^2} R\right)$$

where R is the bend radius of curvature; a is the fiber radius; β is the propagation constant, determined by b , normalized propagation constant, which is related to V , $\beta = n_2 k [1 + b \Delta]$; $k = 2\pi/\lambda$ is the free-space wave vector; $\gamma = [\beta^2 - n_2^2 k^2]^{1/2}$; $\kappa = [n_1^2 k^2 - \beta^2]^{1/2}$; and $K_1(x)$ is a first-order modified Bessel function, available in math software packages. The normalized propagation constant b can be found from $b = (1.1428 - 0.996 V^{-1})^2$. Consider a single-mode fiber with $n_1 = 1.450$, $n_2 = 1.446$, $2a$ (diameter) = $3.9 \mu\text{m}$. Plot α_B vs. R for $\lambda = 633 \text{ nm}$ and 790 nm from $R = 2 \text{ mm}$ to 15 mm . Figure 2.56 shows the experimental results on an SMF that has the same properties as the fiber above. While the data in Figure 2.56 are for a 10 cm fiber, the equation above gives the attenuation per meter. We need to divide the calculated values by 10 to compare the two sets of data. What is your conclusion? (Note that while the data in Figure 2.56 are for a 10 cm fiber, the equation above gives the attenuation per meter. We need to divide the calculated attenuation values by 10 to compare the two sets of data. You might wish to compare your calculations with the experiments of A. J. Harris and P. F. Castle, *IEEE J. Light Wave Technol.*, LT4, 34, 1986).

Attenuation for 10 cm of bend

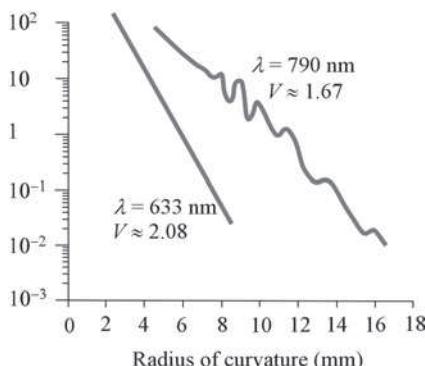


FIGURE 2.56 Measured microbending loss (attenuation) for a 10-cm fiber bent by different amounts of radius of curvature R . Single-mode fiber with a core diameter of $3.9 \mu\text{m}$, cladding radius of $48 \mu\text{m}$, $\Delta = 0.00275$, $\text{NA} \approx 0.10$, $V \approx 1.67$ and 2.08 . (Source: Data extracted from A. J. Harris and P. F. Castle, "Bend Loss Measurements on High Aperture Single-Mode Fibers as a Function of Wavelength and Bend Radius," *IEEE J. Light Wave Technol.*, LT14, 34, 1986, and replotted with a smoothed curve; [see original article for the discussion of peaks in α_B vs. R at 790 nm].)

- 2.38 Fiber Bragg Grating (FBG)** A silica fiber-based FBG is required to operate at 850 nm . What should be the periodicity of the grating Λ ? If the amplitude of the index variation Δn is 2×10^{-5} and total length of the FBG is 5 mm , what are the maximum reflectance at the Bragg wavelength and the bandwidth of the FBG? Assume that the effective refractive index \bar{n} is 1.460 . What are the reflectance and the bandwidth if Δn is 2×10^{-4} ?



Fiber Bragg Grating (FBG)-based optical strain sensor for monitoring strains in civil structures, for example, bridges, dams, buildings, and tunnels. The fiber is mounted on a steel carrier and maintained stretched. (Courtesy of Micron Optics.)

- 2.39 Fiber Bragg Grating sensor array** Consider a FBG strain sensor array embedded in a silica fiber that is used to measure strain at various locations on an object. Two neighboring sensors have grating periodicities of $\Lambda_1 = 534.5 \text{ nm}$ and $\Lambda_2 = 539.7 \text{ nm}$. The effective refractive index is 1.450 and the photoelastic coefficient is 0.22 . What is the maximum strain that can be measured? What would be the main problem with this sensor array? What is the strain at fracture if the fiber fractures roughly at an applied stress of 700 MPa and the elastic modulus is 70 GPa ? What is your conclusion?



Fiber Bragg Grating (FBG)-based optical temperature sensors for use from -200°C to 250°C . FBGs are mounted in different packages: top left, all dielectric; top right, stainless steel; and bottom, copper. The sensors operate over 1510 – 1590 nm . (Courtesy of Lake Shore Cryotronics Inc.)

I wanted visible light, because I knew that if I could get red light, other colors would be possible.

—Nick Holonyak Jr.



Nick Holonyak Jr. carried out the early work in the development of practical light-emitting diodes (LEDs) in the visible spectrum during the 1960s while working as a consulting research scientist for General Electric Co. in Syracuse. He made his first visible laser-LED in 1962, which emitted red light.¹ It served as the basis for the concept of the “ultimate lamp.” Since 1963, he has been at the University of Illinois at Urbana-Champaign, where he currently holds the John Bardeen Endowed Chair. In the February 1963 issue of *Readers Digest*, Nick Holonyak Jr. suggested that the incandescent light bulb will eventually be replaced by the LED. The above photo was taken around 1970–1975. (*Courtesy of University of Illinois at Urbana-Champaign.*)

¹N. Holonyak Jr. and S. F. Bevacqua, “Coherent (Visible) Light Emission from a $\text{Ga}(\text{As}_x\text{P}_{1-x})$ Junction,” *Appl. Phys. Letts.*, 1, 82, 1962.

3

SEMICONDUCTOR SCIENCE AND LIGHT-EMITTING DIODES

3.1 REVIEW OF SEMICONDUCTOR CONCEPTS AND ENERGY BANDS

A. Energy Band Diagrams, Density of States, Fermi–Dirac Function and Metals

We know from modern physics that the energy of the electron in an atom is quantized and can only have certain discrete values as illustrated for the Li atom in Figure 3.1. The isolated Li atom has two electrons in the 1s shell and one electron in the 2s subshell. The same concept also applies to the electron energy in a molecule with several atoms. Again the electron energy is quantized. However, when we bring together something like 10^{23} Li atoms to form the metal crystal, the interatomic interactions result in the formation of electron **energy bands**. The 2s energy level splits into some 10^{23} closely spaced energy levels that effectively form an **energy band**, which is called the 2s band. Similarly other higher energy levels also form bands as illustrated in Figure 3.1. These energy bands overlap to form one continuous energy band which represents the energy band structure of a metal. The 2s energy level in the Li atom is half full (2s subshell needs 2 electrons but has 1) which means that the 2s band in the crystal will also be half full. Metals characteristically have partially filled energy bands.

In the case of metals, the energy band diagram is straightforward and essentially one band of energies in which the energy levels from the bottom of the band to a certain level are all filled at zero Kelvin (0 K), as indicated in Figure 3.1. Each energy level in this band corresponds to a wavefunction for the electron inside the crystal, and represents a **quantum state** for the electron. Thus, the energy levels are often referred to as energy states. In the solid, the bottom of the band is normally taken as the zero of energy so that electrons fill all the states until a certain energy called the **Fermi energy**, denoted as E_F —the highest occupied energy level at 0 K. The states above E_F are empty up to the vacuum level. The energy required to remove an electron from the metal is the energy involved in taking an electron from E_F to the vacuum level, which is called the **work function** Φ of the metal. For example, a photon with an energy $h\nu$ greater than Φ that is incident on a metal can cause an electron at E_F to become ejected out from the metal to become free, a phenomenon called the **photoelectric effect**. The electron is said to be **photoemitted**. Many photomultiplier tubes in optoelectronics have a cathode material with the right Φ that allows electrons to be photoemitted when light within the desired range of wavelengths is incident on the cathode.

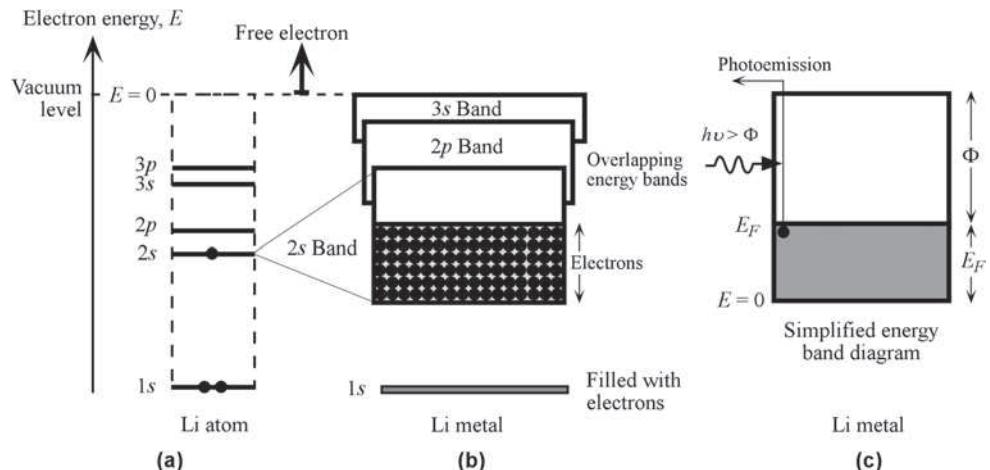


FIGURE 3.1 (a) Energy levels in a Li atom are discrete. (b) The energy levels corresponding to outer shells of isolated Li atoms form an energy band inside the crystal—for example, the 2s level forms a 2s band. Energy levels form a quasi-continuum of energy within the energy band. Various energy bands overlap to give a single band of energies that is only partially full of electrons. There are states with energies up to the vacuum level where the electron is free. (c) A simplified energy band diagram for a metal, and the photoelectric effect.

The photoemitted electrons are then accelerated by the applied field inside the multiplier tube and are made incident on other electrodes to cause secondary electron emission and hence multiplication. (One incident photon can give rise to 10^6 collected electrons.)

The electrons around E_F control many of the properties of metals such as electrical and thermal conductivity. When an electric field E is applied, an electron at E_F can gain energy from the external force eE and be able to move to a higher energy state since these are empty and available. Electrical conduction occurs by the drift of electrons at around E_F .

Many important properties of both metals (and semiconductors below) are described by considering the behavior of electrons within an energy band. In the case of metals, this is called the **free electron model**; that is, the electrons are free inside the metal and hence electron potential energy is constant. These electrons are what are normally called conduction electrons.

There are two important concepts in describing free electrons in an energy band. The **density of states** ($g(E)$) represents the number of electronic states (electron wavefunctions) in a band per unit energy per unit volume of the crystal. We can use quantum mechanics to calculate the DOS by considering how many electron wavefunctions there are within a given energy range per unit volume of the crystal. Figure 3.2 (a) shows the energy band diagram of a metal, and (b) shows, in a simplified way, how $g(E)$ depends on the electron energy in the band. According to quantum mechanics, for an electron confined within a three-dimensional potential energy well, as one might expect for a conduction electron in the crystal of a metal, DOS $g(E)$ increases with energy from the bottom of the band and is given by

Density of states in a band

$$g(E) = 4\pi (2m_e)^{3/2} h^{-3} E^{1/2} = AE^{1/2} \quad (3.1.1)$$

where E is the electron energy from the bottom of the band, m_e is the mass of the electron, h is Planck's constant, and A is a constant that represents the quantities multiplying $E^{1/2}$ in the second term. The DOS gives information on available states, that is, wavefunctions, for the electron and not on their actual occupation.

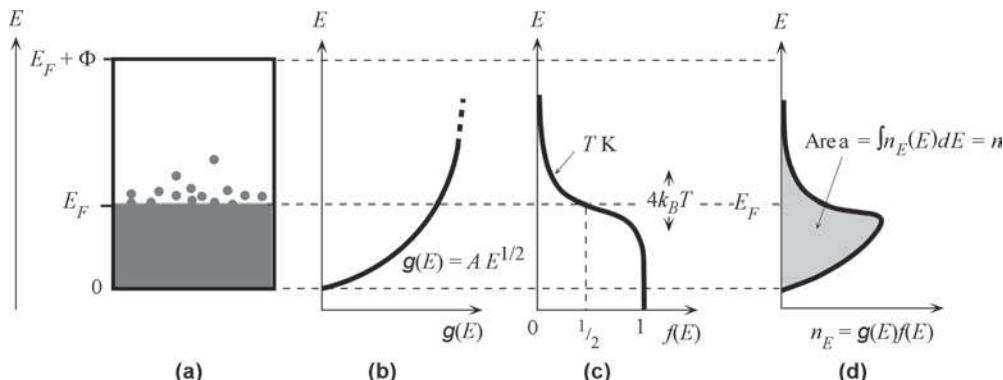


FIGURE 3.2 (a) Above 0 K, due to thermal excitation, some of the electrons are at energies above E_F . (b) The density of states, $g(E)$ vs. E in the band. (c) The probability of occupancy of a state at an energy E is $f(E)$. The product $g(E)f(E)$ is the number of electrons per unit energy per unit volume or electron concentration per unit energy. The area under the curve with the energy axis is the concentration of electrons in the band, n .

The **Fermi–Dirac function** $f(E)$ is the probability of finding an electron in a quantum state with energy E (state implies a wavefunction). It is a fundamental property of a collection of interacting electrons in *thermal equilibrium*. It is given by

$$f(E) = \frac{1}{1 + \exp\left(\frac{E - E_F}{k_B T}\right)} \quad (3.1.2)$$

Fermi–
Dirac
function

where k_B is the Boltzmann constant, T is the temperature (K), and E_F is the Fermi energy. Figure 3.2 (c) shows the dependence of $f(E)$ on the energy E . At low energies well below E_F , all states are occupied and at high energies, well above E_F , the states are empty. At E_F , $f(E_F) = 1/2$, half the states are occupied. Changes in $f(E)$ occur around E_F over several $k_B T$ s of energy, as indicated in Figure 3.2 (c). The higher the temperature, the greater is the reach of $f(E)$ to higher energies, and therefore more electrons at higher energies.

Consider the product $g(E)f(E)$, density of states at $E \times$ probability of finding an electron in a state at E . The product $g(E)f(E)$ represents the number electrons per unit energy per unit volume, $n_E(E)$, in the band which is shown in Figure 3.2 (d). It is essentially the energy distribution of electrons in the band. Thus, $n_E dE = g(E)f(E)dE$ is the number of electrons in the energy range E to $E + dE$. Integrating this from the bottom to the top of the band gives the conduction electron concentration n in the band,

$$n = \int_0^{E_F + \Phi} g(E)f(E)dE \quad (3.1.3)$$

Equations (3.1.1) and (3.1.2) can be substituted in Eq. (3.1.3) to evaluate the integration. Once integrated, the resulting equation will relate n to the Fermi energy E_F at a temperature T . Put differently, E_F depends on n , the total number of electrons per unit volume in the band. At 0 K, the Fermi energy E_{FO} is

$$E_{FO} = \left(\frac{\hbar^2}{8m_e}\right)\left(\frac{3n}{\pi}\right)^{2/3} \quad (3.1.4)$$

Fermi
energy at
0 K for
metals

E_{FO} is typically a few electron volts, that is, orders of magnitude greater than $k_B T$. Above 0 K, E_F decreases slightly with T , though we ignore this tiny change in using metal band diagrams in optoelectronics. Inasmuch as the electrons in the metal are free, they have kinetic energies only. An electron at the Fermi energy E_{FO} has several eVs of kinetic energy, which is orders of magnitude greater than the classical kinetic theory (roughly $3k_B T/2$ for a free particle). The reason is that the electrons in the metal must obey the *Pauli exclusion principle*, which prevents more than two electrons (with spins up and down) from occupying a given energy state. Thus, the electrons must distribute themselves among available states and need to reach higher energies to satisfy the Pauli exclusion principle.

B. Energy Band Diagrams of Semiconductors

The electron energies in a semiconductor crystal, however, are distinctly different than that for metals. Figure 3.3 (a) shows a simplified two-dimensional view of the silicon crystal which has each Si atom bonding to four neighbors. All the four valence electrons per atom are used in these bonds. The interactions between the Si atoms and their valence electrons result in the electron energy in the crystal falling into two distinct energy bands called the **valence band (VB)** and **conduction band (CB)** that are separated by an energy gap, **bandgap (E_g)**, as shown in Figure 3.3 (b). There are no allowed electron energies in the **bandgap**; it represents the forbidden electron energies in the crystal. The valence band represents electron wavefunctions in the crystal that correspond to bonds between the atoms. Electrons that occupy these wavefunctions are the valence electrons. Since at absolute zero of temperature all the bonds are occupied by valence electrons (there are no broken bonds), all the energy levels in the VB are normally filled with these electrons. The CB represents electron wavefunctions in the crystal that have higher energies than those in the VB and are normally empty at zero Kelvin. The top of the VB is labeled E_v , bottom of conduction band E_c , so that $E_g = E_c - E_v$ is the bandgap. The width of the CB is called the **electron affinity χ** .

An electron placed in the CB is free to move around the crystal and also to respond to an electric field because there are plenty of neighboring empty energy levels. This electron can easily gain energy from the field and move to higher energy levels because these states are

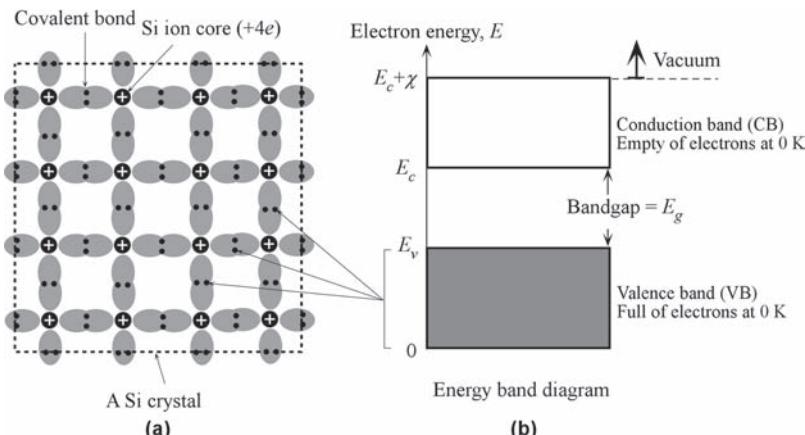


FIGURE 3.3 (a) A simplified two-dimensional view of a region of the Si crystal showing covalent bonds. (b) The energy band diagram of electrons in the Si crystal at absolute zero of temperature. The bottom of the VB has been assigned a zero of energy.

empty in the CB. Generally we can treat an electron in the CB as if it were free within the crystal by simply assigning an **effective mass** m_e^* to it. This effective mass is a quantum mechanical quantity, which takes into account that the electron in the CB interacts with a periodic potential energy as it moves through the crystal so that its inertial resistance to acceleration (definition of mass) is not the same as if it were free in vacuum.

Since the only empty states are in the CB, the excitation of an electron from the VB requires a minimum energy of E_g . Figure 3.4 illustrates what happens when an incident photon of energy $h\nu > E_g$ interacts with an electron in the VB. This electron absorbs the incident photon and gains sufficient energy to surmount the energy gap E_g and reach the CB. Consequently, a free electron in the CB and a “hole,” corresponding to a missing electron in the VB, are created. In some semiconductors, such as Si and Ge, the photon absorption process also involves lattice vibrations (vibrations of the Si atoms) which we have not shown in Figure 3.4.

The empty electronic state, or the missing electron in the bond, is what we call a **hole** in the valence band. The free electron, which is in the CB, can wander around the crystal and contribute to the electrical conduction when an electric field is applied. The region remaining around the hole in the VB is positively charged because a charge of $-e$ has been removed from an otherwise neutral region of the crystal. This hole, denoted as h^+ , can also wander around the crystal as if it were “free.” This is because an electron in a neighboring bond can “jump,” that is, tunnel into the hole, to fill the vacant electronic state at this site and thereby create a hole at its original position as illustrated in Figure 3.5. This is effectively equivalent to the hole being displaced in the opposite direction. Thus, conduction in semiconductors occurs by both electrons and holes with charges $-e$ and $+e$ and their own effective masses m_e^* and m_h^* .

When a wandering electron in the CB meets a hole in the VB, it has found an empty electronic state of lower energy and it therefore occupies it. The electron falls from the CB to the VB to fill the hole. This is called **recombination**, which results in the annihilation of an electron in the CB and a hole in the VB as in Figure 3.5. The excess energy of the electron falling from CB to VB in certain semiconductors such as GaAs and InP is emitted as a photon. In Si and Ge the excess energy is lost as lattice vibrations (heat), that is, as phonons. In the steady state, the thermal generation rate is balanced by the recombination rate so that the electron concentration n in the CB and hole concentration p in the VB remain constant; both n and p depend on the temperature.

Although in the specific example in Figure 3.4, a photon of energy $h\nu > E_g$ creates an electron–hole pair, other sources of energy can also lead to an electron–hole pair creation. In fact, in the absence of radiation, there is still an electron and hole pair generation process going on in the sample as a result of **thermal generation**. Due to thermal energy, the atoms in the crystal

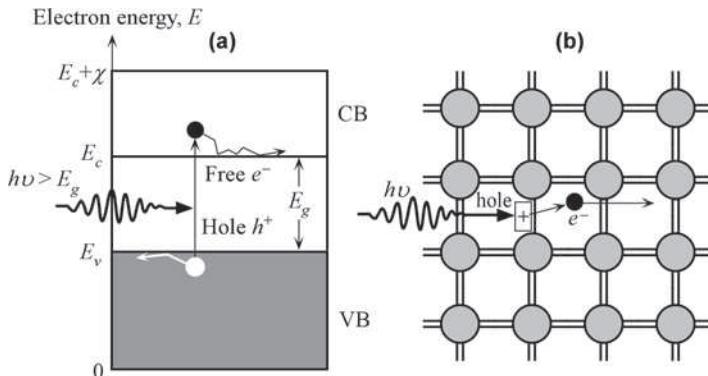


FIGURE 3.4 (a) A photon with an energy $h\nu$ greater than E_g can excite an electron from the VB to the CB. (b) Each line with Si-Si atoms is a valence electron in a bond. When a photon breaks a Si-Si bond, a free electron and a hole in the Si-Si bond are created. The result is the photogeneration of an electron–hole pair (EHP).

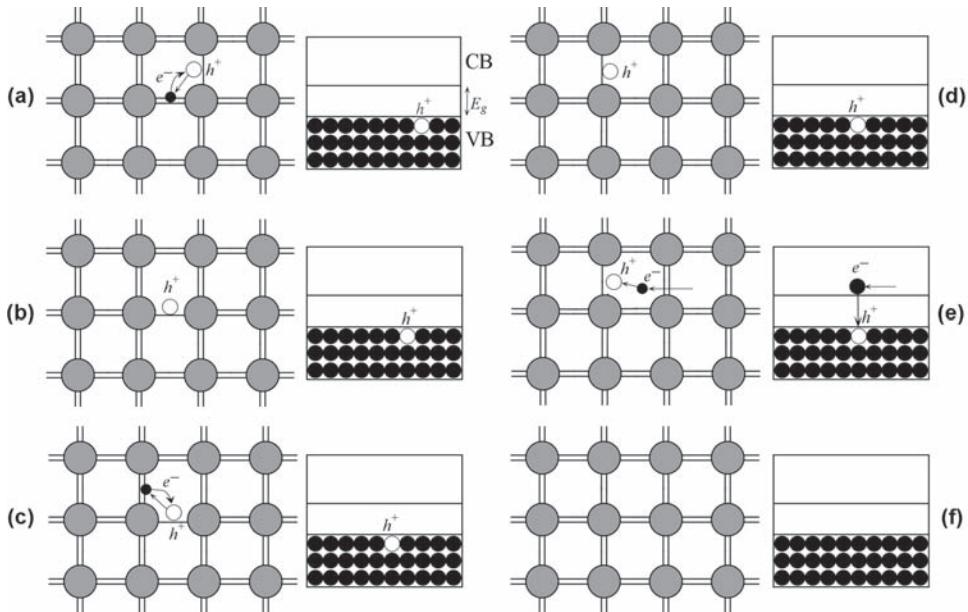


FIGURE 3.5 A pictorial illustration of a hole in the valence band wandering around the crystal due to the tunneling of electrons from neighboring bonds, and its eventual recombination with a wandering electron in the conduction band. A missing electron in a bond represents a hole as in (a). An electron in a neighboring bond can tunnel into this empty state and thereby cause the hole to be displaced as in (a) to (d). The hole is able to wander around in the crystal as if it were free but with a different effective mass than the electron. A wandering electron in the CB meets a hole in the VB in (e), which results in the recombination and the filling of the empty VB state as in (f).

are constantly vibrating which corresponds to the bonds between the Si atoms being periodically deformed with a distribution of energies. Energetic vibrations can rupture bonds and thereby create electron–hole pairs, which correspond to the excitation of electrons from the VB to the CB.

3.2 SEMICONDUCTOR STATISTICS

Many important properties of semiconductors are described by considering the behavior of electrons in the CB and holes in the VB. We need to consider the density of states (DOS) $g_{\text{CB}}(E)$ of electronic states in the CB, $g_{\text{VB}}(E)$ in the VB and their occupation statistics. Figure 3.6 (a) shows the energy band diagram of a semiconductor, and (b) shows, in a simplified way, how $g_{\text{CB}}(E)$ depends on the electron energy in the CB and VB near the band edges E_c and E_v , respectively. The DOS $g_{\text{CB}}(E)$ increases with energy from the CB edge as

Density
of states
near E_c

$$g_{\text{CB}}(E) = 4\pi(2m_e^*)^{3/2}h^{-3}(E - E_c)^{1/2} \quad (3.2.1)$$

where $(E - E_c)$ is the electron energy from the bottom of the CB. The DOS curves usually have various peaks at certain energies for crystalline solids, which we have neglected as these are not needed for the present discussion.

To find the energy distribution of electrons in the CB, we need the Fermi–Dirac function $f(E)$, the probability of finding an electron in a quantum state with energy E . The behavior of $f(E)$ is shown in Figure 3.6 (c) assuming that the Fermi level E_F is located in the bandgap. (Its position

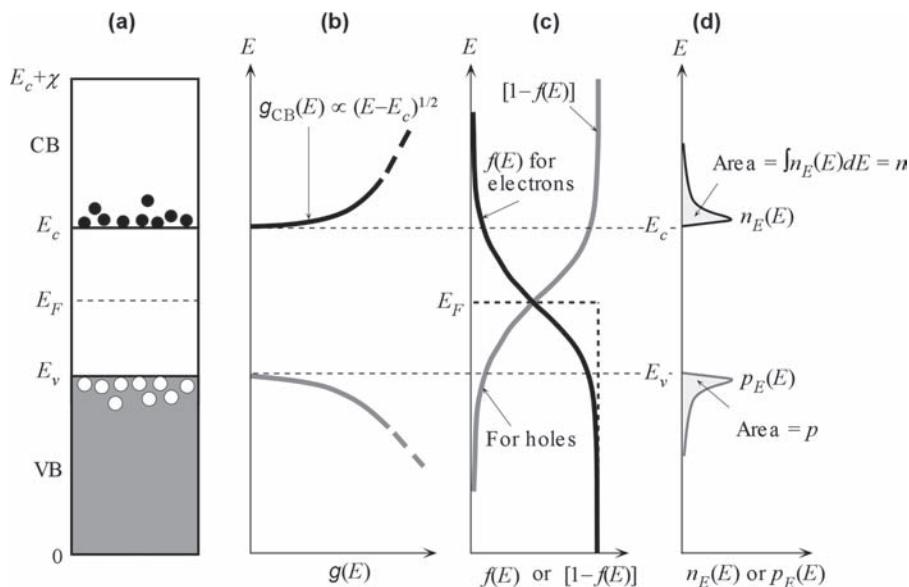


FIGURE 3.6 (a) Energy band diagram of a semiconductor, assuming that the Fermi level is near the middle of the bandgap. (b) Density of states (number of states per unit energy per unit volume). (c) Fermi–Dirac probability function (probability of occupancy of a state). (d) The product of $g(E)$ and $f(E)$ is the energy density of electrons in the CB (number of electrons per unit energy per unit volume). The area under $n_E(E)$ vs. E is the electron concentration.

in the energy gap is discussed later.) When we multiply $g(E)$ with $f(E)$, we get the number of electrons per unit energy per unit volume at E in the CB, $n_E(E)$, which is shown in Figure 3.6 (d). Thus, $n_E dE = g_{\text{CB}}(E)f(E)dE$ is the number of electrons in the energy range [Figure 3.6 (d)] E to $E + dE$. Integrating this from the bottom (E_c) to the top ($E_c + \chi$) of the CB gives the electron concentration

$$n = \int_{E_c}^{E_c + \chi} g_{\text{CB}}(E)f(E)dE \quad (3.2.2)$$

Electron concentration in CB

Whenever $(E_c - E_F) \gg k_B T$, that is, E_F is at least a few $k_B T$ below E_c , then $f(E) \approx \exp[-(E - E_F)/k_B T]$. That is, the Fermi–Dirac statistics can be replaced by Boltzmann statistics. Such semiconductors are called **nondegenerate**. It implies that the number of electrons in the CB is far less than the number states in this band. For nondegenerate semiconductors, the above integration leads to

$$n = N_c \exp\left(-\frac{E_c - E_F}{k_B T}\right) \quad (3.2.3)$$

Electron concentration in CB

where $N_c = 2[2\pi m_e^* k_B T / h^2]^{3/2}$ is a constant at a given temperature for a given material, called the **effective density of states at the CB edge**. The result of the integration in Eq. (3.2.3) seems too simple but it is, however, an approximation as it assumes that $(E_c - E_F) \gg k_B T$. We can interpret Eq. (3.2.3) as follows. If we take all the states in the conduction band and replace them with an effective concentration N_c (number of states per unit volume) at E_c and then multiply this simply by the Boltzmann probability function, $f(E_c) = \exp[-(E_c - E_F)/k_B T]$, we would

obtain the concentration of electrons at E_c , that is, in the conduction band. N_c is thus an effective density of states at the CB band edge.

We can carry out a similar analysis for the concentration of holes in the VB as illustrated in Figures 3.6 (a)–(d). Multiplying the density of states $g_{VB}(E)$ in the VB with the probability of occupancy by a hole, $[1 - f(E)]$, gives p_E , the hole concentration per unit energy. Note that the probability of finding a hole, a missing electron, in a state with energy E is $1 - f(E)$. Integrating $p_E(E)$ over the VB gives the hole concentration by assuming that E_F is a few $k_B T$ above E_v , we obtain

Hole
concentration in VB

$$p \approx N_v \exp\left(-\frac{E_F - E_v}{k_B T}\right) \quad (3.2.4)$$

where $N_v = 2[2\pi m_h^* k_B T / h^2]^{3/2}$ is the **effective density of states at the VB edge**, and m_h^* is the effective mass of holes in the VB.

There are no specific assumptions in our derivations above, except for E_F being a few $k_B T$ away from the band edges, which means that Eqs. (3.2.3) and (3.2.4) are generally valid. It is apparent from Eqs. (3.2.3) and (3.2.4) that the location of E_F determines the electron and hole concentrations. Thus, E_F can be viewed as a useful semiconductor property that allows us to represent electron and hole concentrations in the energy band diagram. In an intrinsic semiconductor (a pure crystal), $n = p$, by using Eqs. (3.2.3) and (3.2.4) we can show that the Fermi level E_{Fi} in the *intrinsic crystal* is above E_v and located in the bandgap at

Fermi level
in intrinsic
crystal

$$E_{Fi} = E_v + \frac{1}{2} E_g - \frac{1}{2} kT \ln\left(\frac{N_c}{N_v}\right) \quad (3.2.5)$$

Typically N_c and N_v values are comparable and both occur in the logarithmic term so that E_{Fi} is approximately in the middle of the bandgap as originally sketched in Figure 3.6.

There is a useful semiconductor relation between n and p , called the **mass action law**. From Eqs. (3.2.3) and (3.2.4), the product np is

Mass
action law

$$np = N_c N_v \exp\left(-\frac{E_g}{k_B T}\right) = n_i^2 \quad (3.2.6)$$

where $E_g = E_c - E_v$ is the bandgap energy and n_i^2 has been *defined as* $N_c N_v \exp(-E_g/k_B T)$. The quantity n_i is a material property; for example, it depends on E_g , but not on the position of the Fermi level. It also depends on the temperature. The intrinsic concentration n_i corresponds to the concentration of electrons or holes in an undoped (pure) crystal, that is, intrinsic semiconductor. In such a semiconductor $n = p = n_i$, which is therefore called the **intrinsic concentration**. The mass action law is valid whenever we have thermal equilibrium and the sample is in the dark (without any photogeneration).

Equations (3.2.3) and (3.2.4) determine the total concentration of electrons and holes in the CB and VB, respectively. The average energy of the electrons in the CB can be calculated by using $n_E(E)$, their energy distribution. The result gives the average energy as $(3/2)k_B T$ above E_c . Since the electron in the CB is “free” in the crystal with an effective mass m_e^* , it wanders around the crystal with an average kinetic energy $(3/2)k_B T$, which is the same as the kinetic energy of a free atom in a gas or vapor in a tank. This is not surprising as both particles are wandering freely (without interacting with each other) and obey Boltzmann statistics. If v is the electron velocity and triangular brackets represent an average, then $\langle(1/2)m_e^* v^2\rangle$ must be $(3/2)k_B T$. We can thus calculate the *root mean square velocity* $(\langle v^2 \rangle)^{1/2}$ which is called the **thermal velocity** (v_{th}) and is typically $\sim 10^5 \text{ m s}^{-1}$. The same ideas apply to holes in the VB with an effective hole mass m_h^* .

It is important to understand what the **Fermi energy** represents. First, it is a convenient way to represent free carrier concentrations (n in the CB and p in the VB) on the energy band diagram. However, the most useful property of E_F is in terms of a change in E_F . Any change ΔE_F across a material system represents electrical work input or output per electron.² If V is the potential difference between two points in a material system, then

$$\Delta E_F = eV \quad (3.2.7)$$

For a semiconductor system in equilibrium, in the dark, and with no applied voltage or no EMF generated, $\Delta E_F = 0$ and E_F must be uniform across the system.

Fermi
energy and
electrical
work

3.3 EXTRINSIC SEMICONDUCTORS

A. *n*-Type and *p*-Type Semiconductors

By introducing small amounts of impurities into an otherwise-pure crystal, it is possible to obtain a semiconductor in which the concentration of carriers of one polarity is much in excess of the other type. Such semiconductors are referred to as **extrinsic semiconductors** vis-à-vis the intrinsic case of a pure and perfect crystal. For example, by adding pentavalent impurities, such as arsenic, which have a valency one more than Si, we can obtain a semiconductor in which the electron concentration is much larger than the hole concentration. In this case we will have an ***n*-type semiconductor**. If we add trivalent impurities, such as boron, which have a valency of one less than four, we then have an excess of holes over electrons—a ***p*-type semiconductor**.

An arsenic (As) atom has five valence electrons whereas Si has four. When the Si crystal is doped with small amounts of As, each As atom substitutes for one Si atom and is surrounded by four Si atoms. When an As atom bonds with four Si atoms, it has one electron left unbound. This fifth electron cannot find a bond to go into so it is left orbiting around the As atom, which looks like an As^+ , as illustrated in Figure 3.7 (a). The As^+ ionic center, with an electron e^- orbiting

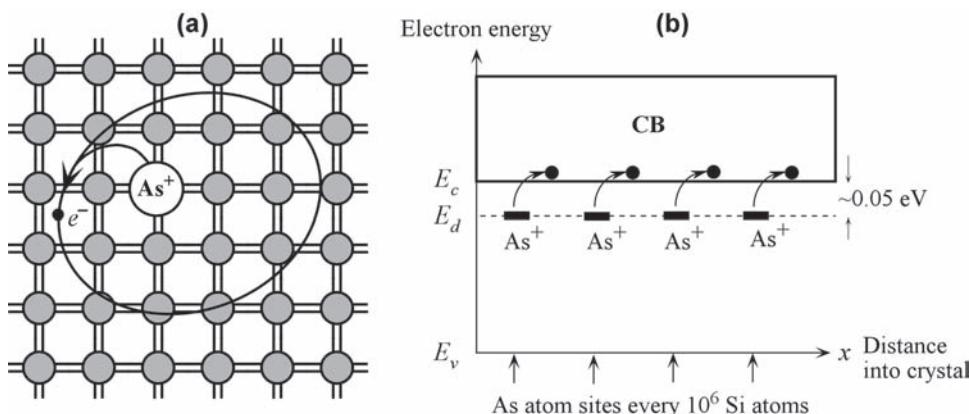


FIGURE 3.7 (a) The four valence electrons of As allow it to bond just like Si but the fifth electron is left orbiting the As site. The energy required to release to free fifth electron into the CB is very small. (b) Energy band diagram for an *n*-type Si doped with 1 ppm As. There are donor energy levels just below E_c around As^+ sites.

²For readers familiar with thermodynamics, its rigorous definition is that E_F is the *chemical potential per electron*, that is, Gibbs free energy per electron. The definition in Eqs. (3.2.7) is in terms of a change in E_F .

it, resembles a hydrogen atom in a silicon environment. We can easily calculate how much energy is required to free this electron away from the As site, thereby ionizing the As impurity by using our knowledge on the ionization of a hydrogen atom (removing the electron from the H-atom). This energy turns out to be a few hundredths of an electronvolt, that is, ~ 0.05 eV, which is comparable to the thermal energy at room temperature ($\sim k_B T = 0.025$ eV). Thus, the fifth valence electron can be readily freed by thermal vibrations of the Si lattice. The electron will then be “free” in the semiconductor, or in other words, it will be in the CB. The energy required to excite the electron to the CB is therefore ~ 0.05 eV. The addition of As atoms introduces localized electronic states at the As sites because the fifth electron has a localized wavefunction, of the hydrogenic type, around As^+ . The energy of these states, E_d , is ~ 0.05 eV below E_c because this is how much energy is required to take the electron away into the CB. Thermal excitation by lattice vibrations at room temperature is sufficient to ionize the As atom, that is, excite the electron from E_d into the CB. This process creates free electrons; however, the As^+ ions remain immobile as shown in the energy band diagram of an *n*-type semiconductor in Figure 3.7 (b).

Because the As atom donates an electron into the CB, it is called a **donor** impurity. E_d is the electron energy around the donor atom and it is below E_c by ~ 0.05 eV as in Figure 3.7 (b). If N_d is the donor atom concentration in the crystal, provided that $N_d \gg n_i$, then at room temperature, the electron concentration in the CB will nearly be equal to N_d , that is, $n \approx N_d$. The hole concentration will be $p = n_i^2/N_d$, which is less than the intrinsic concentration. The reason is that a small fraction of the large number of electrons in the CB recombine with holes in the VB so as to maintain $np = n_i^2$. The mass action law must be maintained in equilibrium.

The conductivity σ of a semiconductor depends on both electrons and holes as both contribute to charge transport. If μ_e and μ_h are the drift mobilities of the electrons and holes, respectively, then

Semi-conductor conductivity

$$\sigma = en\mu_e + ep\mu_h \quad (3.3.1)$$

which for an *n*-type semiconductor becomes,

n-type conductivity

$$\sigma = eN_d\mu_e + e\left(\frac{n_i^2}{N_d}\right)\mu_h \approx eN_d\mu_e \quad (3.3.2)$$

We should, by similar arguments to the above, anticipate that doping a Si crystal with a trivalent atom (valency of 3) such as B (boron) will result in a *p*-type Si that has an excess of holes in the crystal. Consider doping Si with small amounts of B as shown in Figure 3.8 (a). Because B has only three valence electrons, when it shares them with four neighboring Si atoms one of the bonds has a missing electron which is of course a “hole.” A nearby electron can tunnel into this hole and displace the hole further away from the B atom. As the hole moves away it gets attracted by the negative charge left behind on the B atom. The binding energy of this hole to the B^- ion (a B atom that has accepted an electron) can be calculated using the hydrogenic atom analogy just like in the *n*-type Si case. This binding energy also turns out to be very small, ~ 0.05 eV, so that at room temperature the thermal vibrations of the lattice can free the hole away from the B^- site. A free hole, we recall, exists in the VB. The escape of the hole from the B^- site involves the B atom accepting an electron from a neighboring Si-Si bond (from the VB) which effectively results in the hole being displaced away, and its eventual escape to freedom in the VB. The B atom introduced into the Si crystal therefore acts as an electron **acceptor** impurity. The electron accepted by the B atom comes from a nearby bond. On the energy band diagram, an electron leaves the VB and gets accepted by a B atom which becomes negatively charged. This process leaves a hole in the VB which is free to wander away as illustrated in Figure 3.8 (b).

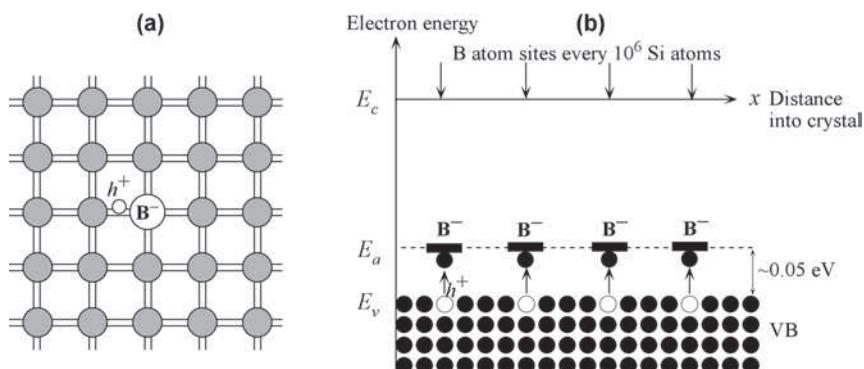


FIGURE 3.8 (a) Boron-doped Si crystal. B has only three valence electrons. When it substitutes for a Si atom one of its bonds has an electron missing and therefore a hole. (b) Energy band diagram for a *p*-type Si doped with 1 ppm B. There are acceptor energy levels just above E_v around B^- sites. These acceptor levels accept electrons from the VB and therefore create holes in the VB.

It is apparent that doping a silicon crystal with a trivalent impurity results in a *p*-type material. We have many more holes than electrons for electrical conduction since the negatively charged B atoms are immobile and hence cannot contribute to the conductivity. If the concentration of acceptor impurities N_a in the crystal is much greater than the intrinsic concentration n_i , then at room temperature all the acceptors would have been ionized and thus $p \approx N_a$. The electron concentration is then determined by the mass action law, $n = n_i^2/N_a$, which is much smaller than p , and consequently the conductivity is simply given by $\sigma = eN_a\mu_h$, similar to Eq. (3.3.2).

Figures 3.9 (a)–(c) show the energy band diagrams of an intrinsic, an *n*-type, and a *p*-type semiconductor. The energy distance of E_F from E_c and E_v determines the electron and hole concentrations by virtue of Eqs. (3.2.3) and (3.2.4). Note the locations of the Fermi level in each case: E_{Fi} for intrinsic; E_{Fn} for *n*-type; and E_{Fp} for *p*-type.

The position of the Fermi level in the energy band diagram plays an important role in understanding device principles. It is clear from Eqs. (3.2.3) and (3.2.4) that its location in the energy band diagram acts as a lever in finding the free electron and hole concentrations in the CB and VB, respectively, as highlighted in Figures 3.9 (a)–(c). It represents the electron

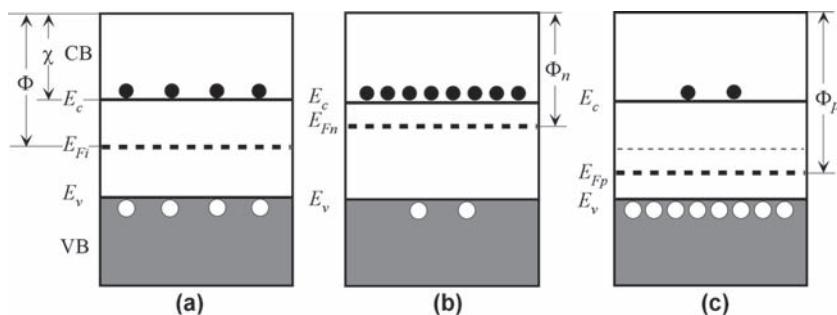


FIGURE 3.9 Energy band diagrams for (a) intrinsic, (b) *n*-type, and (c) *p*-type semiconductors. In all cases, $np = n_i^2$. Note that donor and acceptor energy levels are not shown. CB = conduction band, VB = valence band, E_c = CB edge, E_v = VB edge, E_{Fi} = Fermi level in intrinsic semiconductor, E_{Fn} = Fermi level in *n*-type semiconductor, E_{Fp} = Fermi level in *p*-type semiconductor. χ is the electron affinity. Φ , Φ_n , and Φ_p are the work functions for the intrinsic, *n*-type, and *p*-type semiconductors, respectively.

TABLE 3.1 Selected typical properties of various semiconductors at 300 K

a (nm)	E_g (eV)	χ (eV)	N_c (cm $^{-3}$)	N_v (cm $^{-3}$)	n_i (cm $^{-3}$)	ϵ_r	μ_e (cm 2 V $^{-1}$ s $^{-1}$)	μ_h (cm 2 V $^{-1}$ s $^{-1}$)
Ge (DI)	0.5650	0.66 (I)	4.13	1.04×10^{19}	6.0×10^{19}	2.3×10^{13}	16.0	3900
Si (DI)	0.5431	1.11 (I)	4.05	2.8×10^{19}	1.2×10^{19}	1.0×10^{10}	11.8	1450
InP (ZB)	0.5868	1.35 (D)	4.50	5.2×10^{17}	1.1×10^{19}	3.0×10^7	12.6	4600
GaAs (ZB)	0.5653	1.42 (D)	4.07	4.7×10^{17}	7.0×10^{18}	2.1×10^6	13.0	8500
AlAs (ZB)	0.5661	2.17 (I)	3.50	1.5×10^{19}	1.7×10^{19}	10	10.1	200

Notes: Data combined from a number of sources. I and D represent indirect and direct bandgap. DI, diamond crystal; ZB, zinc blend; a , lattice constant. (Note that there are variations in the values of certain properties among books, for example, n_i for Si, ϵ_r for GaAs, etc. Most commonly used or recent values have been selected.)

and hole concentrations in a semiconductor in the energy band diagram. The work function Φ of a semiconductor is defined as the energy distance from the Fermi level to the vacuum level as shown in Figures 3.9 (a)–(c) as Φ , Φ_n , and Φ_p for the intrinsic, n -type, and p -type semiconductor, respectively. Although there are no states at E_F , the work function nonetheless represents the *average energy* required to remove an electron from a semiconductor.³

The following definitions and notation are generally used to describe extrinsic semiconductors. Electrons in an n -type semiconductor ($n > p$) are **majority carriers** whereas holes are **minority carriers**. The electron, that is, majority carrier, concentration in this n -type semiconductor in equilibrium is n_{no} , where the subscripts refer to the n -type semiconductor and equilibrium (which excludes photoexcitation). The hole, that is, minority carrier, concentration is denoted p_{no} . In this notation, $n_{no} = N_d$ and the mass action law is $n_{no}p_{no} = n_i^2$. Similarly, hole or majority carrier concentration in a p -type semiconductor ($p > n$) is p_{po} and electron, minority carrier, concentration is n_{po} . Further $p_{po} = N_a$ and $n_{po}p_{po} = n_i^2$.

Table 3.1 summarizes some of the properties of useful semiconductors, including Si and GaAs, at 300 K.

B. Compensation Doping

Compensation doping describes the doping of a semiconductor with both donors and acceptors to control the properties. For example, a p -type semiconductor doped with N_a acceptors can be converted to an n -type semiconductor by simply adding donors until the concentration N_d exceeds N_a . The effect of donors compensates for the effect of acceptors and vice versa. The electron concentration is then given by $N_d - N_a$ provided the latter is larger than n_i . When both acceptors and donors are present, what essentially happens is that electrons from donors recombine with the holes from the acceptors so that the mass action law $np = n_i^2$ is obeyed. Remember that we cannot simultaneously increase the electron and hole concentrations because that leads to an increase in the recombination rate which returns the electron and hole concentrations to values that satisfy $np = n_i^2$. Suppose that we have more donors than acceptors. When an acceptor atom accepts a valence band electron, a hole is created in the VB. This hole then recombines with an electron from the CB; the electron would have been supplied by a donor. If we take the initial electron concentration as $n = N_d$ then the recombination between the electrons from the donors and N_a holes generated by N_a acceptors

³The following intuitive argument may help. The easiest electron to remove from an intrinsic semiconductor needs an energy χ . To maintain equilibrium, we must also remove an electron from the VB, which needs an energy $E_g + \chi$ so that the average energy is $\frac{1}{2}E_g + \chi$, which is the work function of the intrinsic semiconductor.

results in the electron concentration reduced by N_a to $n = N_d - N_a$. The hole concentration is then $p = n_i^2/(N_d - N_a)$. By a similar argument, if we have more acceptors than donors, the hole concentration becomes $p = N_a - N_d$, with electrons from N_d donors recombining with holes from N_a acceptors. The electron concentration is then $n = n_i^2/(N_a - N_d)$.

C. Nondegenerate and Degenerate Semiconductors

In nondegenerate semiconductors, the number of states in the CB far exceeds the number of electrons so that the likelihood of two electrons trying to occupy the same state is almost nil. This means that the Pauli exclusion principle can be neglected and the electron statistics can be described by the Boltzmann statistics. N_c is a measure of the density of states in the CB. The Boltzmann expression in Eq. (3.2.3) for n is valid only when $n \ll N_c$. Those semiconductors for which $n \ll N_c$ and $p \ll N_v$ are termed **nondegenerate** semiconductors.

When the semiconductor has been excessively doped with donors, n may be so large, typically $10^{19} - 10^{20} \text{ cm}^{-3}$, that it may be comparable to N_c . In that case the Pauli exclusion principle becomes important in the electron statistics and we have to use the Fermi–Dirac statistics. Such a semiconductor exhibits properties that are more metal-like than semiconductor. Semiconductors that have $n \geq N_c$ or $p \geq N_v$ are called **degenerate semiconductors**. For degenerate semiconductors, the Fermi level is inside the band that is involved in electrical conduction. For example, the Fermi level in an n -type degenerate semiconductor is above E_c .

The large carrier concentration in a degenerate semiconductor is due to the heavy doping. For example, as the donor concentration in an n -type semiconductor is increased, at sufficiently high doping levels, the donor atoms become so close to each other that their orbitals overlap to form a narrow energy band, which overlaps and becomes part of the conduction band as illustrated in Figure 3.10 (a). This situation is reminiscent to the valence electrons filling overlapping energy bands in a metal. In a degenerate n -type semiconductor, the Fermi level is therefore within the CB, or above E_c just like E_F is within the band in a metal. Majority of the states between E_c and E_F are full of electrons as indicated in Figure 3.10 (a). Notice that E_c has been pushed down so that the bandgap is *narrower* in heavily doped semiconductors. The random distribution of a large number of donors in the crystal also introduces a random variation in the potential energy (PE) of the electron; the PE is no longer perfectly periodic. Put differently, there are spatial variations in the density of states in the crystal. Such random fluctuations in the PE give rise to a small tail (an extension) in the density of states $g(E)$ in the bandgap, called **band tailing**, so that E_c is no longer sharp in heavily doped crystals.

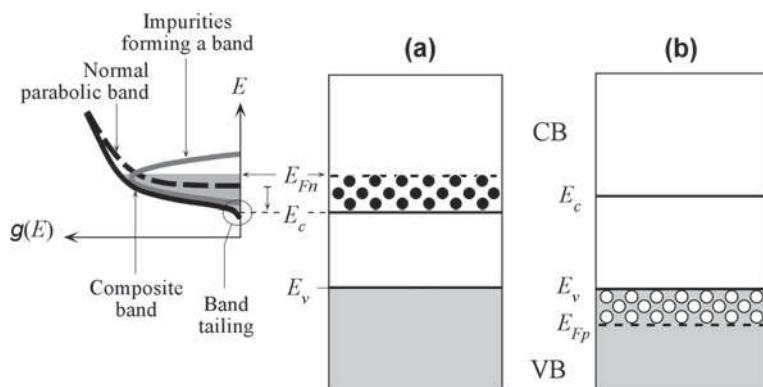


FIGURE 3.10
 (a) Degenerate n -type semiconductor. Large number of donors form a band that overlaps the CB. E_c is pushed down and E_{Fn} is within the CB.
 (b) Degenerate p -type semiconductor.

In the case of a *p*-type degenerate semiconductor, the Fermi level lies in the VB below E_v as in Figure 3.10 (b). One cannot simply assume that $n = N_d$ or $p = N_a$ in a degenerate semiconductor because the dopant concentration is so large that they interact with each other. Not all dopants are able to become ionized and the carrier concentration eventually reaches a saturation typically around $\sim 10^{20} \text{ cm}^{-3}$. Furthermore, the mass action law, $np = n_i^2$, is not valid for degenerate semiconductors.

D. Energy Band Diagrams in an Applied Field

Consider an *n*-type semiconductor that is connected to a voltage supply of V and is carrying a current as shown in Figure 3.11. From basic electrical conduction we know that there is an electric field E inside the semiconductor that drifts the conduction electrons and hence causes a current flow. Consider now the energy band diagram. The Fermi level E_F is above that for the intrinsic case (E_{Fi}), closer to E_c than E_v . The applied voltage drops uniformly along the semiconductor so that the electrons in the semiconductor now also have an imposed electrostatic potential energy, which decreases toward the positive terminal as illustrated in Figure 3.11. The whole band structure, the CB and the VB, therefore tilts. When an electron drifts from *A* toward *B*, its PE decreases because it is approaching the positive terminal.

For a semiconductor system in the dark, in equilibrium and with no applied voltage or no EMF generated, E_F must be uniform across the system since $\Delta E_F = eV = 0$. However, when electrical work is done on the system, for example, when a battery is connected to a semiconductor, E_F is not uniform throughout the whole system. A change ΔE_F in E_F within a material system is equivalent to electrical work per electron or eV . The Fermi level E_F therefore follows the electrostatic PE behavior. The change in E_F from one end to the other, $E_F(A) - E_F(B)$, is just eV , the energy needed in taking an electron through the semiconductor as shown in Figure 3.11. The electron concentration in the semiconductor is uniform so that $E_c - E_F$ must be constant from one end to the other. Thus the CB, VB, and E_F all bend by the same amount.

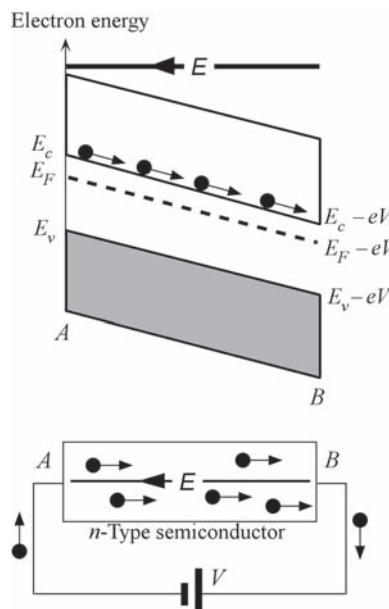


FIGURE 3.11 Energy band diagram of an *n*-type semiconductor connected to a voltage supply of V volts. The whole energy diagram tilts because the electron now has an electrostatic potential energy as well.

EXAMPLE 3.3.1 Fermi levels in semiconductors

An *n*-type Si wafer has been doped uniformly with 10^{16} phosphorus (P) atoms cm^{-3} . Calculate the position of the Fermi energy with respect to the Fermi energy E_{Fi} in intrinsic Si. The above *n*-type Si sample is further doped with 2×10^{17} boron atoms cm^{-3} . Calculate the position of the Fermi energy with respect to the Fermi energy E_{Fi} in intrinsic Si at room temperature (300 K), and hence with respect to the Fermi energy in the *n*-type case above.

Solution

P (Group V) gives *n*-type doping with $N_d = 10^{16} \text{ cm}^{-3}$, and since $N_d \gg n_i (= 10^{10} \text{ cm}^{-3}$ from Table 3.1), we have $n = N_d = 10^{16} \text{ cm}^{-3}$. For intrinsic Si

$$n_i = N_c \exp[-(E_c - E_{Fi})/k_B T]$$

whereas for doped Si,

$$n = N_c \exp[-(E_c - E_{Fn})/k_B T] = N_d$$

where E_{Fi} and E_{Fn} are the Fermi energies in the intrinsic and *n*-type Si. Dividing the two expressions,

$$N_d/n_i = \exp[(E_{Fn} - E_{Fi})/k_B T]$$

so that

$$E_{Fn} - E_{Fi} = k_B T \ln(N_d/n_i) = (0.0259 \text{ eV}) \ln(10^{16}/10^{10}) = 0.358 \text{ eV}$$

When the wafer is further doped with boron, the acceptor concentration, $N_a = 2 \times 10^{17} \text{ cm}^{-3} > N_d = 10^{16} \text{ cm}^{-3}$. The semiconductor is compensation doped and compensation converts the semiconductor to a *p*-type Si. $p = N_a - N_d = 2 \times 10^{17} - 10^{16} = 1.9 \times 10^{17} \text{ cm}^{-3}$.

For intrinsic Si

$$p = n_i = N_v \exp[-(E_{Fi} - E_v)/k_B T],$$

whereas for doped Si

$$p = N_v \exp[-(E_{Fp} - E_v)/k_B T] = N_a - N_d$$

where E_{Fi} and E_{Fp} are the Fermi energies in the intrinsic and *p*-type Si, respectively. Dividing the two expressions

$$p/n_i = \exp[-(E_{Fp} - E_{Fi})/k_B T]$$

so that

$$E_{Fp} - E_{Fi} = -k_B T \ln(p/n_i) = -(0.0259 \text{ eV}) \ln(1.9 \times 10^{17}/1.0 \times 10^{10}) = -0.434 \text{ eV}$$

The negative value indicates that E_{Fp} is below E_{Fi} .

EXAMPLE 3.3.2 Conductivity of *n*-Si

Consider a pure intrinsic Si crystal. What would be its intrinsic conductivity at 300 K? What are the electron and hole concentrations in an *n*-type Si crystal that has been doped with 10^{16} cm^{-3} phosphorus (P) donors. What is the conductivity if the drift mobility of electrons is about $1200 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ at this concentration of dopants?

Solution

The intrinsic concentration $n_i = 1 \times 10^{10} \text{ cm}^{-3}$, so that the intrinsic conductivity is

$$\begin{aligned}\sigma &= en_i(\mu_e + \mu_h) = (1.6 \times 10^{10} \text{ C})(1 \times 10^{10} \text{ cm}^{-3})(1450 + 490 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}) \\ &= 3.1 \times 10^{-6} \Omega^{-1} \text{ cm}^{-1} \quad \text{or} \quad 3.1 \times 10^{-4} \Omega^{-1} \text{ m}^{-1}\end{aligned}$$

Consider n -type Si. $N_d = 10^{16} \text{ cm}^{-3} > n_i (= 10^{10} \text{ cm}^{-3})$, the electron concentration $n = N_d = 10^{16} \text{ cm}^{-3}$ and $p = n_i^2/N_d = (10^{10} \text{ cm}^{-3})^2/(10^{16} \text{ cm}^{-3}) = 10^4 \text{ cm}^{-3}$; and negligible compared to n .

The conductivity is

$$\sigma = eN_d\mu_e = (1.6 \times 10^{-19} \text{ C})(1 \times 10^{16} \text{ cm}^{-3})(1200 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}) = 1.92 \Omega^{-1} \text{ cm}^{-1}$$

3.4 DIRECT AND INDIRECT BANDGAP SEMICONDUCTORS: E - k DIAGRAMS

We know from quantum mechanics that when the electron is within an infinite potential energy well of spatial width L , its energy is quantized and given by

$$E_n = \frac{(\hbar k_n)^2}{2m_e}$$

where m_e is the mass of the electron and the electron's wave vector k_n is essentially a quantum number determined by

$$k_n = \frac{n\pi}{L}$$

where $n = 1, 2, 3, \dots$. The energy increases parabolically with the wave vector k_n . We also know that the electron momentum is given by $\hbar k_n$. This description can be used to represent the behavior of electrons in a metal within which their average potential energy can roughly be taken to be zero. In other words, we take $V(x) = 0$ within the metal crystal and $V(x)$ to be large, for example, $V(x) = V_o$ (several electron volts) outside, so that the electron is contained within the metal. This is the **nearly free electron model** of a metal which has been quite successful in interpreting many of the metallic properties. Indeed, we typically calculate the density of states $g(E)$ based on the three-dimensional potential well problem. However, it is quite obvious that this model is too simple because it does not take into account the actual variation of the electron potential energy in the crystal.

The potential energy of the electron depends on its location within the crystal and is periodic due to the regular arrangement of the atoms. How does a periodic potential energy affect the relationship between E and k ? It will no longer be simply $E_n = (\hbar k_n)^2/2m_e$.

To find the energy of the electron in a crystal, we need to solve the Schrödinger equation for a periodic potential energy function in three dimensions. We first consider the hypothetical one-dimensional crystal shown in Figure 3.12. The electron potential energy functions for each atom add to give an overall potential energy function $V(x)$, which is clearly periodic in x with the periodicity of the crystal, a . Thus, $V(x) = V(x + a) = V(x + 2a) = \dots$ and so on. Our task is therefore to solve the Schrödinger equation for the electron wavefunction $\psi(x)$,

$$\frac{d^2\psi}{dx^2} + \frac{2m_e}{\hbar^2} [E - V(x)]\psi = 0 \tag{3.4.1}$$

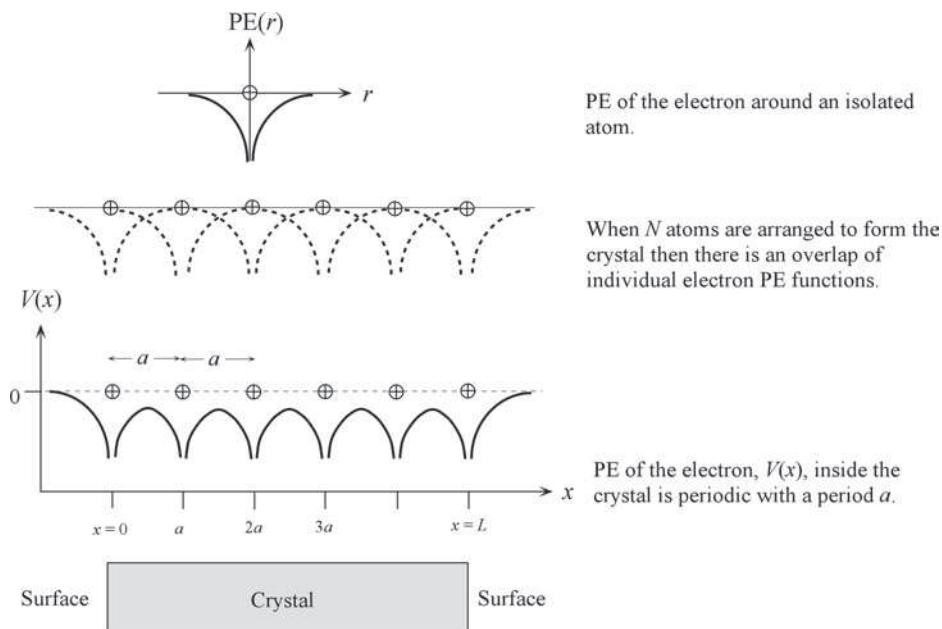


FIGURE 3.12 The electron potential energy (PE), $V(x)$, inside the crystal is periodic with the same periodicity as that of the crystal, a . Far away outside the crystal, by choice, $V = 0$ (the electron is free and PE = 0).

subject to the condition that the potential energy, $V(x)$, is periodic in a , that is,

$$V(x) = V(x + ma); \quad m = 1, 2, 3, \dots \quad (3.4.2)$$

Periodic
PE in the
crystal

The solution of Eq. (3.4.1) will give the electron wavefunction in the crystal and hence the electron energy. Since $V(x)$ is periodic, we should expect, by intuition at least, the solution $\psi(x)$ to be periodic. It turns out that the solutions to Eq. (3.4.1), which are called **Bloch wavefunctions**, are of the form

$$\psi_k(x) = U_k(x) \exp(jkx) \quad (3.4.3)$$

Bloch
wave in
the crystal

where $U_k(x)$ is a periodic function which depends on $V(x)$ and has the same periodicity a as $V(x)$. The term $\exp(jkx)$, of course, represents a traveling wave whose wave vector is k . We should remember that we have to multiply this by $\exp(-jEt/\hbar)$, where E is the energy, to get the overall wavefunction $\Psi(x, t)$. Thus, the electron wavefunction in the crystal is a traveling wave which is modulated by $U_k(x)$. Further, both $\exp(jkx)$ and $\exp(-jkx)$ are possible and represent left and right traveling waves.

There are many such Bloch wavefunction solutions to the one-dimensional crystal, each identified with a particular k value, say k_n , which acts as a kind of quantum number. Each $\psi_k(x)$ solution corresponds to a particular k_n and represents a state with an energy E_k . The dependence of the energy E_k on the wave vector k is illustrated in what is called an **E - k diagram**. Figure 3.13 shows a typical E - k diagram for the hypothetical one-dimensional solid for k values in the range $-\pi/a$ to $+\pi/a$. Just as $\hbar k$ is the momentum of a free electron, $\hbar k$ for the Bloch electron is the momentum involved in its interaction with external fields, for example, those involved in the photon absorption processes. Indeed the rate of change of $\hbar k$ is the externally applied force F_{ext} on

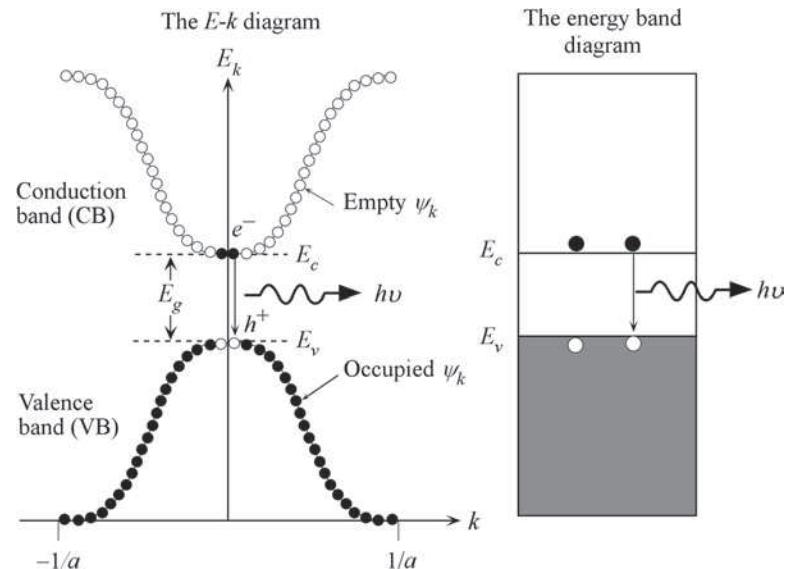


FIGURE 3.13 The E - k diagram of a direct bandgap semiconductor such as GaAs. The E - k curve consists of many discrete points with each point corresponding to a possible state, wavefunction $\psi_k(x)$, that is allowed to exist in the crystal. The points are so close that we normally draw the E - k relationship as a continuous curve. In the energy range E_v to E_c there are no points, that is, no $\psi_k(x)$ solutions.

the electron such as that due to an electric field E (*i.e.*, $F_{\text{ext}} = eE$). Thus, for the electron within the crystal, $d(\hbar k)/dt = F_{\text{ext}}$ and consequently we call $\hbar k$ the **crystal momentum** of the electron.⁴

Inasmuch as the momentum of the electron in the x -direction in the crystal is given by $\hbar k$, the E - k diagram is an energy vs. crystal momentum plot. The states $\psi_k(x)$ in the lower E - k curve constitute the wavefunctions for the valence electrons and thus correspond to the states in the valence band. Those in the upper E - k curve, on the other hand, correspond to the states in the conduction band since they have higher energies. All the valence electrons at 0 K therefore fill the states (particular k_n values) in the lower E - k diagram.

It should be emphasized that an E - k curve in the diagram consists of many discrete points, each point corresponding to a possible state, wavefunction $\psi_k(x)$, that is allowed to exist in the crystal. The points are so close that we draw the E - k relationship as a continuous curve. It is clear from the E - k diagram that there is a range of energies, from E_v to E_c , for which there are no solutions to the Schrödinger equation and hence there are no $\psi_k(x)$ with energies in E_v to E_c . Furthermore, we also note that the E - k behavior is not a simple parabolic relationship except near the bottom of the CB and the top of the VB.

Above absolute zero of temperature, due to thermal excitation, however, some of the electrons from the top of the valence band will be excited to the bottom of the conduction band. According to the E - k diagram in Figure 3.13, when an electron and hole recombine, the electron simply drops from the bottom of the CB to the top of the VB without any change in its k value so that this transition is quite acceptable in terms of momentum conservation. We should recall that the momentum of the emitted photon is negligible compared with the momentum of the electron. The E - k diagram in Figure 3.13 is therefore for a **direct bandgap semiconductor**. The minimum of the CB is directly above the maximum of the VB. As shown in Figure 3.13, the electron transition from the bottom of the CB to the top of the VB emits a photon.

⁴The actual momentum of the electron, however, is not $\hbar k$ because $d(\hbar k)/dt \neq F_{\text{external}} + F_{\text{internal}}$. The true momentum p_e satisfies $dp_e/dt = F_{\text{external}} + F_{\text{internal}}$ (all forces on the electron). However, as we are interested in interactions with external forces such as an applied field, we treat $\hbar k$ as if it were the momentum of the electron in the crystal and use the name *crystal momentum*.

The simple E - k diagram sketched in Figure 3.13 is for the hypothetical one-dimensional crystal in which each atom simply bonds with two neighbors. In real crystals, we have a three-dimensional arrangement of atoms with $V(x, y, z)$ showing periodicity in more than one direction. The E - k curves are then not as simple as that in Figure 3.13 and often show unusual features. The E - k diagram for GaAs, which is shown in Figure 3.14 (a), as it turns out, has general features that are quite similar to that sketched in Figure 3.13. The lowest CB minimum is right above the top of the VB at the same k -value. GaAs is therefore a direct bandgap semiconductor in which electron–hole pairs can recombine directly and emit a photon. The majority of light-emitting devices use direct bandgap semiconductors to make use of direct recombination.

In the case of Si, the diamond crystal structure leads to an E - k diagram which has the essential features illustrated in Figure 3.14 (b). We notice that the minimum of the CB is *not* directly above the maximum of the VB, but it is displaced on the k -axis. Such crystals are called **indirect bandgap semiconductors**. An electron at the bottom of the CB cannot therefore recombine directly with a hole at the top of the VB, because for the electron to fall down to the top of the VB its momentum must change from k_{CB} to k_{VB} , which is not allowed by the law of conservation of momentum. Thus, direct electron–hole recombination does not normally take place in Si and Ge. The recombination process in these elemental semiconductors occurs via a **recombination center** at an energy level E_r , within the bandgap as illustrated in Figure 3.14 (c). These recombination centers may be crystal defects or impurities. The electron is first captured by the defect at E_r . The change in the energy and momentum of the electron by this capture process is transferred to lattice vibrations, that is, to **phonons**. As much as an electromagnetic radiation is quantized in terms of photons, lattice vibrations in the crystal are quantized in terms of phonons. Lattice vibrations travel in the crystal just like a wave and these waves are called phonons. The captured electron at E_r can readily fall down into an empty state at the top of the VB and thereby recombine with a hole as in Figure 3.14 (c). Typically, the electron transition from E_r to E_v involves the emission of further lattice vibrations.

In some indirect bandgap semiconductors such as GaP, however, the recombination of the electron with a hole at certain recombination centers results in photon emission. The E - k diagram is similar to that shown in Figure 3.14 (c) except that the recombination centers at E_r are generated by the purposeful addition of nitrogen (N) impurities to GaP, written as GaP:N. The electron transition from E_r to E_v involves photon emission in the green.

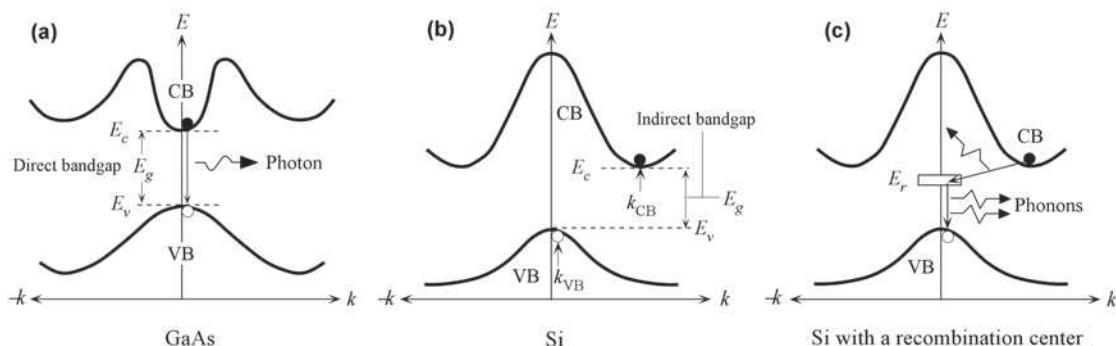


FIGURE 3.14 (a) In GaAs the minimum of the CB is directly above the maximum of the VB. GaAs is therefore a direct bandgap semiconductor. (b) In Si, the minimum of the CB is displaced from the maximum of the VB and Si is an indirect bandgap semiconductor. (c) Recombination of an electron and a hole in Si involves a recombination center.

3.5 *pn* JUNCTION PRINCIPLES

A. Open Circuit

Consider what happens when one side of a sample of Si is doped *n*-type and the other *p*-type as shown in Figure 3.15 (a). Assume that there is an *abrupt* discontinuity between the *p* and *n* regions which we call the **metallurgical junction**, *M*, as indicated in Figure 3.15 (a). There are fixed (immobile) ionized donors and free electrons (in the conduction band, CB) in the *n*-region and fixed ionized acceptors and holes (in the valence band, VB) in the *p*-region.

Due to the hole concentration gradient from the $p = p_{po}$ to the *n*-side where $p = p_{no}$, holes diffuse toward the right and enter the *n*-region and recombine with the electrons (majority carriers)

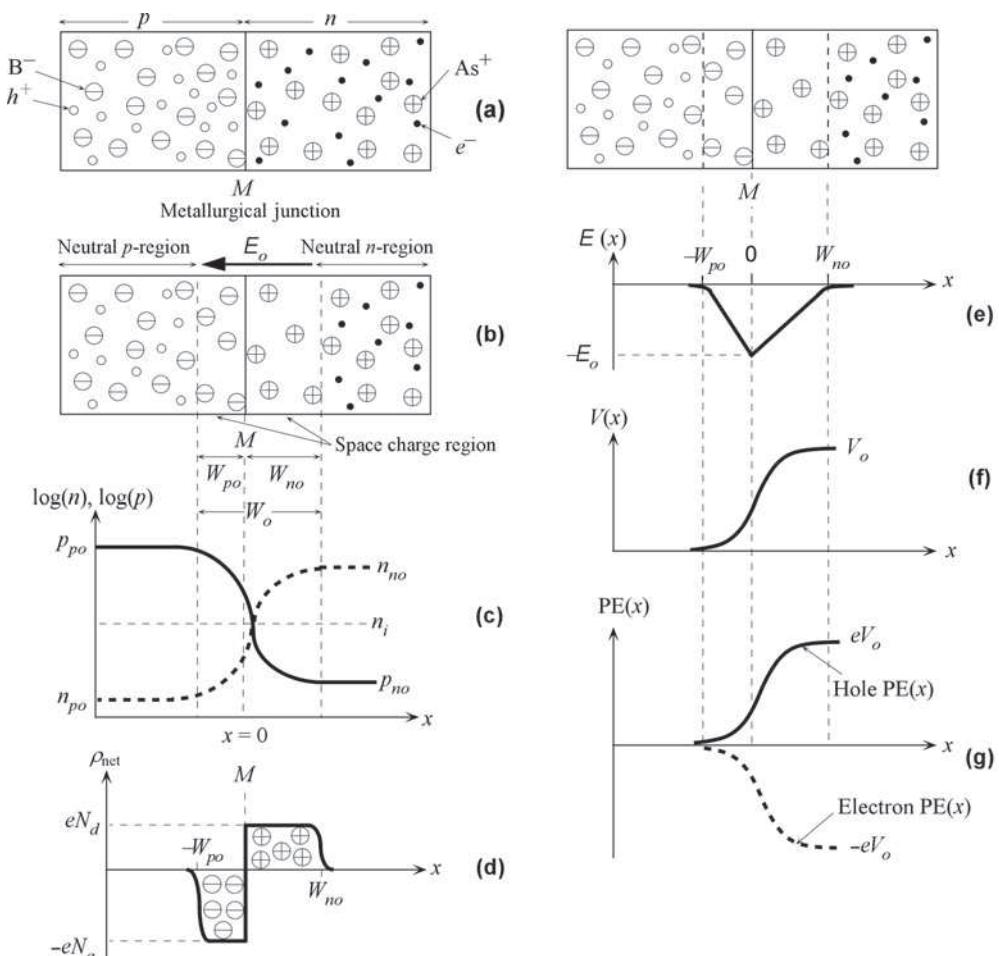


FIGURE 3.15 Properties of the *pn* junction. (a) The *p*- and *n*-sides of the *pn* junction before the contact. (b) The *pn* junction after contact, in equilibrium and in open circuit. (c) Carrier concentrations along the whole device through the *pn* junction. At all points, $n_{po}p_{po} = n_{no}p_{no} = n_i^2$. (d) Net space charge density ρ_{net} across the *pn* junction. (e) The electric field across the *pn* junction is found by integrating ρ_{net} in (d). (f) The potential $V(x)$ across the device. Contact potentials are not shown at the semiconductor–metal contacts. (g) Hole and electron potential energy (PE) across the *pn* junction. Potential energy is charge \times potential $= qV$.

in this region. The *n*-side near the junction therefore becomes depleted of majority carriers and therefore has exposed positive donor ions (As^+) of concentration N_d . Similarly, the electron concentration gradient drives the electrons by diffusion toward the left. Electrons diffusing into the *p*-side recombine with the holes (majority carriers), which exposes negative acceptor ions (B^-) of concentration N_a in this region. The regions on both sides of the junction M consequently become *depleted* of free carriers in comparison with the bulk *p*- and *n*-regions far away from the junction. There is therefore a **space charge layer** (SCL) around M . Figure 3.15 (b) shows the SCL, also known as the **depletion region**, around M . Figure 3.15 (c) illustrates the hole and electron concentration profiles in which the vertical concentration scale is logarithmic. Note that we must, under equilibrium conditions (e.g., no applied bias or photoexcitation), have $pn = n_i^2$ everywhere. The depletion region extends from about $x = -W_{po}$ to $x = W_{no}$. (The subscripts indicate the *p*- or *n*-side and open circuit.). The total width of the depletion region is $W_o = W_{po} + W_{no}$.

It is clear that there is an internal electric field E_o from positive ions to negative ions, that is, in the $-x$ direction, which tries to drift the holes back into the *p*-region and electrons back into the *n*-region. This field drives the holes in the opposite direction to their diffusion. As shown in Figure 3.15 (b), E_o imposes a drift force on holes in the $-x$ direction whereas the hole diffusion flux is in the $+x$ direction. A similar situation also applies for electrons with the electric field attempting to drift the electrons against diffusion from *n*- to the *p*-region. It is apparent that as more and more holes diffuse toward the right, and electrons toward the left, the internal field around M will increase until eventually an *equilibrium* is reached when the rate of holes diffusing toward the right is just balanced by the rates of holes drifting back to the left, driven by the field E_o . The electron diffusion and drift fluxes will also be balanced in equilibrium. For uniformly doped *p*- and *n*-regions, the net space charge density $\rho_{\text{net}}(x)$ across the semiconductor will be as shown in Figure 3.15 (d). The net space charge density ρ_{net} is negative and equal to $-eN_a$ in the SCL from $x = -W_{po}$ to $x = 0$ (M is at $x = 0$) and then positive and equal to $+eN_d$ from $x = 0$ to W_{no} . The total charge on the left-hand side must equal to that on the right-hand side for overall charge neutrality, so that

$$N_a W_{po} = N_d W_{no} \quad (3.5.1)$$

Depletion
region
widths

Figure 3.15 arbitrarily assumes that the acceptor concentration is greater than the donor concentration, $N_a > N_d$. From Eq. (3.5.1) this implies that $W_{no} > W_{po}$, that is, the depletion region penetrates the *n*-side (lightly doped side) more than the *p*-side (heavily doped side). Indeed, if $N_a \gg N_d$, then the depletion region is almost entirely on the *n*-side. We generally indicate heavily doped regions with the superscript plus sign as p^+ . The electric field $E(x)$ and the net space charge density $\rho_{\text{net}}(x)$ at a point are related in electrostatics⁵ by $dE/dx = \rho_{\text{net}}(x)/\epsilon$, where $\epsilon = \epsilon_0 \epsilon_r$, is the permittivity of the medium, and ϵ_0 and ϵ_r are the absolute permittivity and relative permittivity of the semiconductor material, respectively. We can thus integrate $\rho_{\text{net}}(x)$ across the device and thus determine the electric field. The variation of the electric field across the *pn* junction is shown in Figure 3.15 (e). The negative field means that it is in the $-x$ direction. Note that $E(x)$ reaches a maximum magnitude E_o at M .

The potential $V(x)$ at any point x can be found by integrating the electric field since by definition $E = -dV/dx$. Taking the potential on the *p*-side far away from M as zero (we have no applied voltage), which is an arbitrary reference level, $V(x)$ increases in the depletion region

⁵This is called *Gauss's law in point form* and comes from Gauss's law in electrostatics. The integration of the electric field E over a closed surface S is related to the total charge Q_{enclosed} enclosed, $\int E dS = Q_{\text{enclosed}}/\epsilon$, where ϵ is the permittivity of the medium.

toward the *n*-side as indicated in Figure 3.15 (f). Notice that on the *n*-side the potential reaches V_o , which is called the **built-in potential**.

In an abrupt *pn* junction $\rho_{\text{net}}(x)$ can simply and approximately be described by step functions as displayed in Figure 3.15 (d). Using the step form of $\rho_{\text{net}}(x)$ in Figure 3.15 (d) and integrating it gives the electric field E_o and the built-in potential V_o ,

$$\begin{array}{l} \text{Built-in} \\ \text{field} \end{array} \quad E_o = -\frac{eN_d W_{no}}{\epsilon} = -\frac{eN_a W_{po}}{\epsilon} \quad (3.5.2)$$

and

$$\begin{array}{l} \text{Built-in} \\ \text{potential} \end{array} \quad V_o = -\frac{1}{2} E_o W_o = -\frac{eN_a N_d W_o^2}{2\epsilon(N_a + N_d)} \quad (3.5.3)$$

where $\epsilon = \epsilon_o \epsilon_r$ is the permittivity of the semiconductor medium and $W_o = W_{no} + W_{po}$ is the total width of the depletion region under open circuit. If we know W_o , then W_{no} or W_{po} follows readily from Eq. (3.5.1). Equation (3.5.3) is a relationship between the built-in voltage V_o and the depletion region width W_o . If we know V_o we can calculate W_o .

The simplest way to relate V_o to the doping parameters is through Boltzmann statistics. For the system consisting of *p*- and *n*-type semiconductors together (forming one system), in equilibrium, the Boltzmann statistics⁶ demands that the concentrations n_1 and n_2 of carriers at potential energies E_1 and E_2 be related by

$$\begin{array}{l} \text{Boltzmann} \\ \text{statistics} \end{array} \quad n_2/n_1 = \exp[-(E_1 - E_2)/k_B T] \quad (3.5.4)$$

in which E is the potential energy—that is, qV , where q is charge and V voltage. Considering electrons, $q = -e$, we see from Figure 3.15 (g) that $E = 0$ on the *p*-side far away from M where $n = n_{po}$; and $E = -eV_o$ on the *n*-side away from M where $n = n_{no}$. Thus,

$$n_{po}/n_{no} = \exp(-eV_o/k_B T) \quad (3.5.5a)$$

Equation (3.5.5a) shows that V_o depends on n_{no} and n_{po} and hence on N_d and N_a . The corresponding equation for hole concentrations is

$$p_{no}/p_{po} = \exp(-eV_o/k_B T) \quad (3.5.5b)$$

Thus, rearranging Eqs. (3.5.5a) and (3.5.5b) we obtain

$$V_o = (k_B T/e) \ln(n_{no}/n_{po}) \quad \text{and} \quad V_o = (k_B T/e) \ln(p_{po}/p_{no})$$

We can now write p_{po} and p_{no} in terms of the dopant concentrations inasmuch as $p_{po} = N_a$, $p_{no} = n_i^2/n_{no} = n_i^2/N_d$, so that V_o becomes

$$\begin{array}{l} \text{Built-in} \\ \text{potential} \end{array} \quad V_o = \frac{k_B T}{e} \ln\left(\frac{N_a N_d}{n_i^2}\right) \quad (3.5.6)$$

⁶We use Boltzmann statistics, *i.e.*, $n(E) \propto \exp(-E/k_B T)$, because the concentration of electrons in the conduction band whether on the *n*- or *p*-side is never so large that the Pauli exclusion principle becomes important. As long as the carrier concentration in the conduction band is much smaller than N_c , we can use Boltzmann statistics.

Clearly V_o has been conveniently related to the dopant and parent material properties via N_a , N_d , and n_i^2 , which is given by $(N_c N_v) \exp(-E_g/k_B T)$. The built-in voltage (V_o) is the potential across a *pn* junction, going from *p*- to *n*-type semiconductor, in an open circuit. V_o is not the actual voltage across the diode. The voltage across the diode is made up of V_o and the contact potentials at the metal-to-semiconductor junctions at the electrodes. If we add V_o and the contact potentials at the electrode ends, we will find zero. Once we know the built-in potential V_o from Eq. (3.5.6), we can then calculate the width W_o of the depletion region from Eq. (3.5.3). The term $k_B T/e$ in Eq. (3.5.6) is called the **thermal voltage** V_{th} , and is about 26 mV at 300 K.

B. Forward Bias and the Shockley Diode Equation

Consider what happens when a battery with a voltage V is connected across a *pn* junction so that the positive terminal of the battery is attached to the *p*-side and the negative terminal to the *n*-side (forward bias). The negative polarity of the supply will reduce the potential barrier V_o by V , as shown in Figure 3.16 (a). The reason is that the bulk regions outside the SCL have high conductivities, due to the plenty of majority carriers in the bulk, in comparison with the depletion region in which there are mainly immobile ions. Thus, the applied voltage drops mostly across the depletion width W . The applied bias V now directly opposes V_o . The potential barrier against diffusion therefore becomes reduced to $(V_o - V)$ as illustrated in Figure 3.16 (b). This has drastic consequences because the probability that a hole in the *p*-side will surmount this potential barrier and diffuse to the *n*-side now becomes proportional to $\exp[-e(V_o - V)/k_B T]$. In other words, the applied voltage effectively reduces the built-in potential and hence the built-in field which acts against diffusion. Consequently, many holes can now diffuse across the depletion region and enter the *n*-side. This results in the **injection of excess minority carriers**, that is, holes into the *n*-region. Similarly, excess electrons can now diffuse toward the *p*-side and enter this region and thereby become injected minority carriers.

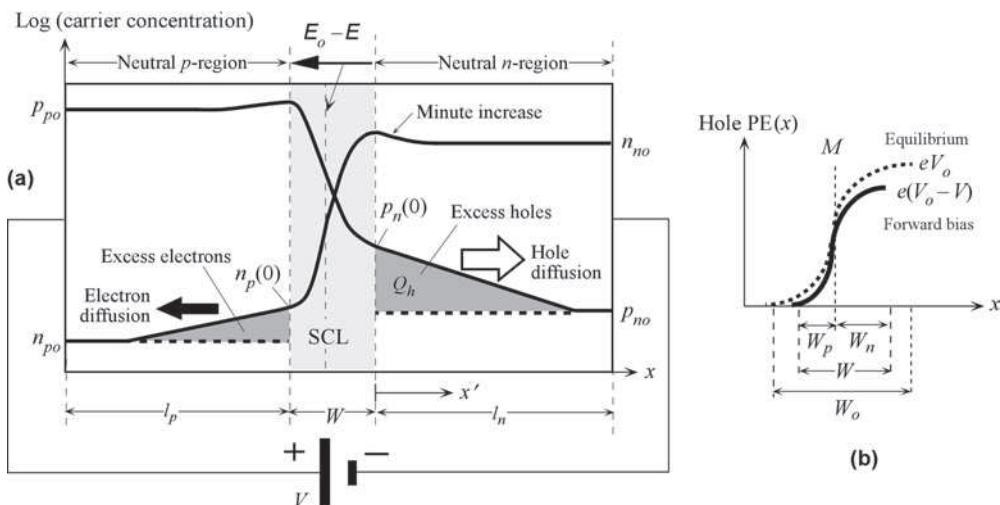


FIGURE 3.16 Forward-biased *pn* junction and the injection of minority carriers (a) Carrier concentration profiles across the device under forward bias. (b) The hole potential energy with and without an applied bias. W is the width of the SCL with forward bias.

The hole concentration $p_n(0)$ just outside the depletion region at $x' = 0$ (x' is measured from W_n) in the n -side is very large due to the injection of minority carriers as shown in Figure 3.16 (a). The hole concentration at the end of the neutral n -side at $x' = l_n$ at the negative terminal is always the equilibrium concentration p_{no} , because any additional holes at $x' = l_n$ would be neutralized immediately by electrons flowing from the negative terminal of the battery. It is obvious that there is a very large hole concentration gradient in the n -side as a result of minority carrier injection under forward bias, as shown in Figure 3.16 (a). The hole concentration gradient causes hole diffusion flux toward the negative terminal, and hence gives rise to an electric current. There is a similar electron diffusion flux and hence an electric current in the p -side. These two minority carrier diffusion fluxes constitute the forward bias pn junction current.

When holes are injected into the neutral n -side, they draw some electrons from the bulk of n -side (and hence from the battery) so that there is a small increase in the electron concentration. This small increase in the majority carriers is necessary to balance the hole charges and maintain neutrality in the n -side.

We now derive the pn junction equation. Consider the hole concentration profile in the n -side. The hole concentration, $p_n(0) = p_n(x' = 0)$, just outside the depletion region is due to the excess of holes diffusing through the SCL as a result of the reduction in the built-in potential barrier. This concentration, $p_n(0)$, is determined by the probability of surmounting the new potential energy barrier $e(V_o - V)$,

Boltzmann statistics and applied voltage

$$p_n(0) = p_{po} \exp\left[-\frac{e(V_o - V)}{k_B T}\right] \quad (3.5.7)$$

This follows directly from the Boltzmann equation, by virtue of the hole potential energy changing by $e(V_o - V)$ from $x = -W_p$ to $x = W_n$, as indicated in Figure 3.16 (b), and at the same time the hole concentration falling from p_{po} to $p_n(0)$. By dividing Eq. (3.5.7) by Eq. (3.5.5b), we get the effect of the applied voltage out directly, which shows how the voltage V determines the amount of *excess* holes diffusing and arriving at the n -region,

Law of the junction

$$p_n(0) = p_{no} \exp\left(\frac{eV}{k_B T}\right) \quad (3.5.8)$$

which is called the **law of the junction**. Equation (3.5.8) describes the effect of the applied voltage V on the injected minority carrier concentration just outside the depletion region, $p_n(0)$. Obviously, with no applied voltage, $V = 0$ and $p_n(0) = p_{no}$ which is exactly what we expect, Eq. (3.5.5b).

Injected holes diffuse in the n -region and eventually recombine with electrons in this region; there are many electrons in the n -side. Those electrons lost by recombination are readily replenished by the negative terminal of the battery connected to this side. The current due to holes diffusing in the n -region can be sustained because more holes can be supplied by the p -region, which itself can be replenished by the positive terminal of the battery.

Electrons are similarly injected from the n -side to the p -side. The electron concentration $n_p(0)$ just outside the depletion region at $x = -W_p$ is given by the equivalent of Eq. (3.5.8) for electrons, that is

Law of the junction

$$n_p(0) = n_{po} \exp\left(\frac{eV}{k_B T}\right) \quad (3.5.9)$$

In the *p*-region, the injected electrons diffuse toward the positive terminal looking to be collected. As they diffuse, they recombine with some of the many holes in this region. Those holes lost by recombination can be readily replenished by the positive terminal of the battery connected to this side. The current due to the diffusion of electrons in the *p*-side can be maintained by the supply of electrons from the *n*-side, which itself can be replenished by the negative terminal of the battery. It is apparent that an electric current can be maintained through a *pn* junction under forward bias, and that the current flow seems to be due to the *diffusion of minority carriers*. There is, however, some drift of majority carriers as well. The semiconductor *pn* junction is a minority carrier device in the sense that minority carriers play an important role.

When a hole (minority carrier) is injected into the *n*-side, it will diffuse in this region inasmuch as there is very little electric field in the neutral *n*-region to give rise to drift. Eventually the hole will recombine with an electron (majority carrier). The average time it takes for a hole (minority carrier) to recombine with an electron in the *n*-side is called the **minority carrier recombination time** or **lifetime** τ_h . The reciprocal $1/\tau_h$ is the mean probability per unit time that a hole will recombine and disappear. The ability of holes to diffuse in a semiconductor is represented by their **diffusion coefficient** D_h , which is related to their drift mobility μ_h through the Einstein relation.⁷ An average distance a hole diffuses in the *n*-side before it recombines with an electron is called the **diffusion length** L_h and is given by $L_h = (D_h \tau_h)^{1/2}$. We can interpret the reciprocal $1/L_h$ in a similar way to $1/\tau_h$ as the probability per unit distance that a hole recombines with an electron in the *n*-side as it diffuses along x .

If the length of the *n*-region l_n is much longer than the minority carrier diffusion length L_h , most (nearly all) injected holes will eventually recombine before they reach the negative electrode. The hole concentration profile $p_n(x')$ vs. x' therefore decays from $p_n(0)$ toward the thermal equilibrium value, p_{no} , as illustrated in Figure 3.16 (a). Consider $\Delta p_n(x') = p_n(x') - p_{no}$, which is the **excess carrier concentration**, then the change $\delta\Delta p_n$ in excess concentration over a distance δx due to recombination would depend on Δp_n itself (how many excess holes we have) and also on the probability of recombination ($\delta x'/(L_h)$) so that $\delta\Delta p_n = -\Delta p_n(\delta x'/L_h)$, which implies an exponential decay,

$$\Delta p_n(x') = \Delta p_n(0) \exp(-x'/L_h) \quad (3.5.10)$$

Equation (3.5.10) can be derived by rigorous means using the Continuity Equation⁸ rather than the above intuitive arguments. The *pn* junction in which $l_n \gg L_h$ is called a **long diode**.

The hole diffusion current density $J_{D,\text{hole}}$ is the *hole diffusion flux* multiplied by the hole charge,⁹

$$J_{D,\text{hole}} = -eD_h \frac{dp_n(x')}{dx'} = -eD_h \frac{d\Delta p_n(x')}{dx'}$$

Long
diode
minority
carrier
profile

⁷In nondegenerate semiconductors, the diffusion coefficient D and the drift mobility μ for a given type of carrier are related by $D/\mu = kT/e$ so that we can easily calculate D from μ . The proof is not difficult; see, for example, S. O. Kasap, *Principles of Electronic Materials and Devices*, 3rd Edition (McGraw-Hill, 2006), Ch. 5.

⁸The intuitive derivation here for Eq. (3.5.10) was actually based on knowing the outcome from the continuity equation and using the fact that we have a very long *n*-side. In the latter case *only*, the oversimplified intuitive arguments above do indeed lead to Eq. (3.5.10); but not in a general treatment. One must also consider the change in the diffusion flux (the current). (See, for example, S. O. Kasap, *Principles of Electronic Materials and Devices*, 3rd Edition, McGraw-Hill, 2006), Ch. 6.

⁹The hole diffusion flux is $-D_h(dp/dx) = -D_h(d\Delta p/dx)$ and the diffusing charge is $+e$.

i.e.,

$$J_{D,\text{hole}} = \left(\frac{eD_h}{L_h} \right) \Delta p_n(0) \exp\left(-\frac{x'}{L_h}\right) \quad (3.5.11)$$

The above equation shows that hole diffusion current depends on location x' and decreases due to recombination. The total current at any location, however, is the sum of hole and electron contributions. The total current is independent of x' as indicated in Figure 3.17. The decrease in the minority carrier diffusion current with x' is made up by the increase in the current due to the drift of the majority carriers as schematically shown in Figure 3.17. The field in the neutral region is not totally zero but a small value, just sufficient to drift the huge number of majority carriers there and maintain a constant current.

We can now use the law of the junction to substitute for $\Delta p_n(0)$ in Eq. (3.5.11) in terms of the applied voltage V in Eq. (3.5.8). Further, we can eliminate p_{no} by $p_{no} = n_i^2/n_{no} = n_i^2/N_d$. Thus, at $x' = 0$, just outside the depletion region, from Eq. (3.5.11) the hole diffusion current density is

Hole diffusion current

$$J_{D,\text{hole}} = \left(\frac{eD_h n_i^2}{L_h N_d} \right) \left[\exp\left(\frac{eV}{k_B T}\right) - 1 \right] \quad (3.5.12)$$

There is a similar expression for the electron diffusion current density $J_{D,\text{elec}}$ in the p -region. We will assume that the electron and hole currents do not change across the depletion region because, in general, the width of this region is quite narrow (and, for the time being, we neglect the recombination in the SCL). The electron current at $x = -W_p$ is the same as that at $x = W_n$. The total current density is then simply given by $J_{D,\text{hole}} + J_{D,\text{elec}}$, that is,

Shockley long diode equation

$$J = \left(\frac{eD_h}{L_h N_d} + \frac{eD_e}{L_e N_a} \right) n_i^2 \left[\exp\left(\frac{eV}{k_B T}\right) - 1 \right] \quad (3.5.13a)$$

or

Shockley long diode equation

$$J = J_{so} \left[\exp\left(\frac{eV}{k_B T}\right) - 1 \right] \quad (3.5.13b)$$

This is the familiar diode equation with $J_{so} = [(eD_h/L_h N_d) + (eD_e/L_e N_a)] n_i^2$. It is frequently called the **Shockley equation**. It represents the *diffusion of minority carriers* in the neutral regions. The constant J_{so} depends not only on the doping, N_d , N_a , but also on the material

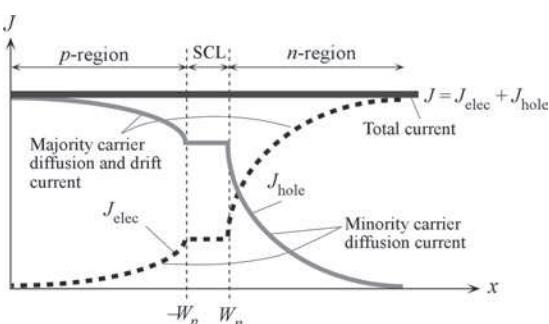


FIGURE 3.17 The total current anywhere in the device is constant. Just outside the depletion region it is primarily due to the diffusion of minority carriers and near the contacts it is primarily due to the drift of majority carriers.

via n_i , D_h , D_e , L_h , L_e . It is known as the **reverse saturation current density**, because if we apply a reverse bias $V = -V_r$ greater than the thermal voltage $k_B T/e$ (25 mV), Eq. (3.5.13b) becomes $J = -J_{so}$.

We note that since the applied voltage drops across the depletion region, the built-in field is reduced by the applied field imposed by the bias, that is, the field at the junction becomes $E_o - E$, where E is the applied field at the junction. Because the field is reduced, we need less charges on either side of the junction. This can only be accommodated if the depletion regions shrink in width to W_p and W_n on the *p*- and *n*-sides, and the new total width $W = W_p + W_n$ is less than W_o . Equation (3.5.1) remains valid in the form $N_a W_p = N_d W_n$.

The *pn* junction in which the neutral regions l_n and l_p are shorter than the diffusion lengths L_h and L_n , respectively, is called a **short diode**. Its treatment is much easier than the long diode. There is essentially no time for the injected hole (minority carrier) in the *n*-side to recombine because it quickly diffuses and reaches the collection (negative) electrode, which is very close in distance. If we are not losing holes by recombination, then there is no reason for the hole diffusion current to decay along x' in this region as in the long diode where recombination is present. The hole diffusion current would remain constant if the hole concentration gradient is constant, that is, the excess hole concentration profile decreases linearly as shown in Figure 3.18. Thus, the hole diffusion current density $J_{D,\text{hole}}$ on the *n*-side is

$$J_{D,\text{hole}} = -eD_h \frac{d\Delta p_n(x')}{dx'} = eD_h \frac{p_n(0) - p_{no}}{l_n} = eD_h \frac{p_{no}e^{eV/k_B T} - p_{no}}{l_n}$$

so that

$$J_{D,\text{hole}} = \frac{eD_h n_i^2}{l_n N_d} \left[\exp\left(\frac{eV}{k_B T}\right) - 1 \right]$$

A similar argument applies for $J_{D,\text{elec}}$ on the *p*-side so that the total current density is then

$$J = \left(\frac{eD_h}{l_n N_d} + \frac{eD_e}{l_p N_a} \right) n_i^2 \left[\exp\left(\frac{eV}{k_B T}\right) - 1 \right] \quad (3.5.14)$$

Shockley
short
diode
equation

Equation (3.5.14) is identical to Eq. (3.5.13a) with diffusion lengths L_h and L_e replaced by the widths of the neutral regions l_n and l_p , respectively. The short diode is a good example of a

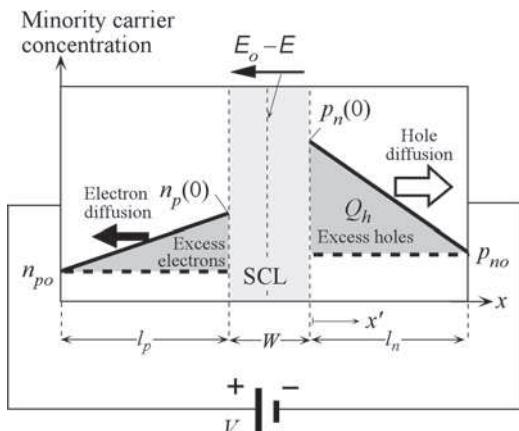


FIGURE 3.18 Minority carrier injection and diffusion in a short diode. Minority recombination in the neutral regions is negligible. $p_n(0)$ is the minority carrier concentration just outside the depletion region, and is controlled by the law of the junction, $p_n(0) = p_{no} \exp(eV/k_B T)$.

device in which the current is almost entirely due to the *diffusion of minority carriers*, and highlights the significant role played by diffusion.

C. Minority Carrier Charge Stored in Forward Bias

So far we have considered a *pn* junction with heavier doping on the *p*-side. The injected minority carriers, holes in this case, represent an *injected excess minority carrier charge*, Q_h , in the neutral region as shown in Figures 3.16 (a) and 3.18. There is, of course, also excess majority carrier charge so the region is neutral; but the excess majority carrier concentration (about the same as $p_n(0)$) is very small compared with the actual majority carrier concentration ($n_{no} = N_d$). For the long diode, the excess holes in the neutral *n*-side diffuse and recombine after τ_h seconds—the minority carrier lifetime. The charge Q_h should therefore disappear after a time τ_h . It must therefore be replenished at a rate Q_h/τ_h if Q_h is to be maintained at steady state DC operation. The external current supplies holes at exactly this rate to maintain the steady state DC conduction. Thus, $I = Q_h/\tau_h$.

In the case of a short diode, a hole that is injected into the *n*-side must diffuse across the width l_n of this neutral region to reach the negative electrode. The time required for holes to diffuse across the width l_n of the neutral *n*-region is called the **transit time** τ_t , or the **diffusion time**. From random walk examples in various materials textbooks,¹⁰ τ_t is given by $\tau_t = l_n^2/2D_h$. To maintain DC operating conditions, the current must replenish holes exactly at a rate Q_h/τ_t to make up for those holes reaching the end of the *n*-side, that is, $I = Q_h/\tau_t$. We can now generalize the above discussion as follows:

DC currents

$$I = Q/\tau \text{ for a long diode} \quad \text{and} \quad I = Q/\tau_t \text{ for a short diode} \quad (3.5.15)$$

where τ is the **effective minority carrier lifetime**, $1/\tau = 1/\tau_h + 1/\tau_e$, and τ_t is the **effective transit time**, that is, $1/\tau_t = 1/\tau_{th} + 1/\tau_{te}$; subscripts refer to holes and electrons in the neutral *n*- and *p*-sides, respectively. Equation (3.5.15) applies under DC conditions only.

D. Recombination Current and the Total Current

So far we have assumed that, under a forward bias, the minority carriers diffusing and recombining in the neutral regions are supplied by the external current. However, some of the minority carriers will recombine in the depletion region. The external current must therefore also supply the carriers lost in the recombination process in the SCL, as illustrated in Figure 3.19 (a). Consider for simplicity a symmetrical *pn* junction as in Figure 3.19 (b) under forward bias. At the junction center *C*, the hole and electron concentrations are p_M and n_M and are equal. We can find the SCL recombination current by considering electrons recombining in the *p*-side in W_p and holes recombining in the *n*-side in W_n as shown by the shaded areas *ABC* and *BCD*, respectively, in Figure 3.19 (b). Suppose that the **mean hole recombination time** in W_n is τ_h and **mean electron recombination time** in W_p is τ_e . The rate at which the electrons in *ABC* are recombining is the area *ABC* (nearly all injected electrons) divided by τ_e . The electrons are replenished by the diode current. Similarly, the rate at which holes in *BCD* are recombining is the area *BCD* divided by τ_h . Thus, the recombination current density is

$$J_{\text{recom}} = \frac{eABC}{\tau_e} + \frac{eBCD}{\tau_h}$$

¹⁰The expression for the mean square distance (l^2) diffused in a time interval τ , $l^2 = 2D\tau$, is derived in S. O. Kasap, *Principles of Electronic Materials and Devices*, 3rd Edition (McGraw-Hill, 2006), Ch. 1.

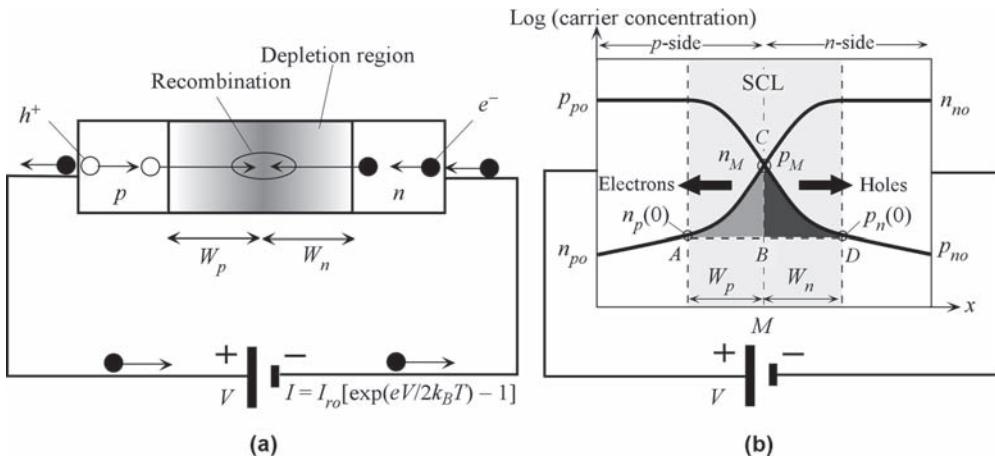


FIGURE 3.19 Forward-biased *pn* junction, the injection of carriers, and their recombination in the SCL. (a) Recombination of electrons and holes in the depletion region involves the current supplying the carriers. (b) A symmetrical *pn* junction for calculating the recombination current.

We can evaluate the areas ABC and BCD by taking them as triangles, $ABC \approx (1/2)W_p n M$, etc., so that

$$J_{\text{recom}} = \frac{e_2^1 W_p n_M}{\tau_e} + \frac{e_2^1 W_n p_M}{\tau_h} \quad (3.5.16)$$

Under steady state and equilibrium conditions, assuming a nondegenerate semiconductor, we can use Boltzmann statistics in Eq. (3.5.4) to relate these concentrations to the potential energy. At A , the hole potential energy is zero ($V = 0$) and at M it is $e(V_o - V)/2$ so that

$$\frac{p_M}{p_{po}} = \exp\left[-\frac{e(V_o - V)}{2k_B T}\right]$$

Since V_o depends on dopant concentrations and n_i as in Eq. (3.5.6) and further $p_{po} = N_a$, we can simplify the above to

$$p_M = n_i \exp\left(\frac{eV}{2k_B T}\right)$$

and there is a similar expression for n_M . We can substitute for p_M and n_M in Eq. (3.5.16) to find the recombination current for $V > k_B T/e$, that is,

$$J_{\text{recom}} = \frac{en_i}{2} \left(\frac{W_p}{\tau_e} + \frac{W_n}{\tau_h} \right) \exp\left(\frac{eV}{2k_B T}\right) \quad (3.5.17a)$$

Recombination current

From a better mathematical treatment, the expression for the recombination current can be shown to be¹¹

$$J_{\text{recom}} = J_{ro} \left[\exp(eV/2k_B T) - 1 \right] \quad (3.5.17b)$$

where J_{ro} is the pre-exponential constant in Eq. (3.5.17a).

SCL
recombi-
*na-
tion current*

¹¹This is generally proved in advanced texts.

Equation (3.5.17) is the current that supplies the carriers that recombine in the depletion region. It does not stipulate what the recombination process is, but only the carriers recombine in this region. The recombination could be direct (as in high quality GaAs *pn* junctions) or indirect through defects or impurities in the SCL.

In many III–V *compound* semiconductors, the carriers also recombine at crystal surfaces and interfaces where, usually, there are defects that encourage recombination; or act as recombination centers. The external current also has to supply carriers that are lost through surface recombination. The exact I – V dependence for surface recombination is difficult to derive but has a voltage dependence quite similar to recombination in the SCL in Eq. (3.5.17b). Thus, surface recombination results in a current that can be described by

*Surface
recombi-
nation
current*

$$I_{\text{surf}} = I_{sro} \left[\exp(eV/2k_B T) - 1 \right] \quad (3.5.18)$$

where I_{sro} is the pre-exponential constant that depends on the surface recombination process, and the surface area involved in the recombination.

The total current into the diode will supply carriers for minority carrier diffusion in the neutral regions and recombination in the space charge layer and at the surfaces and interfaces so that it will be the sum of Eqs. (3.5.13b), (3.5.17b), and (3.5.18). Generally the diode current is written as

*General
diode
equation*

$$I = I_o \left[\exp\left(\frac{eV}{\eta k_B T}\right) - 1 \right] \quad (3.5.19)$$

where I_o is a constant and η , called the **diode ideality factor**, is 1 for diffusion-controlled and 2 for recombination-controlled forward current characteristics.

Figure 3.20 shows typical semilogarithmic plots of the forward I – V characteristics of three types of *pn* junctions based on Ge, Si, and GaAs. The Ge *pn* junction follows the Shockley model with $\eta = 1$ and the current is due to diffusion in the neutral regions. Both Si and GaAs diodes initially follow the recombination model and then the Shockley model. Many Si and GaAs diodes have an η that is between 1 and 2 for the I – V ranges of interest.

Notice that, at a given current, say 1 mA, the voltage across a Ge diode is about 0.2 V, it is 0.7 V across the Si diode, and almost 1.1 V across the GaAs diode. The increase in the turn-on

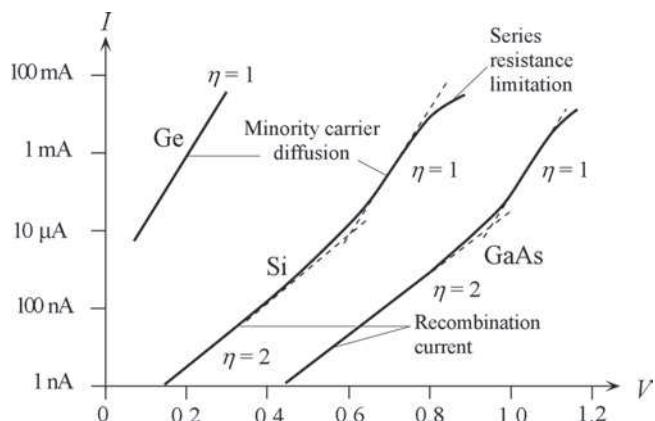


FIGURE 3.20 Schematic representation of the forward I – V characteristics of three typical *pn* junctions based on Ge, Si, and GaAs.

voltage that results in a significant current from 0.2 V to 1.1 V follows the trend for the increase in the bandgap E_g from Ge (0.66 eV) to Si (1.12 eV) to GaAs (1.42 eV). The pre-exponential factor I_o in the diode current equation contains n_i , which depends exponentially on the bandgap E_g . As E_g increases, n_i falls sharply by Eq. (3.2.6), and results in very small I_o values. Thus, a larger applied voltage is needed to bring I_o up to any significant current.

It is clear that, under a forward bias, the *p*-side acts to inject holes into the *n*-side, and the *n*-side acts to inject electrons into the *p*-side. The latter concept will become useful in understanding the operation of heterostructures discussed below.

3.6 *pn* JUNCTION REVERSE CURRENT

When a *pn* junction is reverse biased, the reverse current is typically very small. The reverse bias across a *pn* junction is illustrated in Figure 3.21 (a). The applied voltage drops mainly across the resistive depletion region, which becomes wider. The applied field is in the same direction as the built-in field and adds to it, so that the field at the junction becomes $E_o + E$, and large, where E is the field imposed by the applied forward bias. There is therefore a larger electric field inside the depletion region due to the reverse bias. To accommodate this larger field, the widths of the depletion region on the *p*- and the *n*-side widen to expose more ionized dopants as indicated in Figure 3.21 (a). The movement of electrons in the *n*-region toward the positive battery terminal cannot be sustained because there is no electron supply to this *n*-side. The *p*-side cannot supply electrons to the *n*-side because it has almost none. However, there is a small reverse current due to two causes.

The applied voltage increases the built-in potential barrier, as illustrated in Figure 3.21 (b). The electric field in the SCL is larger than the built-in internal field E_o . The small number of holes on the *n*-side near the depletion region become extracted and swept by the field across the SCL over to the *p*-side. This small current can be maintained by the diffusion of holes from the *n*-side bulk to the SCL boundary. Assume that the reverse bias $V_r > 25 \text{ mV} = k_B T/e$. The hole concentration $p_n(0)$ just outside the SCL is nearly zero by the law of the junction, Eq. (3.5.8), whereas the hole concentration in the bulk (or near the negative terminal) is the equilibrium

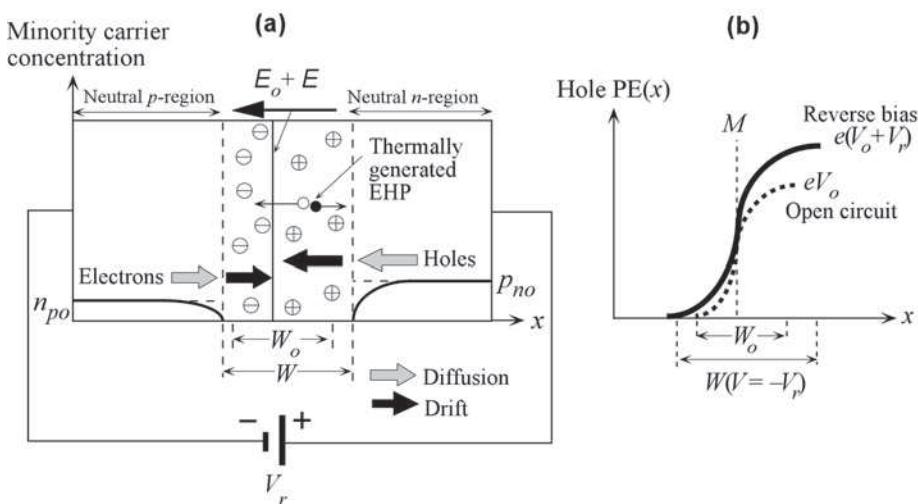


FIGURE 3.21 Reverse-biased *pn* junction. (a) Minority carrier profiles and the origin of the reverse current. (b) Hole PE across the junction under reverse bias.

concentration p_{no} , which is small. There is therefore a small concentration gradient and hence a small hole diffusion current toward the SCL as shown in Figure 3.21 (a). Similarly, there is a small electron diffusion current from the bulk p -side to the SCL. Within the SCL, these carriers are drifted by the field. This minority carrier diffusion is accounted in the Shockley model with $V = -V_r$. The reverse current is given by Eq. (3.5.13) with a negative voltage which leads to a diode current density of $-J_{so}$ called the **reverse saturation current density**. The value of J_{so} depends only on the material via n_i, μ_h, μ_e , the dopant concentrations, etc., but not on the voltage ($V_r > k_B T/e$). Furthermore, as J_{so} depends on n_i^2 , it is strongly temperature dependent. In some books, it is stated that the thermal generation of minority carriers in the neutral region within a diffusion length to the SCL, the diffusion of these carriers to the SCL, and their subsequent drift through the SCL is the cause of the reverse current. This description, in essence, is identical to the Shockley model, and the description above.

The thermal generation of electron–hole pairs in the SCL, as shown in Figure 3.21 (a), can also contribute to the observed reverse current since the internal field in this layer will separate the electron and hole, and drift them toward the neutral regions. This drift will result in an external current in addition to the reverse current due to the diffusion of minority carriers. The theoretical evaluation of SCL generation current involves an in-depth knowledge of the charge carrier generation processes via recombination centers, which is discussed in advanced texts. Suppose that τ_g is the **mean time to generate an electron–hole pair** by virtue of thermal vibrations of the lattice; τ_g is also called the **mean thermal generation time**. Given τ_g , the rate of thermal generation per unit volume must be n_i/τ_g because it takes on average τ_g seconds to create n_i number of EHPs per unit volume. Furthermore, since WA , where A is the cross-sectional area, is the volume of the depletion region, the rate of EHP, or charge carrier, generation is $(AWn_i)/\tau_g$. Both holes and electrons drift in the depletion region and both contribute to the current. The observed current density must be $e(Wn_i)/\tau_g$. Therefore, the reverse current density component due to thermal generation of electron–hole pairs within the SCL should be given by

EHP thermal generation in SCL

$$J_{\text{gen}} = \frac{eWn_i}{\tau_g} \quad (3.6.1)$$

The total reverse current density J_{rev} is the sum of the diffusion and generation components, that is,

Total reverse current

$$J_{\text{rev}} = \left(\frac{eD_h}{L_h N_d} + \frac{eD_e}{L_e N_a} \right) n_i^2 + \frac{eWn_i}{\tau_g} \quad (3.6.2)$$

which is shown schematically in Figure 3.22 (a). The thermal generation component J_{gen} in Eq. (3.6.1) increases with the reverse bias V_r because the SCL width W widens with V_r .

The terms in the reverse current in Eq. (3.6.2) are predominantly controlled by n_i^2 and n_i . Their relative importance depends not only on the semiconductor properties but also on the temperature since $n_i \propto \exp(-E_g/2k_B T)$. Figure 3.22 (b) shows the reverse current I_{rev} in the dark in a Ge pn junction plotted as $\ln(I_{\text{rev}})$ vs. $1/T$ to highlight the two different processes in Eq. (3.6.2). The measurements in Figure 3.22 (b) show that above 238 K, I_{rev} is controlled by n_i^2 because the slope of $\ln(I_{\text{rev}})$ vs. $1/T$ yields an E_g of approximately 0.63 eV, close to the expected E_g of about 0.66 eV in Ge. Below 238 K, I_{rev} is controlled by n_i because the slope of $\ln(I_{\text{rev}})$ vs. $1/T$ is

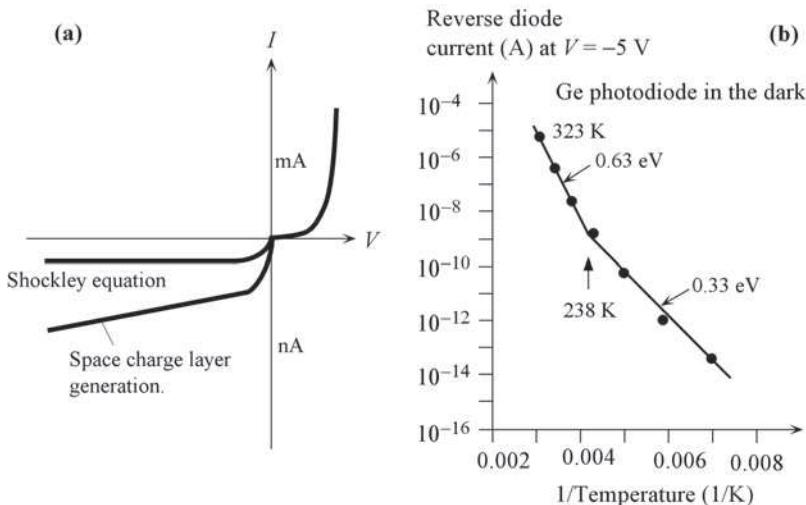


FIGURE 3.22 (a) Schematic illustration of the reverse current of a pn junction and the two contributing factors: diffusion and thermal generation. (The I -axis changes unit for forward and reverse currents.) (b) Reverse diode current in a Ge pn junction as a function of temperature in a $\ln(I_{rev})$ vs. $1/T$ plot. Above 238 K, I_{rev} is controlled by n_i^2 , and below 238 K it is controlled by n_i . The vertical axis is a logarithmic scale with actual current values. (Source: Data plotted from D. Scansen and S. O. Kasap, *Cnd. J. Physics.*, 70, 1070, 1992.)

equivalent to an $E_g/2$ of approximately 0.33 eV. In this range, the reverse current is due to EHP generation in the SCL via defects and impurities (recombination centers).

3.7 ***pn* JUNCTION DYNAMIC RESISTANCE AND CAPACITANCES**

A. Depletion Layer Capacitance

It is apparent that the depletion region or space charge layer of a pn junction has positive charges in W_n and negative charges in W_p , which have been separated from each other, similar to a parallel plate capacitor as indicated in Figures 3.23 (a) and 3.15 (d). We also know that an applied voltage to the pn junction modifies the width W of the SCL; it increases with the reverse bias. If A is the cross-sectional area, the stored charge in the depletion region is $+Q = eN_d W_n A$ on the n -side and $-Q = -eN_a W_p A$ on the p -side. Unlike in the case of a parallel plate capacitor, Q does not depend linearly on the voltage V across the device.

It is useful to define an **incremental capacitance** which relates the incremental change in the charge stored in the depletion region to an incremental voltage change across the pn junction. When the voltage V across a pn junction changes by dV to $V + dV$, then W also changes and, as a result, the amount of charge in the depletion region becomes $Q + dQ$. The **depletion layer capacitance** C_{dep} is defined by¹²

$$C_{dep} = \left| \frac{dQ}{dV} \right| \quad (3.7.1)$$

Depletion layer capacitance

¹²This is actually an *incremental* capacitance; this is implied by the definition in Eq. (3.7.1).

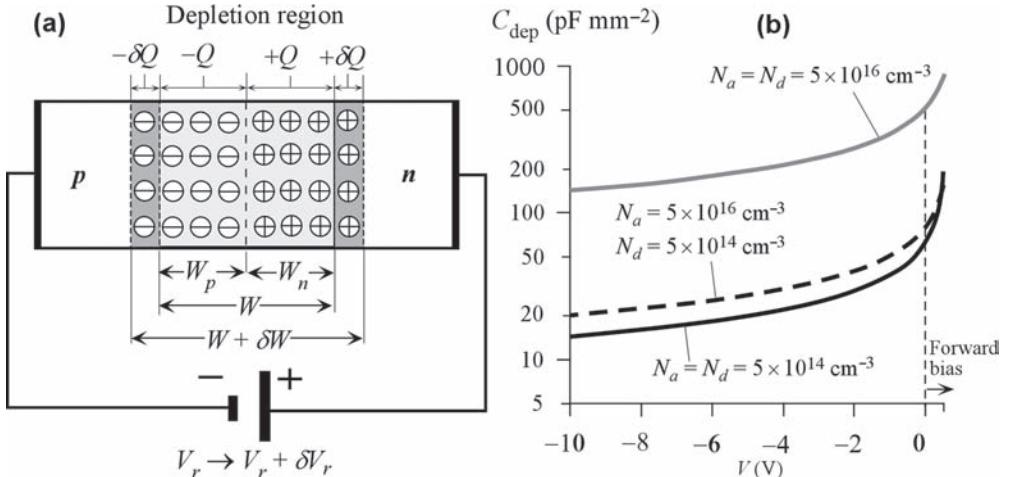


FIGURE 3.23 (a) Depletion region has negative ($-Q$) charges in W_p and positive ($+Q$) charges in W_n , which are separated as in a capacitor. Under a reverse bias V_r , the charge on the n -side is $+Q$. When the reverse bias is increased by δV_r , the charge Q increases by δQ . (b) C_{dep} vs. voltage across an abrupt pn junction for three different sets of dopant concentrations. (Note that the vertical scale is logarithmic.)

If the applied voltage is V then the voltage across the depletion layer W is $V_o - V$ and the depletion region width is

SCL width
and
voltage

$$W = \left[\frac{2\epsilon(N_a + N_d)(V_o - V)}{eN_a N_d} \right]^{1/2} \quad (3.7.2)$$

The amount of charge (on any one side of the depletion layer) is $|Q| = eN_d W_n A = eN_a W_p A$ and $W = W_n + W_p$. We can therefore substitute for W in Eq. (3.7.2) in terms of Q and then differentiate it to obtain dQ/dV . The final result for the depletion capacitance is

Depletion
layer
capacitance

$$C_{\text{dep}} = \frac{\epsilon A}{W} = \frac{A}{(V_o - V)^{1/2}} \left[\frac{e\epsilon(N_a N_d)}{2(N_a + N_d)} \right]^{1/2} \quad (3.7.3)$$

We should note that C_{dep} is given by the same expression as that for the parallel plate capacitor, $\epsilon A/W$, but with W being voltage dependent by virtue of Eq. (3.7.2). Putting a reverse bias $V = -V_r$ in Eq. (3.7.3) shows that C_{dep} decreases with increasing V_r in a $C_{\text{dep}} \propto 1/V_r^{1/2}$ fashion. The reason is that the reverse bias widens W and hence increases the average separation of $+Q$ and $-Q$. Typically C_{dep} under reverse bias is in the $10\text{--}10^3 \text{ pF mm}^{-2}$ depending on N_a and N_d as shown in Figure 3.23 (b).

If $N_a \gg N_d$ as in the case of a p^+n junction, then

Depletion
layer
capacitance
for p^+n

$$C_{\text{dep}} = A \left[\frac{e\epsilon N_d}{2(V_o - V)} \right]^{1/2} \quad (3.7.4)$$

which does not depend on the p -side doping. C_{dep} is present in both forward- and reverse-biased junction.

B. Dynamic Resistance and Diffusion Capacitance for Small Signals

Consider a long *pn* junction diode in which the *p*-side is heavily doped. Under forward bias, minority carriers are injected and diffuse through the neutral regions as shown in Figure 3.16 (a). The excess hole concentration profile $\Delta p_n(x)$ in the *n*-side due to the applied forward bias is shown in Figure 3.24 (a). The injected charge Q is the area under the excess hole concentration profile $\Delta p_n(x)$. It is called the stored minority carrier charge, which is due to diffusing holes in the *n*-side. Recall that the hole concentration profile is a steady state profile. If the hole lifetime is τ_h , the recombination rate of the holes in the *n*-side is determined by Q/τ_h . The current must replenish the holes lost exactly at this rate to maintain a steady state profile. Thus, $I = Q/\tau_h$.

Suppose we increase the diode voltage from V to $V + \delta V$. (This can also be achieved by increasing the drive current from I to $I + \delta I$, which is normally the case in practice.) More holes will be injected into the *n*-side as a result of the law of the junction as shown in Figure 3.24 (a). A change in the diode voltage V to $V + \delta V$, as shown in Figure 3.24 (b), results in change in the forward current. For small changes, we can relate the two because the diode current $I = I_o \exp(eV/k_B T)$, which can be differentiated with respect to V . We define the **dynamic resistance**, also known as the **incremental resistance**, of the diode as

$$r_d = \frac{dV}{dI} = \frac{V_{th}}{I} \quad (3.7.5) \quad \text{Dynamic resistance}$$

where $V_{th} = k_B T/e$ is the thermal voltage (26 mV); r_d can be quite small, 2.5Ω at $I = 10 \text{ mA}$. The dynamic resistance r_d does not represent a true resistance for heat dissipation, but rather a relationship between current and voltage increments imposed by the *pn*-junction behavior. It represents the *diode action* as embedded in $I = I_o \exp(eV/k_B T)$.

When the diode voltage is increased from V to $V + \delta V$, the excess hole concentration just inside the *n*-side will increase from $\Delta p_n(0)$ to $\Delta p_n'(0)$. Additional charge δQ is injected into the *n*-side as shown in Figure 3.24 (a). The increase in the minority carrier charge Q due to the increase

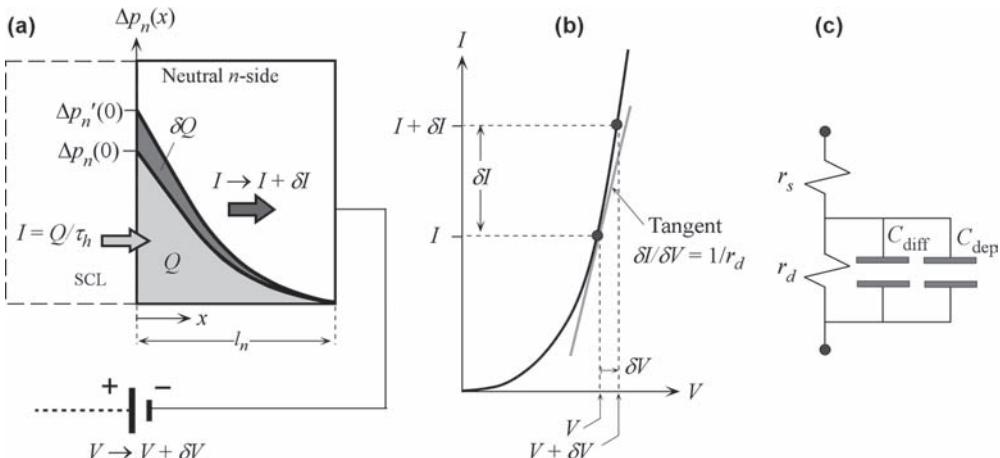


FIGURE 3.24 (a) The forward voltage across a *pn* junction increases by δV , which leads to further minority carrier injection and a larger forward current, which increases by δI . Additional minority carrier charge δQ is injected into the *n*-side. The increase δQ in charge stored in the *n*-side with δV appears as if there is a capacitance across the diode. (b) The increase δV results in an increase δI in the diode current. The dynamic or incremental resistance $r_d = \delta V/\delta I$. (c) A simplified equivalent circuit for a forward-biased *pn* junction for small signals.

in the voltage across the *pn* junction represents a *capacitive behavior* in the sense it takes time to store the additional charge δQ . The additional stored charge δQ has to be removed when $V + \delta V$ returns back to the original value V . The excess charge δQ disappears by recombination, which takes time; again capacitance-like behavior. The **diffusion or storage capacitance** is defined as $C_{\text{diff}} = dQ/dV$. We can easily find C_{diff} . Given $I = Q/\tau_h$ (τ_h being the minority carrier lifetime) under forward bias as in Eq. (3.5.15), we can differentiate Q with respect to V to find

Diffusion capacitance

$$C_{\text{diff}} = \frac{\tau_h I}{V_{\text{th}}} = \frac{\tau_h}{r_d} \quad (3.7.6)$$

It should be mentioned that although under forward bias, C_{dep} can be significant (a few hundred pF mm⁻²) the diffusion or storage capacitance, C_{diff} , due to the injection minority carriers that are diffusing in the neutral regions far exceeds C_{dep} . In fact, C_{diff} can be several nF mm⁻², and a limiting factor in the speed of *pn* junction devices operating in the forward bias regime. We can easily develop a small signal equivalent circuit for a forward-biased junction carrying a current I . There will be a dynamic resistance r_d with two capacitances C_{diff} and C_{dep} in parallel. We should also include a small resistance r_s in series as in Figure 3.24 (c) to represent the resistance of the neutral bulk semiconductor regions. The current has to flow through r_s , which gives rise to a voltage drop through the bulk semiconductor so that not all of applied V drops across the depletion region. Instead of V_o being reduced by V , it is reduced by $V - Ir_s$.

It is important to note that Eq. (3.7.6) is grossly simplified. A proper analysis should impose an AC signal of the form $\delta V = V_m \cos(\omega t)$ on top of the DC operating conditions, where ω is the angular frequency of the modulating signal, and derive the signal current δI , and hence the admittance.¹³ The final result makes r_d as above in Eq. (3.7.5), but C_{diff} is different by a factor 2, that is,

Diffusion capacitance, long diode (AC)

$$C_{\text{diff}} = \frac{\tau_h I}{2V_{\text{th}}} = \frac{\tau_h}{2r_d} \quad (3.7.7)$$

The lifetime τ_h is the minority carrier lifetime in the lesser doped region (holes in the *n*-side). Further Eq. (3.7.7) is valid when $\omega < 1/\tau_h$. At high frequencies ($\omega > 1/\tau_h$), both r_d and C_{diff} become frequency dependent. For a short diode, r_d remains the same. The diffusion capacitance, however, becomes

Diffusion capacitance, short diode (AC)

$$C_{\text{diff}} = \frac{\tau_t I}{V_{\text{th}}} = \frac{\tau_t}{r_d} \quad (3.7.8)$$

where τ_t is the minority carrier transit (diffusion) time. If we have a short diode as in Figure 3.18, and we decrease the voltage, the excess holes will have to diffuse out. The time it takes for a hole to diffuse across the neutral region width l_n is τ_t , and is given by $\tau_t = l_n^2/2D_h$. For a short diode, $C_{\text{diff}} = \tau_t/r_d$.

3.8 RECOMBINATION LIFETIME

A. Direct Recombination

Consider the recombination of an electron and hole in a direct bandgap semiconductor, for example, doped GaAs. Recombination involves a direct meeting of an electron and a hole. Suppose that excess electrons and holes have been injected, as would be in a *pn*-junction under

¹³J. Gower, *Optical Communications Systems*, 2nd Edition (Prentice Hall, Pearson Education, 1993), Ch. 15.

forward bias, and that Δn_p is the excess electron concentration and Δp_p is the excess hole concentration in the *neutral p*-side of a GaAs *pn* junction. Injected electron and hole concentrations would be the same to maintain charge neutrality, that is, $\Delta n_p = \Delta p_p$. Thus, at any instant

$$n_p = n_{po} + \Delta n_p = \text{instantaneous minority carrier concentration}$$

and

$$p_p = p_{po} + \Delta n_p = \text{instantaneous majority carrier concentration}$$

The instantaneous recombination rate will be proportional to both the electron and hole concentrations at that instant, that is, $n_p p_p$. Suppose that the thermal generation rate of EHPs is G_{thermal} . The net rate of *change* of Δn_p is¹⁴

$$d\Delta n_p/dt = -Bn_p p_p + G_{\text{thermal}} \quad (3.8.1)$$

where B is called the **direct recombination capture coefficient**. In equilibrium $d\Delta n_p/dt = 0$ so that setting Eq. (3.8.1) to zero and using $n_p = n_{po}$ and $p_p = p_{po}$, where the subscript *o* refers to thermal equilibrium concentrations, we find $G_{\text{thermal}} = Bn_{po}p_{po}$. Thus, the rate of change in Δn_p is

$$d\Delta n_p/dt = -B(n_p p_p - n_{po} p_{po}) \quad (3.8.2)$$

In many instances the rate of change $d\Delta n_p/dt$ is proportional to Δn_p . It is therefore useful to define an **excess minority carrier recombination time (lifetime)** τ_e by

$$d\Delta n_p/dt = -\Delta n_p/\tau_e \quad (3.8.3)$$

Excess carrier rate of change due to recombination

Recombination time definition

Consider two practical cases where injected excess minority carrier concentration Δn_p is either much less than the actual equilibrium majority carrier concentration p_{po} , or greater. The two cases correspond to weak and strong injection based on Δn_p compared with p_{po} .

In **weak injection**, $\Delta n_p \ll p_{po}$. Then $n_p \approx \Delta n_p$ and $p_p \approx p_{po} + \Delta p_p \approx p_{po} = N_a$, that is, the acceptor concentration. Therefore with these approximations in Eq. (3.8.2) we obtain

$$d\Delta n_p/dt = -BN_a\Delta n_p \quad (3.8.4)$$

Thus, comparing with Eq. (3.8.3), we find the lifetime to be

$$\tau_e = 1/BN_a \quad (3.8.5)$$

Weak injection lifetime

and is constant under weak injection conditions as here.

In **strong injection**, $\Delta n_p \gg p_{po}$. It is easy to show that with this condition, Eq. (3.8.2), becomes

$$d\Delta n_p/dt = -B\Delta p_p\Delta n_p = B(\Delta n_p)^2 \quad (3.8.6)$$

which gives

$$\tau_e = 1/B\Delta n_p = 1/Bn_p \quad (3.8.7)$$

Strong injection lifetime

¹⁴Normally partial derivatives $\partial/\partial t$ would be used since the excess concentration can also depend on x . We ignore this dependence for now to derive the expressions we need for the recombination lifetime.

so that under high-level injection conditions, the lifetime τ_e is inversely proportional to the injected carrier concentration. When a light-emitting diode (LED) is modulated under high injection levels, for example, the lifetime of the minority carriers is therefore not constant, which in turn leads to the distortion of the modulated light output.

B. Indirect Recombination

Consider the recombination of minority carriers such as electrons in an extrinsic indirect bandgap semiconductor such as Si or Ge that have been doped *p*-type. In an indirect bandgap semiconductor, the recombination mechanism involves a recombination center at an energy E_r in Figure 3.14 (c). Such centers are usually crystal defects or impurities in the crystal. The center acts as a third body in the electron and hole recombination process to satisfy the requirements of conservation of momentum. We can view the recombination process as follows. Recombination occurs when an electron is captured by the recombination center at the energy level E_r . As soon as the electron is captured, it will recombine with a hole because holes are abundant in a *p*-type semiconductor. In other words, since there are many majority carriers, the limitation on the rate of recombination is the actual capture of the minority carrier by the center. Suppose that τ_e is the electron recombination time or lifetime, an average time it takes for an electron to recombine. Then, $1/\tau_e$ is the rate of capture per electron. Since the injected electrons will have to be captured by the centers, $1/\tau_e$ is proportional to the capture (recombination) cross-section S_r of the center, the concentration of centers N_r , and also the mean speed of the electron v_{th} (approximately its thermal velocity) all of which increase the rate of capture. Thus,

*Indirect
recombi-
nation via
centers*

$$\tau_e = \frac{1}{S_r N_r v_{th}} \quad (3.8.8)$$

where Eq. (3.8.8) is valid under small injection conditions, that is, $n_p < p_{po}$. There is a more general treatment of indirect recombination called the Shockley-Read-Hall model of indirect recombination and generation, which is treated in more advanced semiconductor physics textbooks; that theory eventually arrives at Eq. (3.8.8) for low-level injection conditions.

EXAMPLE 3.8.1 A direct bandgap *pn* junction

Consider a symmetrical GaAs *pn* junction which has the following properties. N_d (*p*-side doping) = N_d (*n*-side doping) = 10^{17} cm^{-3} (or 10^{23} m^{-3}), direct recombination coefficient¹⁵ $B \approx 2 \times 10^{-16} \text{ m}^3 \text{ s}^{-1}$, cross-sectional area $A = 1 \text{ mm}^2$. Suppose that the forward voltage V across the diode is 0.80 V. What is the diode current due to minority carrier diffusion at 27°C (300 K) assuming direct recombination? If the mean minority carrier lifetime in the depletion region were to be the same as this lifetime, what would be the recombination component of the diode current? What are the two contributions at $V = 1.05 \text{ V}$? What is your conclusion?

Note that at 300 K, GaAs has an intrinsic concentration (n_i) of $2.1 \times 10^6 \text{ cm}^{-3}$ and a relative permittivity (ϵ_r) of 13.0. The hole drift mobility (μ_h) in the *n*-side is $250 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ and electron drift mobility (μ_e) in the *p*-side is $5000 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ (these are at the doping levels given).

¹⁵The direct recombination coefficient B at room temperature is typically of the order $10^{10} \text{ cm}^3 \text{ s}^{-1}$ for many III-V direct bandgap semiconductors; see, for example, E. F. Schubert, *Light-Emitting Diodes*, 2nd Edition (Cambridge University Press, 2006), Table 3.1.

Solution

Assuming weak injection, we can calculate the recombination times τ_e and τ_h for electrons and holes recombining in the neutral p - and n -regions, respectively. Using S.I. units throughout, taking $k_B T/e = 0.02585 \text{ V}$, and since this is a symmetric device,

$$\tau_h = \tau_e \approx \frac{1}{BN_a} = \frac{1}{(2.0 \times 10^{-16} \text{ m}^3 \text{ s}^{-1})(1 \times 10^{23} \text{ m}^{-3})} = 5.00 \times 10^{-8} \text{ s} \quad \text{or} \quad 50.0 \text{ ns}$$

To find the Shockley current we need the diffusion coefficients and lengths. The Einstein relation⁷ gives the diffusion coefficients as

$$D_h = \mu_h k_B T/e = (0.02585)(250 \times 10^{-4}) = 6.46 \times 10^{-4} \text{ m}^2 \text{ s}^{-1}$$

$$D_e = \mu_e k_B T/e = (0.02585)(5000 \times 10^{-4}) = 1.29 \times 10^{-2} \text{ m}^2 \text{ s}^{-1}$$

where $k_B T/e$ was taken as 0.02585 V. The diffusion lengths are easily calculated as

$$L_h = (D_h \tau_h)^{1/2} = [(6.46 \times 10^{-4} \text{ m}^2 \text{ s}^{-1})(50.0 \times 10^{-9} \text{ s})]^{1/2} = 5.69 \times 10^{-6} \text{ m}$$

$$L_e = (D_e \tau_e)^{1/2} = [(1.29 \times 10^{-2} \text{ m}^2 \text{ s}^{-1})(50.0 \times 10^{-9} \text{ s})]^{1/2} = 2.54 \times 10^{-5} \text{ m}$$

Notice that the electrons diffuse much further in the p -side. The reverse saturation current due to diffusion in the neutral regions is

$$\begin{aligned} I_{so} &= A \left(\frac{D_h}{L_h N_d} + \frac{D_e}{L_e N_a} \right) e n_i^2 \\ &= (10^{-6}) \left[\frac{6.46 \times 10^{-4}}{(5.69 \times 10^{-6})(10^{23})} + \frac{1.29 \times 10^{-2}}{(2.54 \times 10^{-5})(10^{23})} \right] (1.6 \times 10^{-19})(2.1 \times 10^{12})^2 \\ &\approx 4.4 \times 10^{-21} \text{ A} \end{aligned}$$

Thus, the forward diffusion current at $V = 0.80 \text{ V}$ is

$$\begin{aligned} I_{\text{diff}} &= I_{so} \exp\left(\frac{eV}{k_B T}\right) \\ &= (4.4 \times 10^{-21} \text{ A}) \exp\left(\frac{0.80 \text{ V}}{0.02585 \text{ V}}\right) = 1.2 \times 10^{-7} \text{ A} \quad \text{or} \quad 0.12 \mu\text{A} \end{aligned}$$

Recombination component of the current is quite difficult to calculate because we need to know the mean electron and hole recombination times in the SCL. Suppose that, as a first order, we assume that these recombination times are as above.

The built-in voltage V_o is

$$V_o = \frac{k_B T}{e} \ln\left(\frac{N_a N_d}{n_i^2}\right) = (0.02585) \ln\left[\frac{10^{23} 10^{23}}{(2.1 \times 10^{12})^2}\right] = 1.27 \text{ V}$$

Depletion layer width W is

$$\begin{aligned} W &= \left[\frac{2\epsilon(N_a + N_d)(V_o - V)}{eN_a N_d} \right]^{1/2} \\ &= \left[\frac{2(13)(8.85 \times 10^{-12} \text{ F m}^{-1})(10^{23} + 10^{23} \text{ m}^{-3})(1.27 - 0.80 \text{ V})}{(1.6 \times 10^{-19} \text{ C})(10^{23} \text{ m}^{-3})(10^{23} \text{ m}^{-3})} \right]^{1/2} \\ &= 1.16 \times 10^{-7} \text{ m} \quad \text{or} \quad 0.116 \mu\text{m} \end{aligned}$$

As this is a symmetric diode, $W_p = W_n = W/2$. The pre-exponential I_{ro} is

$$\begin{aligned} I_{ro} &= \frac{Aen_i}{2} \left(\frac{W_p}{\tau_e} + \frac{W_n}{\tau_h} \right) = \frac{Aen_i}{2} \left(\frac{W}{\tau_e} \right) \\ &= \frac{(10^{-6})(1.6 \times 10^{-19})(2.1 \times 10^{12})}{2} \left(\frac{1.16 \times 10^{-7}}{5.00 \times 10^{-8}} \right) \approx 3.9 \times 10^{-13} \text{ A} \end{aligned}$$

so that at $V = 0.80 \text{ V}$,

$$\begin{aligned} I_{\text{recom}} &\approx I_{ro} \exp\left(\frac{eV}{2k_B T}\right) \\ &\approx (3.9 \times 10^{-13} \text{ A}) \exp\left[\frac{0.8 \text{ V}}{2(0.02585 \text{ V})}\right] = 2.0 \times 10^{-6} \text{ A} \quad \text{or} \quad 2.0 \mu\text{A} \end{aligned}$$

The recombination current is more than an order of magnitude greater than the diffusion current. If we repeat the calculation for a voltage of 1.05 V across the device, then we would find $I_{\text{diff}} = 1.9 \text{ mA}$ and $I_{\text{recom}} = 0.18 \text{ mA}$, where I_{diff} dominates the current. Thus, as the voltage increases across a GaAs pn junction, the ideality factor η is initially 2 but then becomes 1 as shown in Figure 3.20.

The EHP recombination that occurs in the SCL and the neutral regions in this GaAs pn junction case would result in photon emission, with a photon energy that is approximately E_g . This direct recombination of injected minority carriers and the resulting emission of photons represent the principle of operation of the light-emitting diode (LED).

3.9 pn JUNCTION BAND DIAGRAM

A. Open Circuit

Consider a p -type and an n -type semiconductor from the same material (*e.g.*, Si) that are isolated from each other as in Figure 3.25(a). The vacuum level, where the electron is free and has zero potential energy, is also shown and is common to both. The bandgap E_g and the electron affinity χ are

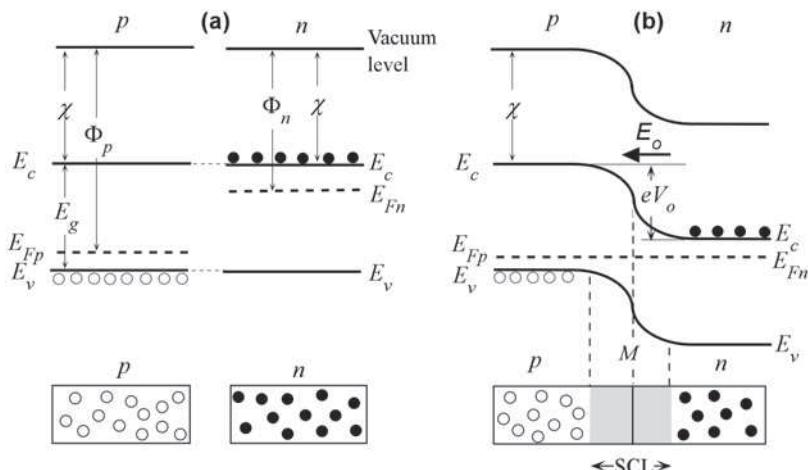


FIGURE 3.25 (a) Consider p - and n -type semiconductor (same material) before the formation of the pn junction, separated from each other and not interacting. (b) After the formation of the pn junction, there is a built-in voltage across the junction.

material properties, and obviously the same for both semiconductors. The Fermi level E_{Fp} on the *p*-side is close to E_v , and E_{Fn} on the *n*-side is close to E_c . The Fermi level is not continuous since the semiconductors are isolated. The work functions Φ_p and Φ_n are also shown.

When we bring the *p*-side and the *n*-side in contact to form the *pn* junction, we form one material system. Equilibrium in the dark in this one-material system requires that the Fermi level be uniform through the two sides of the junction, as illustrated in Figure 3.25 (b). Any change in the Fermi level across the device is equivalent to electrical work done, which must be zero in an open circuit inasmuch as the device is not generating any work and neither is there a voltage applied across it. Consider what happens when we bring the two sides together to form a *pn* junction. Far away from the metallurgical junction *M*, in the bulk of the *n*-type semiconductor we should still have an *n*-type semiconductor, and $E_c - E_{Fn}$ should be the same as in the isolated *n*-type material. Similarly, $E_{Fp} - E_v$ far away from *M* inside the *p*-type material should also be the same as in the isolated *p*-type material. These features are sketched in Figure 3.25 (b) by keeping E_{Fp} and E_{Fn} the same through the whole system and, of course, keeping the bandgap, $E_c - E_v$, the same. Clearly, to draw the energy band diagram we have to *bend the bands*, E_c and E_v , near the junction at *M* because E_c on the *n*-side is close to E_{Fn} , whereas on the *p*-side it is far away from E_{Fp} .

The instant the two semiconductors are brought together to form the junction, electrons diffuse from the *n*-side to the *p*-side and as they do so they deplete the *n*-side near the junction. Thus, E_c must move away from E_{Fn} as we move toward *M* which is exactly what is sketched in Figure 3.25 (b). Holes diffuse from the *p*-side to the *n*-side and the loss of holes in the *p*-type material near the junction means that E_v moves away from E_{Fp} as we move toward *M* as in Figure 3.25 (b). Furthermore, as electrons and holes diffuse toward each other most of them recombine and disappear around *M*, which leads to the formation of a **space charge layer** as we saw in Figure 3.15 (b). The SCL zone around the metallurgical junction has therefore been *depleted* of carriers compared with the bulk. The bending of bands E_c and E_v around *M* accounts for this depletion.

An electron in the *n*-side at E_c must overcome a potential energy (PE) barrier to go over to E_c in the *p*-side. This PE barrier is eV_o where V_o is the built-in potential, the maximum extent E_c has been bent to line up the Fermi levels. Band bending around *M* therefore accounts not only for the variation of electron and hole concentrations in this region but also for the effect of the built-in potential (and hence the built-in field, as the two are related). The diffusion of CB electrons from *n*-side to *p*-side is prevented by the built-in PE barrier eV_o . This barrier also prevents holes from diffusing from the *p*- to the *n*-side. The bending in E_c in the SCL in Figure 3.25 (b) represents the changes in the PE of the electron through this region; and hence it represents the electric field in this region. In fact $|dE_c/dx|$ would be the magnitude of the external force eE on the electron in the SCL. Note that, in the SCL region, the Fermi level is neither close to E_c nor E_v , compared with the bulk or neutral semiconductor regions. This means that both *n* and *p* in this zone are much less than their bulk values n_{no} and p_{po} . The metallurgical junction zone has been depleted of carriers compared with the bulk. Any applied voltage must therefore drop across the SCL.

The extent of band bending depends on how much we need to line up E_{Fp} and E_{Fn} , that is, $\Phi_p - \Phi_n$, as apparent when we compare Figure 3.25 (a) with (b). ($\Phi_p - \Phi_n$ is the $E_{Fn} - E_{Fp}$ difference before the contact.) The built-in potential energy eV_o is therefore $\Phi_p - \Phi_n$. The careful reader will notice that the vacuum level is no longer uniform through the whole device. This is indeed the case and represents the fact that the removal of an electron from the *n*-side will also need some additional work against the fringing electric field. (The exact analysis is quite complicated.)

B. Forward and Reverse Bias

Figure 3.26 (a) shows the energy band diagram of a *pn* junction in open circuit based on the above description (Figure 3.25). We ignored the vacuum level as this is not needed in the following discussion. When the *pn* junction is forward biased, the majority of the applied voltage drops across the depletion region so that the applied voltage is in opposition to the built-in potential, V_o . Figure 3.26 (b) shows the effect of forward bias which is to reduce the PE barrier from eV_o to $e(V_o - V)$. The electrons at E_c in the *n*-side can now readily overcome the PE barrier and diffuse to the *p*-side. The diffusing electrons from the *n*-side can be easily replenished by the negative terminal of the battery connected to this side. Similarly holes can now diffuse from the *p*- to *n*-side.

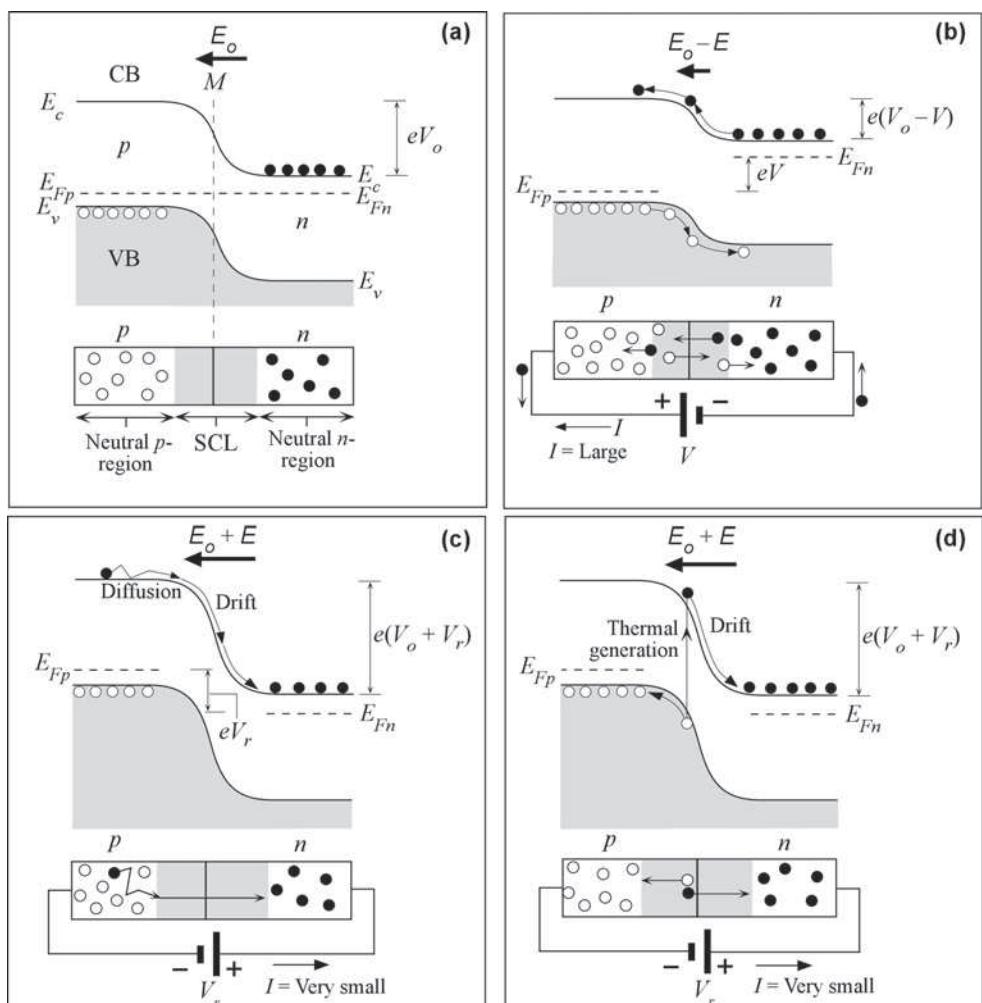


FIGURE 3.26 Energy band diagrams for a *pn* junction under (a) open circuit, (b) forward bias, and (c) and (d) reverse bias conditions. The reverse current in (c) is very small, usually negligible, since it depends on the diffusion of minority carriers to the depletion region. (d) Thermal generation of electron–hole pairs in the depletion region results in a small reverse current that is usually greater than the reverse current in (c).

The positive terminal of the battery can replenish those holes diffusing from the *p*- to the *n*-side. Therefore, a current flows through the junction and around the circuit.

The probability that an electron at E_c in the *n*-side overcomes the new PE barrier and diffuses to E_c in the *p*-side is now proportional to the Boltzmann factor $\exp[-e(V_o - V)/k_B T]$. The latter increases enormously even for small forward voltages. Thus, there is a net diffusion of electrons from the *n*- to *p*-side. Similar ideas also apply to holes at E_v in the *p*-side which also overcome the barrier $e(V_o - V)$ to diffuse into the *n*-side. Since the forward current is due to the number of electrons and holes overcoming the barrier, it is also proportional to $\exp[-e(V_o - V)/k_B T]$ or $\exp(eV/k_B T)$.

When a reverse bias, $V = -V_r$, is applied to the *pn* junction, the voltage again drops across the SCL. In this case, however, V_r adds to the built-in potential V_o so that the PE barrier becomes $e(V_o + V_r)$, as shown in Figure 3.26 (c). The field in the SCL at M increases to $E_o + E$ where E is the applied field (it is not simply V/W). The Shockley model predicts a small reverse saturation current due to the diffusion of minority carriers in neutral regions to the depletion region. Once they reach the depletion region, they will be drifted across by the large field and then collected by the battery. In the energy band diagram in Figure 3.26 (c), an electron in the *p*-side, within a diffusion length (L_e) to the depletion region, can diffuse to the SCL and then *fall down* the PE hill along E_c , over to the *n*-side where it will be collected by the battery. The process of falling down a PE hill is the same process as being driven by a field, in this case by $E_o + E$. The electron lost from the *p*-side is replenished from the negative terminal of the battery (or by thermal generation) to maintain the small reverse current. The same idea applies to the hole in the *n*-side within a diffusion length (L_h) to the SCL. Such minority carrier diffusion is the essence of the Shockley model. The reverse current magnitude is I_{so} . The measured reverse current in Ge diodes at room temperature follow the Shockley model as can be seen in Figure 3.22 (b).

In the case of many semiconductors, however, there is a more significant contribution to the reverse current arising from the thermal generation of electron–hole pairs in the SCL, as shown in Figure 3.26 (d), where the field here separates the pair. The electron falls down the PE hill, down E_c , to the *n*-side to be collected by the battery. Similarly the hole falls down its own PE hill (energy increases downwards for holes) to make it to the *p*-side.

EXAMPLE 3.9.1 The built-in voltage from the band diagram

Derive the expression for the built-in voltage V_o using the energy band diagram in Figure 3.25.

Solution

The extent of band bending, going from (a) to (b) in Figure 3.25, gives the PE barrier eV_o , thus

$$eV_o = \Phi_p - \Phi_n = E_{Fn} - E_{Fp} \text{ (before contact)}$$

Before the contact, on the *n*-side we have

$$n = N_c \exp[-(E_c - E_{Fn})/k_B T] = N_d$$

so that

$$E_c - E_{Fn} = -k_B T \ln(N_d/N_c) \quad (3.9.1)$$

On the *p*-side

$$n = N_c \exp[-(E_c - E_{Fp})/k_B T] = n_i^2/N_a$$

so that

$$E_c - E_{Fp} = -k_B T \ln \left[n_i^2 / (N_a N_c) \right] \quad (3.9.2)$$

Thus, subtracting Eq. (3.9.2) from (3.9.1) gives

Built-in
potential
 V_o

$$eV_o = E_{Fn} - E_{Fp} = k_B T \ln \left[(N_a N_d) / n_i^2 \right] \quad (3.9.3)$$

3.10 HETEROJUNCTIONS

The *pn* junction with the band diagram in Figure 3.25 is a junction within the same crystal (Si), and hence the bandgap does not change along the device; it represents a **homojunction**. A **heterojunction** is a junction between two different semiconductor crystals with different bandgaps E_{g1} and E_{g2} , for example, between GaAs and $\text{Al}_x\text{Ga}_{1-x}$ As ternary alloys. Inasmuch as the bandgaps are now different, their alignment becomes important. If the bandgap difference is $\Delta E_g = E_{g2} - E_{g1}$, then this difference is taken up by a difference $\Delta E_c (= E_{c2} - E_{c1})$ in the CB edges, and $\Delta E_v (= E_{v1} - E_{v2})$ in the VB edges. The energy discontinuities in ΔE_c and ΔE_v are called **band offsets** and play an important role in heterojunction devices.

The terms *heterojunction* and *heterostructure* are frequently used interchangeably, though a heterostructure usually has more than one heterojunction. There may or may not be a change in the doping across the heterojunction. The doping in the wider bandgap semiconductor is usually denoted with a capital letter N or P , and that in the narrower bandgap semiconductor with lower case n or p . There are two cases of particular importance, called Type I and Type II heterojunctions. In a Type I **straddled bandgap alignment heterojunction**, as illustrated in Figure 3.27 (a), the smaller bandgap material offers the lowest energy for both electrons and holes as in GaAs/ $\text{Al}_x\text{Ga}_{1-x}$ As heterostructures in which GaAs has a smaller bandgap than $\text{Al}_x\text{Ga}_{1-x}$ As. Type I is the most common heterostructure in optoelectronic devices, for example, $\text{Ga}_x\text{In}_{1-x}\text{As}/\text{InP}$, $\text{GaAs}/\text{Ga}_x\text{In}_{1-x}\text{P}$. In a Type II **staggered lineup heterojunction**, as illustrated

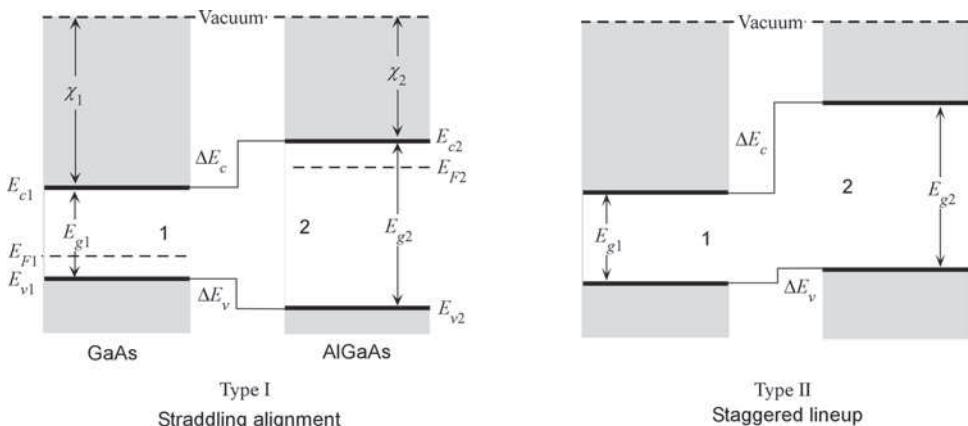


FIGURE 3.27 Two types of heterojunction and the definitions of band offsets Type I and Type II between two semiconductor crystals 1 and 2. Crystal 1 has a narrower bandgap E_{g1} than E_{g2} for crystal 2. Note that the semiconductors are not in contact so that the Fermi level in each is different. In the Type I example, crystal 1 (GaAs) is *p*-type and crystal 2 (AlGaAs) is *N*-type. (Note that the subscripts 1 and 2 refer to semiconductors on the left and right respectively.)

in Figure 3.27 (b), minimum energies for holes and electrons are in the different materials. $\text{Ga}_x\text{In}_{1-x}\text{As}/\text{GaAs}_y\text{Sb}_{1-y}$ heterojunctions over wide compositions follow the Type II behavior.

The band edge profiles for a given heterostructure are determined by the doping levels, carrier transport, and recombination around the junction, as they are for a simple *pn*-homojunction. The open-circuit heterojunction band diagrams are straightforward based on the principle that the Fermi level must be uniform once a contact is made, and we know how ΔE_g is shared between ΔE_c and ΔE_v . Figure 3.28 (a) shows the energy band diagram of an *Np* heterojunction between *n*-type AlGaAs and *p*-type GaAs. First, notice that E_F is uniform through the device as an equilibrium requirement. Far away from the junction on the *N*-side, we have an *n*-type wide bandgap AlGaAs with E_F close to E_c . Far away from the junction on the right, we have a *p*-type narrower bandgap GaAs with E_F close to E_v . In the depletion regions, around the junction, E_c and E_v must bend because there is an internal field, as we know from the ordinary *pn* junction. E_{c1} and E_{v1} of AlGaAs bend upwards and E_{c2} and E_{v2} of GaAs bend downwards as in the normal *pn* junction. We need to join E_{v1} to E_{v2} but also need to account for ΔE_v , which is easily done by putting ΔE_v between E_{v1} and E_{v2} at the junction as shown in Figure 3.28 (a). Similarly, we need to join E_{c1} to E_{c2} but also need to account for ΔE_c . In this case, we can only join E_{c1} and E_{c2} by having a narrow “spike” whose height must be ΔE_c at the junction as shown in Figure 3.28 (a).

One obvious conclusion is that the potential barrier for hole injection, $(E_{v2} - E_{v1})$, from *p* to *N* is greater than the barrier $(E_{c2} - E_{c1})$ for electron injection from *N* to *p*. Under forward bias, the current will be dominated by the injection of electrons from the *N*-side into the *p*-side. ΔE_v has increased the potential barrier against holes and ΔE_c has decreased the potential barrier against electrons.

The energy band diagram for a *pP*-type heterojunction is shown in Figure 3.28 (b). The basic principle for drawing the diagram is the same as before. In this case, there is a small spike

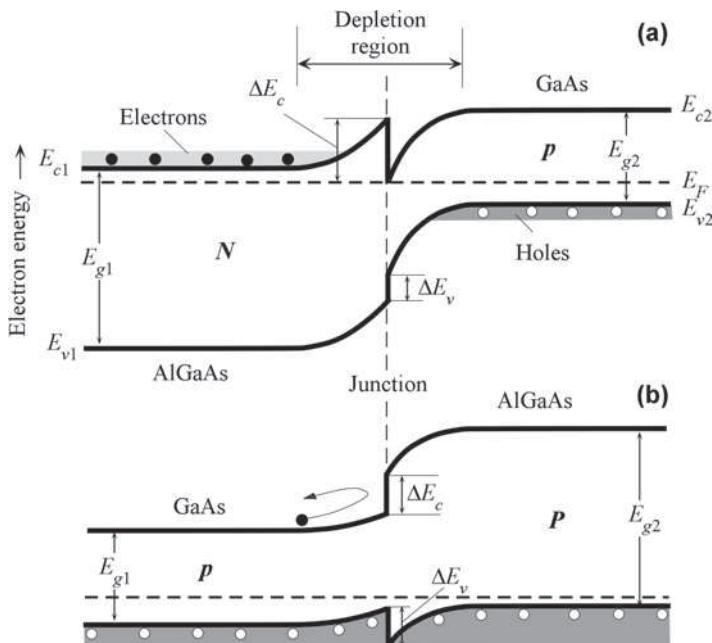


FIGURE 3.28 (a) *nP* and (b) *pP* heterojunctions and their energy band diagrams (schematic only to illustrate general features). Under open circuit and equilibrium conditions, the Fermi level E_F must be uniform, that is, continuous throughout the device. If E_F is close to the conduction band edge, E_c , it results in an *n*-type, and if it is close to the valence band edge, E_v , it results in a *p*-type semiconductor. There is a discontinuity ΔE_c as in E_c , and ΔE_v in E_v , right at the junction.

in E_v at the junction. ΔE_c increases the potential barrier from E_{c1} to E_{c2} . An electron in the CB in the p -side cannot simply overcome this barrier and enter the P -side. Holes in the VB of p - and P -sides can easily cross through the spike by tunneling. Both the Np and pP heterojunctions are used extensively in LED and semiconductor laser diode heterostructures.

3.11 LIGHT-EMITTING DIODES: PRINCIPLES

A. Homojunction LEDs

A light-emitting diode is essentially a pn junction diode typically made from a direct bandgap semiconductor, for example, GaAs, in which the electron–hole pair recombination results in the emission of a photon. The emitted photon energy is therefore approximately equal to the bandgap energy, $h\nu \approx E_g$. Figure 3.29 (a) shows the energy band diagram of an unbiased pn^+ junction device in which the n -side is more heavily doped than the p -side. The band diagram is drawn to keep the Fermi level, E_{Fp} and E_{Fn} on the p - and n -sides, uniform through the device which is a requirement of equilibrium with no applied bias. The depletion region in a pn^+ device extends mainly into the p -side. There is a potential energy (PE) barrier eV_o from E_c on the n -side to E_c on the p -side, that is, $\Delta E_c = eV_o$, where V_o is the *built-in voltage*. The higher concentration of conduction (free) electrons in the n -side encourages the diffusion of these electrons from the n - to the p -side. This net electron diffusion, however, is prevented by the electron PE barrier eV_o .

As soon as a forward bias V is applied, this voltage drops almost entirely across the depletion region since this is the most resistive part of the device. Consequently, the built-in potential V_o is reduced to $V_o - V$, which then allows the electrons from the n^+ side to diffuse, or become injected, into the p -side as illustrated in Figure 3.29 (b). The hole injection component from p into the n^+ side is much smaller than the electron injection component from the n^+ to p -side. The recombination of injected electrons in the depletion region as well as in the neutral p -side results in the *spontaneous emission* of photons. Recombination primarily occurs within the depletion region and within a volume extending over the diffusion length L_e of the electrons in the p -side. (Electron injection is preferred over hole injection in GaAs LEDs because electrons have a higher mobility and hence a larger diffusion coefficient.) This recombination zone is

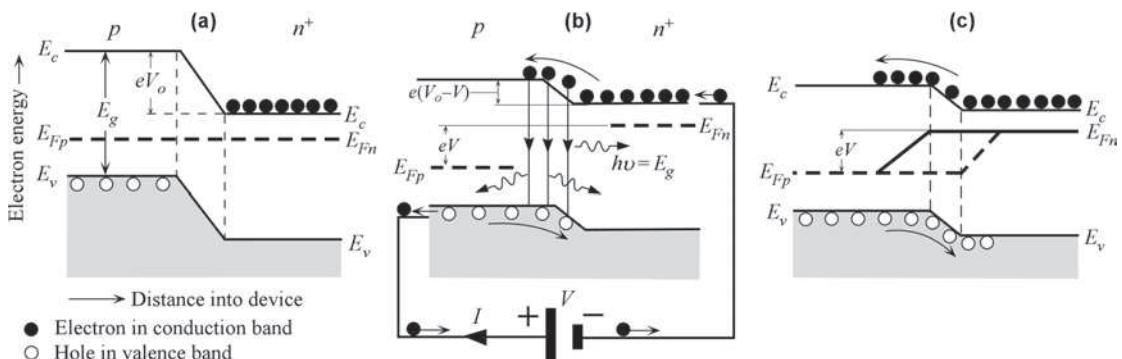


FIGURE 3.29 (a) The energy band diagram of a pn^+ (heavily n -type doped) junction without any bias. Built-in potential V_o prevents electrons from diffusing from n^+ to p -side. (b) The applied bias potential V reduces V_o and thereby allows electrons to diffuse, be injected, into the p -side. Recombination around the junction and within the diffusion length of the electrons in the p -side leads to spontaneous photon emission. (c) Quasi-Fermi levels E_{Fp} and E_{Fn} for holes and electrons across a forward-biased pn -junction.

frequently called the **active region**.¹⁶ The phenomenon of light emission from EHP recombination as a result of minority carrier injection, as in this case, is called **injection electroluminescence**.¹⁷ Because of the statistical nature of the natural recombination process between electrons and holes, the emitted photons are in random directions; such randomly emitted radiation is called **spontaneous emission**. The LED structure has to be such that the emitted photons can escape the device without being reabsorbed by the semiconductor material. This means the *p*-side has to be sufficiently narrow, or we have to use heterostructure devices as discussed below. Further, the rate of recombination depends on the product np , which must be increased to enhance the emitted photon flux. We need to confine the carriers and increase the carrier concentration, again achievable by using heterostructures.

Notice that in Figure 3.29 (b), under forward bias, E_{Fn} and E_{Fp} are separated by eV inasmuch as the electrical work done per electron, eV , must be ΔE_F or $E_{Fn} - E_{Fp}$. Also notice that E_{Fn} and E_{Fp} extend into the depletion region, where there are now two Fermi levels as highlighted in Figure 3.29 (c). We can use E_{Fn} to represent the electron concentration through the general expression $n = N_c \exp[-(E_c - E_{Fn})/k_B T]$, not only on the *n*-side but also in the depletion region, and even in the *p*-side; E_{Fn} is then called a **quasi-Fermi level for electrons**. Similarly, we can use E_{Fp} as a **quasi-Fermi level for holes** and use $p = N_v \exp[-(E_{Fp} - E_v)/k_B T]$ to represent the hole concentration.¹⁸ In the depletion region, E_{Fn} and E_{Fp} are separated within the same spatial region so that $np > n_i^2$, which is expected given that we are injecting minority carriers and have non-equilibrium. Figure 3.29 (c) shows the variation of E_{Fp} and E_{Fn} across a biased *pn* junction. Such variation in E_{Fp} and E_{Fn} will result in $p(x)$ and $n(x)$ profiles across the device that match the sketch in Figure 3.16 (a) for a pn^+ junction; we have effectively represented carrier concentration profiles in terms of E_{Fp} and E_{Fn} on the energy band diagram. Notice that E_{Fn} on the *p*-side slopes down, away from E_c , to represent the decrease in the electron concentration in the neutral *p*-side, due to recombination of injected electrons with the majority carriers (holes). E_{Fn} reaches E_{Fp} over a distance that is very roughly the diffusion length of injected electrons. The region where E_{Fn} and E_{Fp} have merged and are the same represents the neutral bulk *p*-side. Similar arguments apply to holes injected into the *n*-side. E_{Fp} slopes and merges with E_{Fn} over a distance that is roughly the hole diffusion length (shorter than the electron diffusion length).

It is clear that the emission of photons in the LED involves the direct recombination of electrons and holes; that is, an electron makes a radiative transition from the bottom of the CB to the top of the VB, where there is an empty state (a hole) as in Figure 3.14 (a). It may therefore be thought that one should avoid using indirect semiconductors in LEDs. However, this is not entirely true because it is possible to introduce impurities into certain indirect semiconductors and thereby enable radiative transitions through these impurities. A good example is GaP doped with N, that is, GaP:N, which are used in inexpensive green LEDs in numerous indicator and display applications in consumer optoelectronics. GaP is an indirect bandgap semiconductor with E_g of 2.26 eV (corresponding to green light at 550 nm). When N is added, N substitutes for P,

¹⁶The “active region” term is probably more appropriate for laser diodes in which there is photon amplification in this region. However, the term is also used in LED discussions.

¹⁷In 1907, Henry Round, working for Guglielmo Marconi in England, observed yellow light emission when a current was passed through a metal-SiC (silicon carbide or corborandum) crystal rectifier. Oleg Losev in the former USSR carried out a number of experiments on the emission of light upon passing a current through such metal-SiC rectifiers during the mid-1970s. He reported his systematic experiments in a number of journals, called the effect the “inverse photoelectric effect,” and proposed that emission frequency is $\nu = eV/h$. (See Nikolay Zheludev, *Nat. Photonics*, 1, 189, 2007.)

¹⁸Clearly, quasi-Fermi levels provide a convenient and useful means of representing carrier concentrations on the energy diagram.

and is called an **isoelectronic dopant**. However, N dopants create recombination centers that have a localized energy level at E_r , roughly $0.1 - 0.2\text{ eV}$ below the CB as shown in Figure 3.14 (c). An electron at the bottom of the CB can easily fall into E_r during which phonons are emitted to make up for the change in the momentum from k_{CB} to k_{VB} . The electron at E_r can then drop down to the top of the VB and emit a photon $h\nu = E_r - E_v$, with $h\nu$ slightly below E_g . Since $h\nu$ is less than E_g , there is less likelihood of reabsorbing the photon in the bulk of GaP, especially if N doping is introduced in the junction (recombination) region only.

B. Heterostructure High Intensity LEDs

The *pn*-junction LED shown in Figure 3.29 suffers from a number of drawbacks and has a low efficiency. The *p*-region must be narrow to allow the photons to escape without much reabsorption. When the *p*-side is narrow, some of the injected electrons in the *p*-side reach the surface by diffusion and recombine through crystal defects near the surface. This radiationless recombination process decreases the light output. In addition, if the recombination occurs over a relatively large volume (or distance), due to long electron diffusion lengths, then the chances of reabsorption of emitted photons becomes higher; the amount of reabsorption increases with the material volume.

On the other hand, heterostructure LEDs can have exceptionally high efficiencies along with a number of other advantages; most modern LEDs are heterostructure devices (HDS). First, we note that the refractive index of a semiconductor material depends on its bandgap. A wider bandgap semiconductor has a lower refractive index. This means that by constructing LEDs from heterostructures, we can engineer a dielectric waveguide within the device and thereby channel photons out from the recombination region.

LED constructions for increasing the intensity of the output light make use of the double heterostructure (DH) structure. Figure 3.30 (a) shows a **double-heterostructure** (DH) device based on two heterojunctions between different semiconductor crystals with different bandgaps. In this case the semiconductors are AlGaAs ($\text{Al}_x\text{Ga}_{1-x}\text{As}$) with $E_g \approx 2\text{ eV}$ and GaAs with $E_g \approx 1.4\text{ eV}$. The double heterostructure in Figure 3.30 (a) has an N^+p heterojunction between $N^+\text{-AlGaAs}$ and $p\text{-GaAs}$. There is another heterojunction between $p\text{-GaAs}$ and $P\text{-AlGaAs}$. The $p\text{-GaAs}$ region is a thin layer, typically a fraction of a micron and it is lightly doped. The reader would have noticed that this HD device is made of the heterojunctions in Figure 3.28 (a) and (b) one after the other to form an *N-p-P* device.

The simplified energy band diagram for the whole device in the absence of an applied voltage is shown in Figure 3.30 (b). The Fermi-level E_F is continuous through the whole structure. There is a potential energy barrier eV_o for electrons in the CB of $N^+\text{-AlGaAs}$ against diffusion into $p\text{-GaAs}$. The spike in E_c at the junction, as in Figure 3.28 (a), is ignored in this simplified sketch. There is a bandgap change at the junction between $p\text{-GaAs}$ and $P\text{-AlGaAs}$ which results in a step change ΔE_c in E_c between the two CBs of $p\text{-GaAs}$ and $P\text{-AlGaAs}$ as in Figure 3.28 (b). This step change, ΔE_c , is effectively a potential energy barrier that prevents any electron in the CB in $p\text{-GaAs}$ passing into the CB of $P\text{-AlGaAs}$.

When a forward bias is applied, the majority of this voltage drops between the $N^+\text{-AlGaAs}$ and $p\text{-GaAs}$ (*i.e.*, across the depletion region) and reduces the potential energy barrier eV_o , just as in the normal *pn* junction diode. This allows electrons in the CB of $N^+\text{-AlGaAs}$ to be injected (by diffusion) into $p\text{-GaAs}$ as shown in Figure 3.30 (c). These electrons, however, are *confined* to the CB of $p\text{-GaAs}$ since there is a barrier ΔE_c between $p\text{-GaAs}$ and $P\text{-AlGaAs}$ (see Figure 3.28 (b)). The wide bandgap AlGaAs layer therefore acts as a **confining layer** that restricts injected electrons to the $p\text{-GaAs}$ layer. The recombination of injected electrons with the holes present in

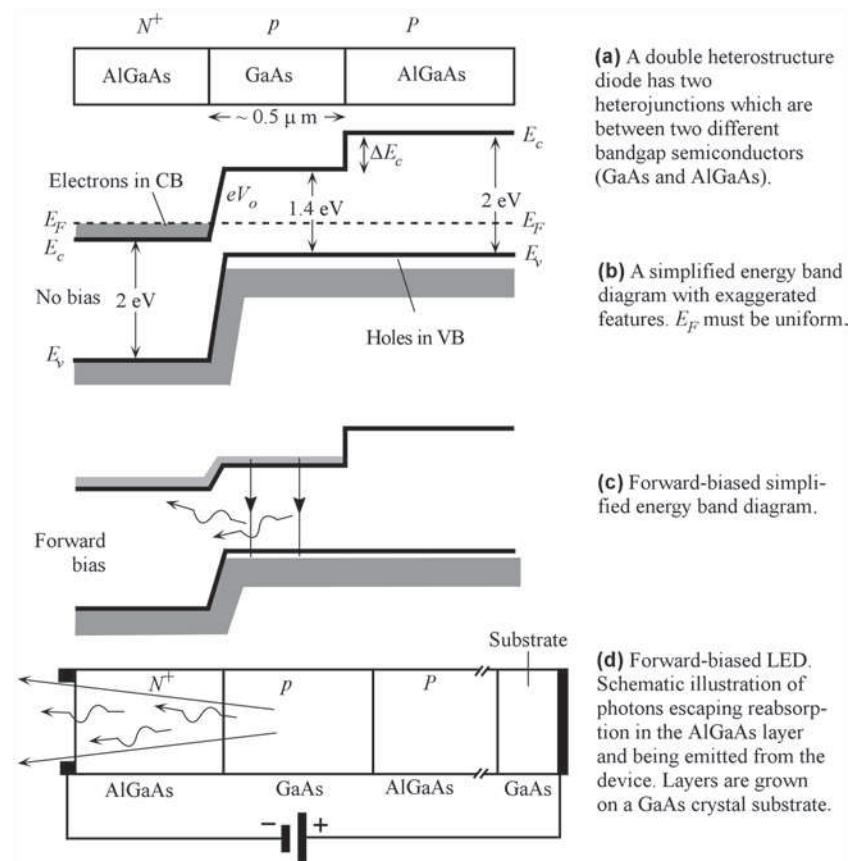
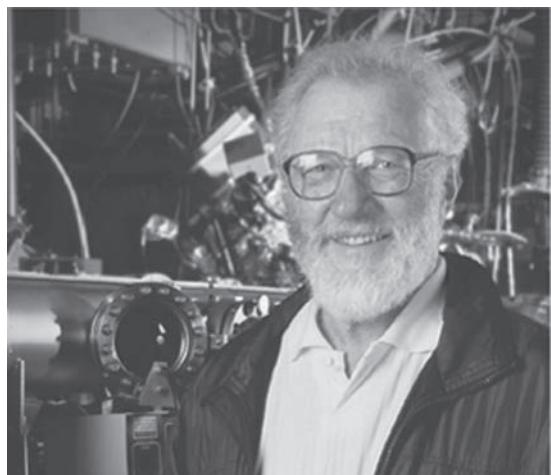


FIGURE 3.30 The double heterostructure AlGaAs/GaAs/AlGaAs LED structure (highly simplified schematic).



Herbert Kroemer (left), along with Zhores Alferov (Chapter 4 opening photo), played a key role in the development of semiconductor heterostructures that are widely used in modern optoelectronics.¹⁹ Herbert Kroemer was also well-recognized for his experimental work on the fabrication of heterostructures by using an atomic layer-by-layer crystal growth technique called Molecular Beam Epitaxy (MBE)—the equipment shown behind Professor Kroemer in the photo. Since 1976, Professor Kroemer has been with the University of California, Santa Barbara, where he continues his research. Herbert Kroemer and Zhores Alferov shared the Nobel Prize in Physics (2000) with Jack Kilby. Their Nobel citation is “*for developing semiconductor heterostructures used in high-speed- and opto-electronics.*” (Courtesy of Herbert Kroemer, University of California, Santa Barbara.)

¹⁹ See H. Kroemer, *Rev. Mod. Phys.*, 73, 783, 2001.

this *p*-GaAs layer results in spontaneous photon emission. Since the bandgap E_g of AlGaAs is greater than that of GaAs, the emitted photons do not get reabsorbed as they escape the active region and can reach the surface of the device as illustrated in Figure 3.30 (d). Since light is also not absorbed in *P*-AlGaAs, it can be *reflected* to increase the light output, by, for example, using a dielectric mirror at the end of *P*-AlGaAs. Another advantage of the AlGaAs/GaAs heterojunction is that there is only a small lattice mismatch between the two crystal structures and hence negligible strain induced interfacial defects (*e.g.*, dislocations) in the device compared with the defects at the surface of the semiconductor in conventional homojunction LED structure. The DH LED is much more efficient than the homojunction LED.

C. Output Spectrum

The energy of an emitted photon from an LED is not simply equal to the bandgap energy E_g because electrons in the conduction band are distributed in energy and so are the holes in the valence band. Suppose that the active region is *p*-type, and excess electrons have been injected by forward bias. Figure 3.31 (a) and (b) illustrates the energy band diagram and the energy distributions of electrons and holes in the CB and VB, respectively, for a *p*-type semiconductor. The electron concentration as a function of energy in the CB is given by $g(E)f(E)$, where $g(E)$ is the density of states in the CB and $f(E)$ is the Fermi–Dirac function (probability of finding an electron in a state with energy E). The product $g(E)f(E)$ represents the electron concentration per unit energy, $n_E(E)$, and is plotted along the horizontal axis in Figure 3.31 (b). There is a similar energy distribution for holes, p_E , in the VB but p_E is enormously larger than n_E . The E – k diagram for a typical direct bandgap semiconductor (such as GaAs) is shown in Figure 3.31 (c). Since the hole concentration is very large, we can assume that the rate of recombination will depend primarily on the concentration of injected electrons.²⁰ The electron concentration in the CB

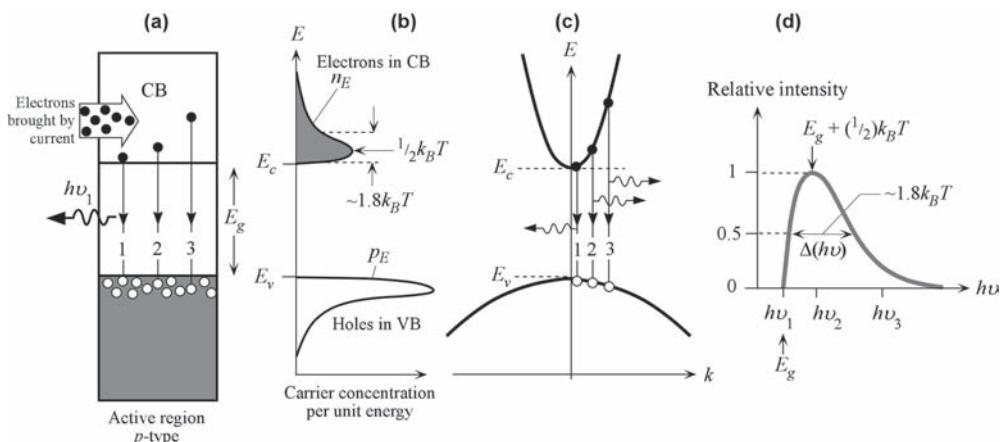


FIGURE 3.31 (a) Energy band diagram with possible recombination paths. (b) Energy distribution of electrons in the CB and holes in the VB. The highest electron concentration is $(1/2)k_B T$ above E_c . (c) A simplified E – k (equivalent to energy vs. momentum) diagram and direct recombination paths in which k (*i.e.*, momentum) is conserved. (d) The relative light intensity as a function of photon energy based on (b) and (c).

²⁰We also need to consider the quantum mechanical transition probability from an occupied state in the CB to an empty state in the VB but, for simplicity, we will take this to be constant, that is, we will ignore it.

as a function of energy is asymmetrical, and has a peak at $\frac{1}{2}k_B T$ above E_c . The energy spread of these electrons is about $1.8k_B T$ between the half-maximum points, as in Figure 3.31 (b), or roughly $\sim 2k_B T$. When an electron at E_c recombines with a hole at E_v , shown as the transition 1 in Figure 3.31 (a), a photon is emitted with an energy $h\nu_1 = E_c - E_v = E_g$. Since there are not many electrons and holes at the band edges, this type of recombination does not occur frequently, and the emitted light intensity from a type 1 transition is small.

The transition that involves the largest electron concentration (the peak in n_E) is shown as 2 in Figure 3.31 (a), and emits a photon with $h\nu_2 > h\nu_1$. Such transitions occur frequently (large n_E), and hence the emitted intensity from type 2 transitions is much larger than that for type 1. Similarly, the transition marked 3, corresponding to $h\nu_3 > h\nu_2$, involves an electron quite high up in the CB where n_E is very small. Such type 3 transitions are infrequent and lead to a small emission intensity. The emission intensity therefore rises to a maximum and then falls with $h\nu$ as shown in Figure 3.31 (d).

One might conclude that the highest emitted intensity should intuitively correspond to the transition from the peak in n_E to the peak in p_E in Figure 3.31 (b) that emits $h\nu = E_g + k_B T$. However, we also need to consider the conservation of momentum, and hence the E - k diagram in Figure 3.31 (c). The momentum of the emitted photon is negligibly small, so that an electron falls straight down in the E - k diagram without changing its k -vector, that is, the electron momentum $\hbar k$ is conserved. The E - k curvatures are different in the CB and the VB. The electron at $E_c + \frac{1}{2}k_B T$ cannot just recombine with the hole at $E_v + \frac{1}{2}k_B T$ because that transition does not satisfy the $\hbar k$ -conservation. As shown in Figure 3.31 (c), direct recombination involves energetic electrons spreading over several $k_B T$ in the CB, more than the holes in the VB, because the E - k curvature is narrower in the CB and broader in the VB. It is apparent that the emission spectrum in this case is determined by n_E , the energy spread in the electrons in the CB, so that the emission has a peak at roughly $E_g + \frac{1}{2}k_B T$. Further the spread $\Delta(h\nu)$ in the emitted photon energies should be roughly the spread in n_E , that is, $\Delta(h\nu) \approx 1.8k_B T$.

The intuitive relative light intensity vs. photon energy characteristic of the output spectrum based on n_E and the E - k diagram is shown in Figure 3.31 (d) and represents an important LED characteristic. Given the spectrum in Figure 3.31 (d), we can also obtain the relative light intensity vs. wavelength characteristic since $\lambda = c/\nu$, which would look like Figure 3.31 (d) flipped horizontally. The **linewidth** of the output spectrum, $\Delta\nu$ or $\Delta\lambda$, is defined as the width between half-intensity points as illustrated in Figure 3.31 (d).

The experimentally observed output spectrum, that is, the relative intensity vs. wavelength characteristics, from an LED depends not only on the semiconductor material, including dopant concentrations, but also on the structure of the pn junction diode. The spectrum in Figure 3.31 (d) represents a highly idealized spectrum without including the effects of heavy doping on the energy bands. For a heavily doped n -type semiconductor there are so many donors that the electron wavefunctions at these donors overlap to generate a narrow impurity band centered at E_d but extending into the conduction band. Thus, the donor impurity band overlaps the conduction band and hence effectively lowers E_c as in Figure 3.10 (a). The minimum emitted photon energy from heavily doped semiconductors is therefore less than E_g and depends on the amount of doping.

Typical output spectrum from an AlGaAs IR LED is shown in Figure 3.32 (a). Notice that the spectrum exhibits significantly less asymmetry than the idealized spectrum in Figure 3.31 (d). The width of the spectrum is about 40 nm, which corresponds to a width of about $2.9k_B T$ in the energy distribution of the emitted photons, more than the expected $1.8k_B T$. The reasons for not observing the theoretical spectrum in Figure 3.31 (d) are essentially twofold. First, higher energy photons become reabsorbed in the material and photogenerate electrons and holes.

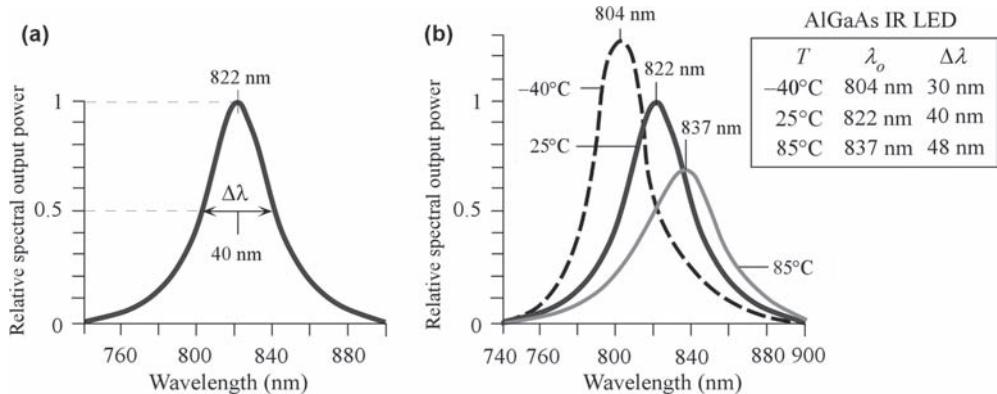


FIGURE 3.32 (a) A typical output spectrum (relative intensity vs. wavelength) from an IR (infrared) AlGaAs LED. (b) The output spectrum of the LED in (a) at three temperatures: 25°C, -40°C, and 85°C. Values normalized to peak emission at 25°C. The spectral widths are FWHM.

These electrons and holes thermalize and end up recombining to emit photons with lower energies, closer to E_g . Thus photons become redistributed. Secondly, the band edges E_c and E_v are not sharp in heavily doped semiconductors, which leads to the smearing of the well-defined E_g for the emission onset. The peak emission frequency ν_o and the spectral width $\Delta\nu$ in photon energy in LEDs with a direct bandgap active region are normally described by

LED
spectrum
in
frequency

$$h\nu_o \approx E_g + \frac{1}{2}k_B T \quad \text{and} \quad h\Delta\nu = mk_B T \quad (3.11.1)$$

where m is a numerical factor that is typically between 1.5 and 3.5, though for many LEDs, $m \approx 3$ is a good value (see Question 3.23). The corresponding peak wavelength λ_o and the spectral width $\Delta\lambda$ can be easily found from Eq. (3.11.1) as in Example 3.11.1. Further, the actual position of the peak is likely to be somewhat more than $\frac{1}{2}k_B T$ in Eq. (3.11.1), given a broader observed spread than $1.8k_B T$ expected from the theory. As the temperature increases, the change in $h\nu_o$ in Eq. (3.11.1) is due mainly to the decrease in the bandgap E_g with temperature. The peak emission wavelength λ_o , corresponding to ν_o , therefore increases with temperature as shown Figure 3.32 (b). In addition, the linewidth $\Delta\lambda$ becomes longer at higher temperatures as electrons are distributed further into the CB. Thus, a wider spectrum of photon energies are emitted upon EHP recombination. The variation of the bandgap E_g with temperature is commonly represented by the **Varshni equation**

Varshni
bandgap
equation

$$E_g = E_{go} - \frac{AT^2}{B + T} \quad (3.11.2)$$

where E_{go} is the bandgap at $T = 0$ K, and A and B are material-specific constants that are listed in various semiconductor handbooks.

Equation (3.11.1) does not apply to an indirect bandgap semiconductor in which a recombination center is involved in the radiative transition, such as GaP:N. The electron localized at the recombination center would have a significant uncertainty in its momentum Δp and hence an uncertainty ΔE in its energy (Heisenberg's uncertainty principle, $\Delta p\Delta x \sim \hbar$). The emitted photon spectrum depends on this ΔE , and is wider than $3k_B T$ that is involved in the direct recombination process in Figure 3.31.

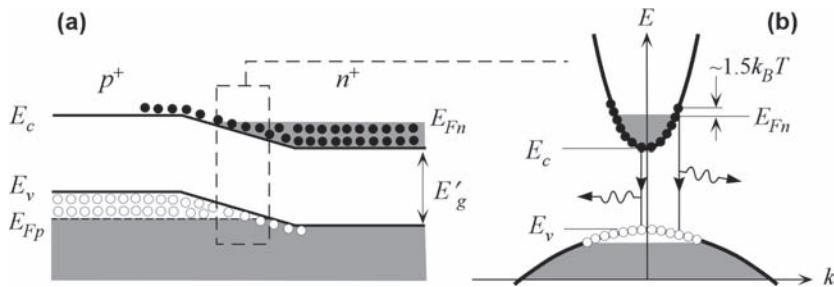


FIGURE 3.33 (a) Forward-biased degenerately doped *pn* junction. The bandgap E'_g is narrower than that in the undoped bulk crystal. The quasi-Fermi levels E_{Fn} and E_{Fp} overlap around the junction. (b) The transitions involved in a degenerately doped *pn* junction.

For degenerately doped junctions, the Fermi level E_{Fn} on the *n*-side and E_{Fp} on the *p*-side will be in the CB and VB, respectively. Under a large forward bias, the active region can have E_{Fn} in the CB, E_{Fp} in the VB around the junction as shown in Figure 3.33 (a). The bandgap E'_g is narrower than E_g in the undoped crystal (see Figure 3.10 (a) and (b)). As shown in the E - k diagram in Figure 3.33 (b), electrons occupy states from the CB edge E_c up to about $\sim 1.5k_B T$ above E_{Fn} . The emission spectrum will extend from $h\nu \approx E_g$ to about $E_g + E_{Fn} + 1.5k_B T$ so the width is roughly $E_{Fn} + 1.5k_B T$.

EXAMPLE 3.11.1 LED spectral linewidth

We know that a spread in the output wavelengths is related to a spread in the emitted photon energies as illustrated in Figure 3.31. The emitted photon energy $h\nu = hc/\lambda$. Assume that the spread in the photon energies $\Delta(h\nu) \approx 3kT$ between the half intensity points. Show that the corresponding linewidth $\Delta\lambda$ between the *half intensity points* in the output spectrum is

$$\Delta\lambda = \lambda_o^2 \frac{3kT}{hc} \quad (3.11.3)$$

*LED
spectral
linewidth*

where λ_o is the peak wavelength. What is the spectral linewidth of an optical communications LED operating at 1310 nm and at 300 K?

Solution

First consider the relationship between the photon frequency ν and λ ,

$$\lambda = \frac{c}{\nu} = \frac{hc}{h\nu}$$

in which $h\nu$ is the photon energy. We can differentiate this

$$\frac{d\lambda}{d(h\nu)} = -\frac{hc}{(h\nu)^2} = -\frac{\lambda^2}{hc} \quad (3.11.4)$$

The negative sign implies that increasing the photon energy decreases the wavelength. We are only interested in changes or spreads, thus $\Delta\lambda/\Delta(h\nu) \approx |\lambda^2/d(h\nu)|$, and this spread should be around $\lambda = \lambda_o$, so that Eq. (3.11.4) gives

$$\Delta\lambda = \frac{\lambda_o^2}{hc} \Delta(h\nu) = \lambda_o^2 \frac{3kT}{hc}$$

where we used $\Delta(h\nu) = 3kT$. We can substitute $\lambda = 1310 \text{ nm}$ and $T = 300 \text{ K}$ to calculate the linewidth of the 1310 nm LED

$$\Delta\lambda = \lambda^2 \frac{3kT}{hc} = (1310 \times 10^{-9})^2 \frac{3(1.38 \times 10^{-23})(300)}{(6.626 \times 10^{-34})(3 \times 10^8)} = 1.07 \times 10^{-7} \text{ m or } 107 \text{ nm}$$

The spectral linewidth of an LED output is due to the spread in the photon energies, which is fundamentally about $3kT$. The only option for decreasing $\Delta\lambda$ at a given wavelength is to reduce the temperature. The output spectrum of a laser, on the other hand, has a much narrower linewidth.

EXAMPLE 3.11.2 LED spectral width

Consider the three experimental points in Figure 3.32 (b) as a function of temperature. By a suitable plot find m and verify Eq. (3.11.3).

Solution

From Example 3.11.1, we can use Eq. (3.11.3) with m instead of 3 as follows

$$\frac{\Delta\lambda}{\lambda_o^2} = \left(\frac{mk}{hc}\right)T \quad (3.11.5)$$

and plot $\Delta\lambda/\lambda_o^2$ vs. T . The slope of the best line forced through zero should give mk/hc and hence m . Using the three λ_o and $\Delta\lambda$ values in the inset of Figure 3.32 (b), we obtain the graph in Figure 3.34. The best line is forced through zero to follow Eq. (3.11.5), and gives a slope of $1.95 \times 10^{-7} \text{ nm}^{-1} \text{ K}^{-1}$ or $195 \text{ m}^{-1} \text{ K}^{-1}$. Thus,

$$\text{Slope} = 195 \text{ m K}^{-1} = \frac{m(1.38 \times 10^{-23} \text{ J K}^{-1})}{(6.626 \times 10^{-34} \text{ J s})(3 \times 10^8 \text{ m s}^{-1})}$$

so that

$$m = 2.81$$

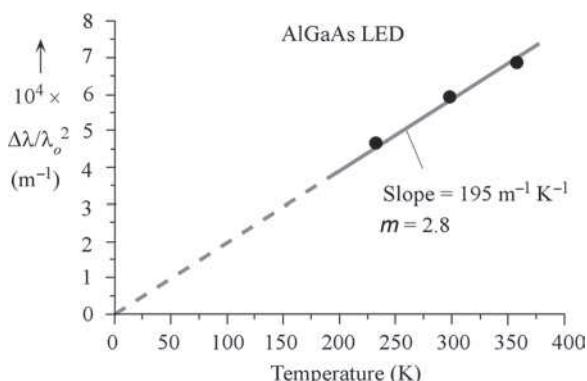
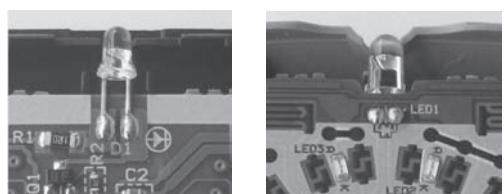


FIGURE 3.34 The plot of $\Delta\lambda/\lambda_o^2$ vs. T for an AlGaAs infrared LED, using the peak wavelength λ_o and spectral width $\Delta\lambda$ at three different temperatures, using the data shown in Figure 3.32 (b).



Infrared LEDs are widely used in various remote controls.

EXAMPLE 3.11.3 Dependence of the emission peak and linewidth on temperature

Using the Varshni equation, Eq. (3.11.2), find the shift in the peak wavelength (λ_o) emitted from a GaAs LED when it is cooled from 25°C to -25°C. The Varshni constants for GaAs are $E_{go} = 1.519 \text{ eV}$, $A = 5.41 \times 10^{-4} \text{ eV K}^{-1}$, $B = 204 \text{ K}$.

Solution

At $T = 298 \text{ K}$, using the Varshni equation

$$\begin{aligned} E_g &= E_{go} - AT^2/(B + T) \\ &= 1.519 \text{ eV} - (5.41 \times 10^{-4} \text{ eV K}^{-1})(298 \text{ K})^2/(204 \text{ K} + 298 \text{ K}) = 1.423 \text{ eV} \end{aligned}$$

At 298 K, $(1/2)k_B T = 0.0128 \text{ eV}$. The peak emission is at $h\nu_o \approx E_g + (1/2)k_B T$. Using $\nu_o = c/\lambda_o$, we get

$$\lambda_o = \frac{ch}{(E_g + \frac{1}{2}k_B T)} = \frac{(3 \times 10^8 \text{ m s}^{-1})(6.626 \times 10^{-34} \text{ J s})/(1.602 \times 10^{-19} \text{ eV J}^{-1})}{(1.4223 \text{ eV} + 0.0128 \text{ eV})} = 864.2 \text{ nm}$$

At -25°C, or 248 K, $(1/2)k_B T = 0.0107 \text{ eV}$, repeating the above calculation

$$E_g = 1.519 \text{ eV} - (5.41 \times 10^{-4} \text{ eV K}^{-1})(248 \text{ K})^2/(204 \text{ K} + 248 \text{ K}) = 1.445 \text{ eV}$$

and the new peak emission wavelength λ'_o is

$$\lambda'_o = \frac{(3 \times 10^8 \text{ m s}^{-1})(6.626 \times 10^{-34} \text{ J s})/(1.602 \times 10^{-19} \text{ eV J}^{-1})}{(1.445 \text{ eV} + 0.0107 \text{ eV})} = 852.4 \text{ nm}$$

The change $\Delta\lambda = \lambda_o - \lambda'_o = 864.2 - 852.4 = 11.8 \text{ nm}$ over 50°C, or $0.24 \text{ nm}/^\circ\text{C}$. The examination of Figure 3.32 (b) shows that the change in the peak wavelength per unit temperature in the range -40°C to 85°C is roughly the same. Because of the small change, we kept four significant figures in E_g and λ_o calculations.

3.12 QUANTUM WELL HIGH INTENSITY LEDs

A typical **quantum well** (QW) device has an ultra-thin, typically less than 50 nm, narrow bandgap semiconductor with a bandgap E_{g1} sandwiched between two wider bandgap semiconductors with a bandgap E_{g2} , as illustrated in Figure 3.35 (a). For example, this could be a thin GaAs (E_{g1}) layer sandwiched between two $\text{Al}_x\text{Ga}_{1-x}\text{As}$ (E_{g2}) layers. The wide bandgap layers are called **confining** layers. We assume that the two semiconductors are lattice matched in the sense that they have the same crystal structure and lattice parameter a . This means that interface defects due to the mismatch of crystal dimensions between the two semiconductor crystals are minimal; and neglected. Since the bandgap, E_g , changes at the interface, there are discontinuities in E_c and E_v at the interfaces as before; these discontinuities, ΔE_c and ΔE_v , are shown in Figure 3.35 (b) and depend on the semiconductor properties.²¹ Because of the potential energy barrier, ΔE_c , conduction electrons in the thin E_{g1} -layer are confined in the x -direction. This confinement length d , the width of the thin E_{g1} -semiconductor, is so small that we can treat the electron as

²¹ In the case of GaAs/AlGaAs heterostructure, ΔE_c is greater than ΔE_v . Very approximately, the change from the wider E_{g2} to narrower E_{g1} is proportioned 60% to ΔE_c and 40% to ΔE_v .

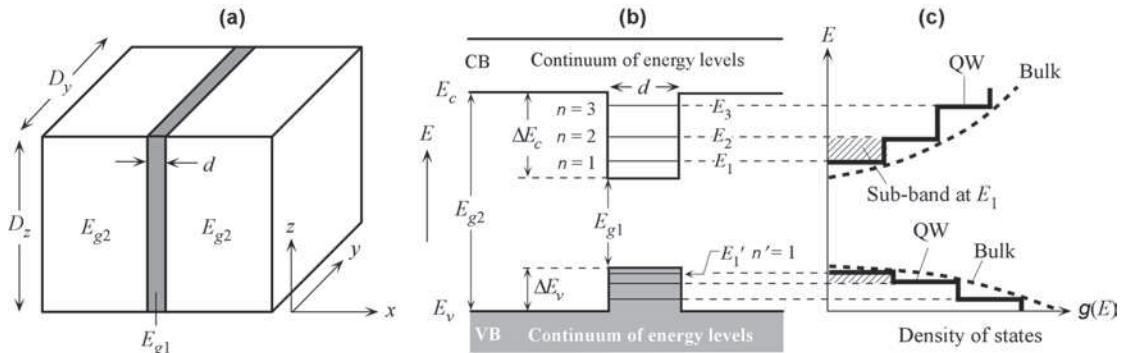


FIGURE 3.35 (a) A single quantum well (SQW) of a smaller bandgap material (E_{g1}) of thickness d along x surrounded by a thicker material of wider bandgap (E_{g2}). (b) The electron energy levels associated with motion along x are quantized as E_1 , E_2 , E_3 , etc. Each level is characterized by a quantum number n . (c) The density of states for a bulk semiconductor and a QW.

in a one-dimensional (1D) potential energy (PE) well in the x -direction but as if it were free in the yz plane.

The energy of the electron in the QW must reflect its 1D quantization in the x -direction, and its freedom in the yz plane. If E_n is the electron energy in the well, then

Energy
in a 1D
quantum
well

$$E_n = E_c + \frac{\hbar^2 n^2}{8m_e^* d^2} + \frac{\hbar^2 k_y^2}{2m_e^*} + \frac{\hbar^2 k_z^2}{2m_e^*} \quad (3.12.1)$$

where n is a quantum number having the values $1, 2, 3, \dots$, and k_y and k_z are the wave vectors of the electron along y - and z -directions. The reason for the E_c in Eq. (3.12.1) is that the potential energy barriers are defined with respect to E_c . These PE barriers are ΔE_c along x and electron affinity (energy required to take the electron from E_c to vacuum) along y and z . The second term is the energy of an electron in an infinite PE well, whereas we have a finite PE well of depth ΔE_c . Thus, the second term is only an approximation. The minimum energy E_1 corresponds to $n = 1$ and is above E_c of the E_{g1} -semiconductor as shown in Figure 3.35 (b). For any given n value, we have a sub-band of energies due to k_y and k_z terms in Eq. (3.12.1). The separation between the energy levels associated with motion in the yz plane in a sub-band is so small that the electron is free to move in the yz plane as if it were in the bulk semiconductor; we assume a continuum of energy as shown in Figure 3.36. We therefore have a *two-dimensional electron gas* which is confined in the x -direction. The holes in the valence band are confined by the potential energy barrier ΔE_v (hole energy is in the opposite direction to electron energy) and behave similarly as illustrated in Figure 3.35 (b). They are characterized by the quantum number $n' = 1, 2, \text{etc.}$ corresponding to the levels $E'_1, E'_2, \text{etc.}$ [only E'_1 is marked in Figure 3.35 (b)].

The density of electronic states for the two-dimensional electron system is not the same as that for the bulk semiconductor. For a given electron concentration n , the density of states $g(E)$, the number of quantum states per unit energy per unit volume, is constant and does not depend on the electron energy. The density of states for the confined electron and that in the bulk semiconductor are shown schematically in Figure 3.35 (c). $g(E)$ is constant at E_1 until E_2 , where it increases as a step and remains constant until E_3 , where again it increases as a step by the same amount and at every value of E_n . Density of states in the valence band behaves similarly as shown in Figure 3.35 (c). Since the electron is free in the yz plane within the well, its kinetic energy will increase parabolically and continuously with the wave vector k_{yz} in this plane as

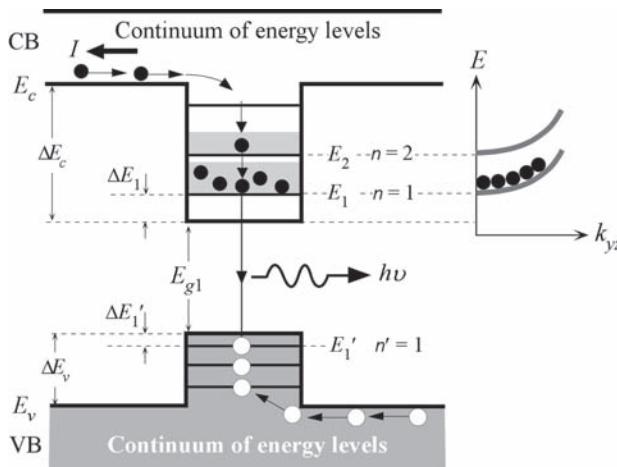


FIGURE 3.36 A QW structure that shows the energy levels in the wells and how charge carriers that are brought in by the current fall into the lowest energy level in the well and then recombine, emitting a photon. The electrons at a particular energy level also have kinetic energies in the yz plane, which is not quantized. The electrons are therefore spread in energy above E_n as shown. The same notion also applies to holes in the ΔE_v well.

shown in Figure 3.36. One has to account for this small kinetic energy, which is shown by shading the region above E_1 , E_2 , etc., in Figure 3.36.

Since at E_1 there is a finite and substantial density of states, the electrons in the conduction band do not have to spread as far in energy as in the bulk to find states. In the bulk semiconductor, on the other hand, the density of states at E_c is zero and increases slowly with energy (as $E^{1/2}$), which means that the electrons are spread more deeply into the conduction band in search for states. A large concentration of electrons can easily occur near E_1 whereas this is not the case in the bulk semiconductor. Similarly, the majority of holes in the valence band will be around E'_1 since there are sufficient states at this energy. Under a forward bias, as indicated in Figure 3.36, electrons are injected into the conduction band of the E_{g1} -layer, which serves as the *active layer*. The injected electrons readily populate the ample number of states at E_1 which means that the electron concentration at E_1 increases rapidly with the current. The direct recombination rate, that is, the radiative recombination rate, which determines the rate of photon emission, depends on the product np , the concentration of electrons *and* holes. Thus, the radiative transitions occur much more readily in the quantum well for the same current as in the bulk device. The quantum mechanical radiative recombination probability is also enhanced because the electrons and holes in the well are brought into close proximity. (However, there are still certain selection rules as there are for photon emission in the H-atom.) Furthermore, since the majority of electrons are spread around E_1 and holes around E'_1 , the range of emitted photon energies can be narrower than the emission spectrum in the corresponding bulk device; unless the well has been flooded with electrons at large currents as explained below. The loss of energy as the electron drops from $n = 3$ to 2 to 1 , etc., to lower levels inside the well occurs by the emission of phonons (lattice vibrations) and is very rapid. The transition from $n = 1$ to $n' = 1$ emits a photon of energy $h\nu \approx E_1 - E'_1$. (An approximate sign is used since the electron and the hole can have kinetic energies in the yz plane). The radiative transition in a QW must obey a **selection rule**, which requires the initial and final quantum numbers, n and n' , to be the same.²² The transition from $n = 1$ to $n' = 1$ is allowed and emits a photon, as well as that from $n = 2$ to $n' = 2$.

²²Quantum mechanical selection rules for radiative transitions are nothing unusual; they also occur in radiative transitions in atoms. A very readable derivation can be found in M. Fox, *Optical Properties of Solids*, 2nd Edition (Oxford University Press, 2010), Ch. 6.

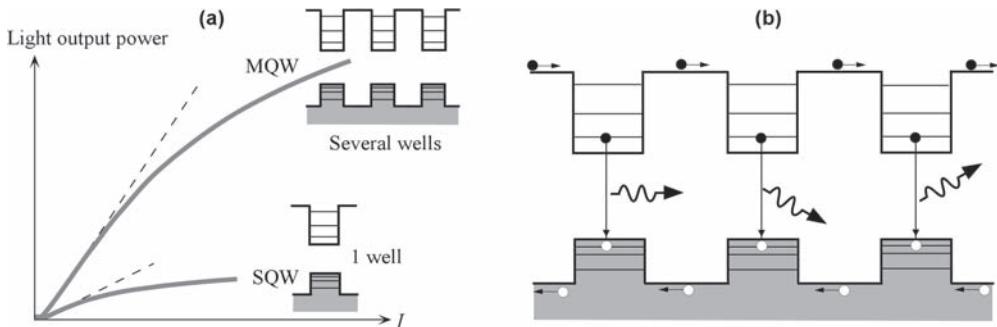


FIGURE 3.37 A schematic illustration of the comparison of light power output vs. current characteristics for an SQW and an MQW LED.

The main problem with the single quantum well (SQW) heterostructure LEDs is that, under a sufficiently large current, the well can be flooded with charge carriers and can overflow. For example, electrons can flood the QW ΔE_c and the well will overflow. The advantages of the QW action (such as confinement that increases the electron concentration n) would be lost. The light output will no longer increase proportionally to the current, and will fall behind the increase in the current as indicated in Figure 3.37 (a). This problem can be resolved by using **multiple quantum wells** (MQWs), in which electrons are shared by a number of quantum wells as in Figure 3.37 (b). Not only is the overflow problem alleviated but more photon flux is generated due to more wells as shown in Figure 3.37 (a). Modern high intensity UV, violet, and blue LEDs use MQW heterostructures. They use a thin $\text{In}_x\text{Ga}_{1-x}\text{N}(E_g)$ QW layer that is sandwiched between GaN (E_{g2}) layers. GaN has a large bandgap of 3.4 eV, and the composition and hence the bandgap of InGaN is chosen for the application, for example, for blue, $E_{g1} = 2.7$ eV. The heterostructure has a number of MQWs to improve the efficiency but the number of QWs is not many, which is limited by the fabrication process.

EXAMPLE 3.12.1 Energy levels in the quantum well

Consider a GaAs QW sandwiched between two $\text{Al}_{0.40}\text{Ga}_{0.60}\text{As}$ layers. Suppose that the barrier height ΔE_c is 0.30 eV, the electron effective mass in the well is $0.067m_e$, and the width of the QW (d) is 12 nm. Calculate the energy levels E_1 and E_2 from the bottom of the well (E_c) assuming an infinite PE well as in Eq. (3.12.1). Compare these with the calculations for a *finite* PE well that give 0.022 eV, 0.088, and 0.186 for $n = 1, 2$, and 3.

Solution

We use Eq. (3.12.1) with $m_e^* = 0.067m_e$, $d = 12 \times 10^{-9}$ nm, so that for $n = 1$

$$\Delta E_n = E_n - E_c = \frac{\hbar^2 n^2}{8m_e^* d^2} = \frac{(6.624 \times 10^{-34} \text{ J s})^2 (1)^2 / (1.602 \times 10^{-19} \text{ J eV}^{-1})}{8(0.067 \times 9.1 \times 10^{-31} \text{ kg})(12 \times 10^{-9} \text{ m})^2} = 0.039 \text{ eV}$$

We can repeat the above calculation for $n = 2$ and 3 to find $\Delta E_2 = 0.156$ eV and $\Delta E_3 = 0.351$ eV. The third level will be above the well depth ($\Delta E_c = 0.3$ eV). Clearly, the infinite QW predicts higher energy levels, by a factor of 1.8, and puts the third level inside the well, not outside. The finite QW calculation is not simple, and involves a numerical solution.²³

²³A similar example is solved for the finite well in M. Fox, *Optical Properties of Solids*, 2nd Edition (Oxford University Press, 2010), pp. 149–152. (This is a highly recommended text on optical properties.)

3.13 LED MATERIALS AND STRUCTURES

A. LED Materials

There are various direct bandgap semiconductor materials that can be readily doped to make commercial *pn* junction LEDs which emit radiation in the red and infrared range of wavelengths. The fabrication of an actual LED would require that the doped crystal layers with the required bandgap can be grown on a suitable **substrate** crystal. The substrate crystal and the LED material will have to be **lattice matched**, that is, they must have the same crystal structure and very close lattice parameters (*a*) to avoid creating dislocations at the interface. (Dislocations and other defects provide a nonradiative recombination path.) Lattice matching is not always perfect and often dictates what material compositions can be used.

An important class of commercial semiconductor materials that cover the visible spectrum is the **III–V ternary alloys** based on alloying GaAs and GaP, which are denoted as $\text{GaAs}_{1-y}\text{P}_y$. In this compound, As and P atoms from group V are distributed randomly at normal As sites in the GaAs crystal structure. When $y < 0.45$, the alloy $\text{GaAs}_{1-y}\text{P}_y$ is a direct bandgap semiconductor and hence the EHP recombination process is direct and efficient. The emitted wavelengths range from about 630 nm, red, for $y = 0.45$ ($\text{GaAs}_{0.55}\text{P}_{0.45}$) to 870 nm for $y = 0$, GaAs.

$\text{GaAs}_{1-y}\text{P}_y$ alloys (which includes GaP) with $y > 0.45$ are indirect bandgap semiconductors. The EHP recombination processes occur through recombination centers and involve lattice vibrations rather than photon emission. As mentioned above, however, if we add isoelectronic impurities such as nitrogen into the semiconductor crystal then the N-dopants can act as recombination centers. An electron is first captured by the N-center, the excess energy is lost to phonons, and then while at the N-center, it recombines with a hole in a radiative transition. The emitted photon energy is only slightly less than E_g . Nitrogen-doped indirect bandgap $\text{GaAs}_{1-y}\text{P}_y$ alloys are widely used in inexpensive green, yellow, and orange LEDs.

There are various commercially important direct bandgap semiconductor materials that emit in the red and infrared wavelengths which are typically **ternary** (containing three elements) and **quaternary** (four elements) alloys based on Group III and V elements, so-called **III–V alloys**. For example, GaAs with a bandgap of about 1.42 eV emits radiation at around 870 nm in the infrared. But ternary alloys based on $\text{Al}_{1-x}\text{Ga}_x\text{As}$ where $x < 0.43$ are direct bandgap semiconductors. The composition can be varied to adjust the bandgap and hence the emitted radiation from about 640–870 nm, from deep red light to infrared.

AlGaInP is a quaternary III–V alloy (In, Ga, Al from III, and P from V) that has a direct bandgap variation with composition over the visible range. It can be lattice-matched to GaAs substrates for compositions $(\text{Al}_x\text{Ga}_{1-x})_{0.5}\text{In}_{0.5}\text{P}$ where $x < 0.53$, that is, $\text{Ga}_{0.50}\text{In}_{0.50}\text{P}$ ($E_g = 1.89$ eV, red) to $\text{Al}_{0.265}\text{Ga}_{0.235}\text{In}_{0.50}\text{P}$ (2.33 eV, green). Many brands of high-intensity LEDs have been based on this material, which is likely to continue to be used in the high-intensity visible LED range, especially for the red, amber, and yellow.

The bandgap of quaternary alloys $\text{In}_{1-x}\text{Ga}_x\text{As}_{1-y}\text{P}_y$ can be varied with composition (*x* and *y*) to span wavelengths from 870 nm (GaAs) to 3.5 μm (InAs) which includes the optical communication wavelengths of 1.3 μm and 1.55 μm . Figure 3.38 summarizes some typical wavelengths that can be emitted for a few selected semiconductor materials over the range from 0.3 μm to 1.7 μm , or from the UV to IR.

GaN is a direct bandgap semiconductor with an E_g of 3.4 eV. The blue GaN LEDs actually use the GaN alloy InGaN with a bandgap of about 2.7 eV, which corresponds to blue emission. One of the most important technological advances in the last two decades has been the development of various III-Nitride LEDs that can emit with high intensities from the UV to green. The alloys

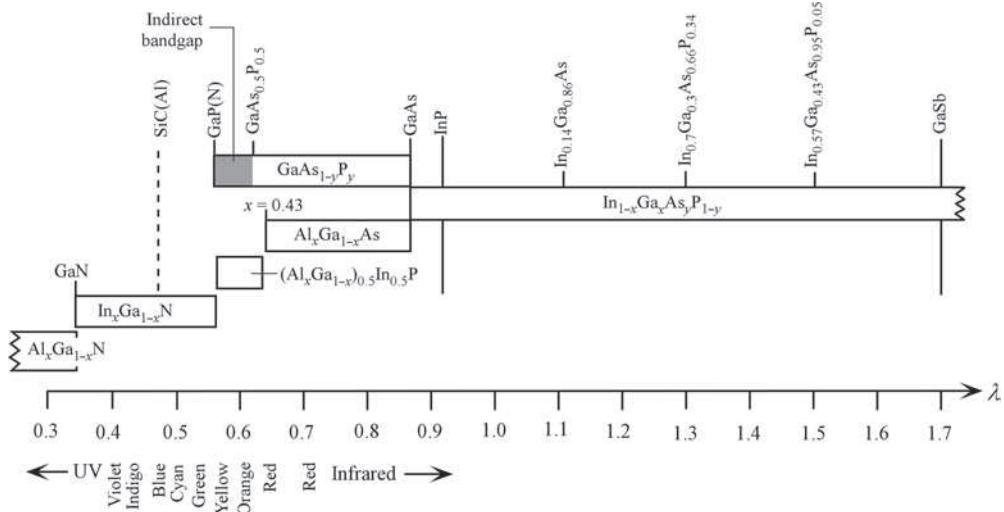


FIGURE 3.38 Free-space wavelength coverage by different LED materials from the visible spectrum to the infrared, including wavelengths used in optical communications. Grey region and dashed lines are indirect E_g materials. Only material compositions of importance have been shown.

of GaN ($E_g = 3.4\text{ eV}$) with InN ($E_g = 0.77\text{ eV}$), $\text{In}_x\text{Ga}_{1-x}\text{N}$ spans wavelengths from the UV up to the IR, though they are currently not used beyond the green wavelength as other semiconductors such as AlGaInP provide better efficiencies. The alloys of AlN ($E_g = 6.2\text{ eV}$) and GaN ($E_g = 3.4\text{ eV}$), AlGaN, provide for wavelengths in the UV. GaN can be doped n -type (e.g., with Si or Ge) and p -type (e.g., with Mg), and the GaN LEDs are generally MQW heterostructures.

B. LED Structures

In its simplest technological form, LEDs are typically fabricated by *epitaxially* growing doped semiconductor layers on a suitable substrate (e.g., GaAs or GaP) as illustrated for a simple homojunction *pn* junction LED in Figure 3.39 (a). This type of planar *pn* junction is formed by the epitaxial growth of first the *n*-layer and then the *p*-layer. The substrate is essentially a mechanical support for the *pn* junction device (the layers) and can be of different crystal. The *p*-side is on the surface from which light is emitted and is therefore made narrow (a few microns)

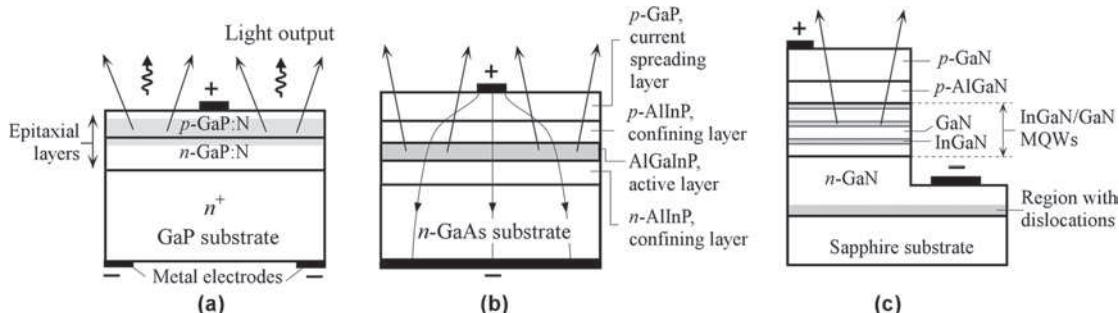


FIGURE 3.39 A schematic illustration of various typical LED structures. (a) A planar surface-emitting homojunction green GaP:N LED. (b) AlGaN high-intensity heterostructure LED. (c) III-Nitride-based (GaN/InGaN) MQW LED for emission from the UV to green.

to allow the photons to escape without being reabsorbed. To ensure that most of the recombination takes place in the *p*-side, the *n*-side is heavily doped (n^+). Those photons that are emitted toward the *n*-side become either absorbed or reflected back at the substrate interface depending on the substrate thickness and the exact structure of the LED. The use of a segmented back electrode as in Figure 3.39 (a) will encourage reflections from the semiconductor–air interface.

If the epitaxial layer and the substrate crystals have different crystal lattice parameters, then there is a lattice mismatch between the two crystal structures. This causes lattice strain in the LED layer and hence leads to crystal defects. Such crystal defects encourage radiationless EHP recombinations. That is, a defect acts as a recombination center. Such defects are reduced by lattice matching the LED epitaxial layer to the substrate crystal. It is therefore important to lattice-match the LED layers to the substrate crystal. For example, $\text{Al}_x\text{Ga}_{1-x}\text{As}$ alloys with $x < 0.43$ are direct bandgap semiconductors that have a bandgap corresponding to the IR to the red emission region (Figure 3.38). $\text{Al}_x\text{Ga}_{1-x}\text{As}$ layers can be grown on GaAs substrates with excellent lattice match, which results in high-efficiency LED devices.

Figure 3.39 (b) shows the structure of a high-intensity AlGaInP heterostructure LED. The layers are grown epitaxially on an *n*-GaAs substrate. There are at least four layers. The active layer is a thin AlGaInP (*e.g.*, $\text{Al}_{0.35}\text{Ga}_{0.15}\text{In}_{0.5}\text{P}$), which is lightly doped. This layer is sandwiched by confining layers that are *p*-type and *n*-type AlInP (*e.g.*, $\text{Al}_{0.5}\text{In}_{0.5}\text{P}$) on the positive and negative terminal sides, respectively. AlInP has a wider bandgap than AlGaInP, and the band offsets confine the carriers to the active region. Remember that under forward bias the *p*-AlInP injects holes and *n*-AlInP injects electrons into the active layer. The top layer is *p*-GaP and serves to spread out the current to regions outside the top contact. Thus, radiative recombinations are avoided right under the top contact from which photons cannot be extracted. Put differently, EHP recombination is spread into regions outside the zone directly under the top contact.

Figure 3.39 (c) shows a simplified III-Nitride-based MQW LED that can be used for emission in the blue as well as green.²⁴ With some modification to compositions, it can also emit in the UV. The *p*-GaN (doped with Mg) is the required *p*-layer used for the injection of holes. The QWs are formed between the narrower E_g InGaN and wider E_g GaN as described above, which are undoped. There is a *p*-AlGaN layer that is called a *buffer* layer. The bandgap of AlGaN is wider than InGaN, so it confines the injected electrons in the QW-region. The *n*-GaN layer is the electron injecting *n*-type semiconductor from which electrons are injected into the MQWs. It is difficult to find a matching crystal substrate for GaN. A sapphire crystal is the most commonly used substrate, though the mismatch is roughly 12% (significant). Special growth techniques have been developed to keep the defects (dislocations) to the initial GaN growth region near the sapphire–GaN interface, away from actual LED heterostructure. Notice that the negative terminal is on high-quality *n*-GaN, away from the defective region.

Not all light rays reaching the semiconductor–air interface, however, can escape because of total internal reflection (TIR). Those rays with angles of incidence greater than the critical angle θ_c become reflected as illustrated in Figure 3.40 (a). For the GaAs–air interface, for example, θ_c is only 17°, which means that much of the light suffers TIR. It is possible to shape the surface of the semiconductor into a dome, or hemisphere, so that light rays strike the surface at angles less than θ_c and therefore do not experience TIR. The main drawback, however, is the additional difficult

²⁴III-Nitride MQW LEDs are currently the most efficient commercial LEDs in the blue, violet, and the UV. An extensive review can be found in Hadis Morkoç, *Handbook of Nitride Semiconductors and Devices*, Vol. 3 (Wiley-VCH, Weinheim, 2008), Ch. 1.

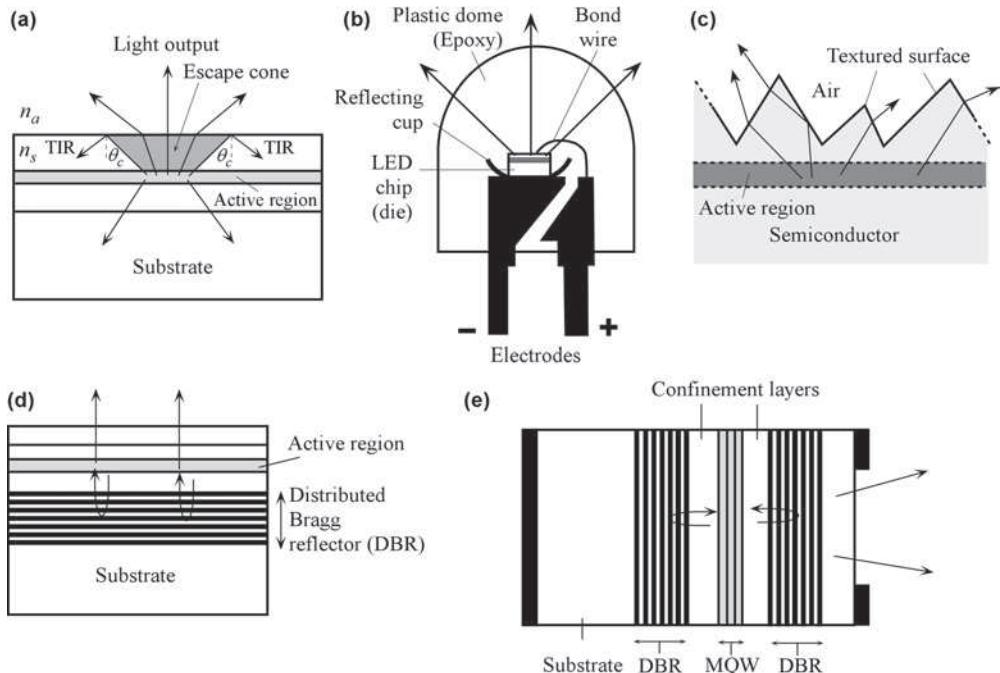


FIGURE 3.40 (a) Some of the internally generated light suffers total internal reflection (TIR) at the semiconductor-air interface and cannot be emitted into the outside. (b) A simple structure that overcomes the TIR problem by placing the LED chip at the centre of a hemispherical plastic dome. The epoxy is refractive index matched to the semiconductor and the rays reaching the dome's surface do not suffer TIR. (c) An example of a textured surface that allows light to escape after a couple of (or more) reflections (highly exaggerated sketch). (d) A distributed Bragg reflector (DBR) under the confining layer (below the active region in grey) acts as a dielectric mirror, and increases the extraction ratio. (e) An RCLED is an LED with an optical resonant cavity (RC) formed by two DBRs has a narrower emission spectrum.

process in fabricating such domed LEDs and the associated increase in expense. An inexpensive and common procedure that reduces TIR is the encapsulation of the semiconductor junction within a transparent plastic medium (an epoxy) which has a higher refractive index than air and, further, also has a domed surface on the emission side of the LED chip as shown in Figure 3.40 (b). Many individual LEDs are sold in similar types of plastic bodies.

Shuji Nakamura obtained his PhD from the University of Tokushima in Japan, and is currently a professor at the University of California at Santa Barbara and the director of Solid State Lighting and Energy Center. He has been credited with the pioneering work that has led to the development of GaN and $In_xGa_{1-x}N$ -based blue and violet light-emitting diodes and laser diodes. He discovered how III-Nitrides could be doped *p*-type, which opened the way to fabricating various UV, violet, blue, and green LEDs. He is holding a blue laser diode that is turned on. (*Courtesy of Shuji Nakamura, University of California, Santa Barbara.*)



Another example of a device structure that improves the light extraction ratio is shown in Figure 3.40 (c). The surface has been textured or nanostructured. Such a textured surface allows light to escape after one or two reflections. The extraction ratio can be greatly improved if the rays traveling toward the substrate could be somehow reflected back toward the exit window. In some LEDs, a *distributed Bragg reflector* (DBR), that is a dielectric mirror, under the confining layer (below the active region) acts as a dielectric mirror at the wavelength of the LED light, and increases the extraction ratio, as shown in Figure 3.40 (d).

Another possible design is to use a *resonant cavity* for the light generated by the active region as shown in Figure 3.40 (e). An RCLED (resonant cavity LED) is an LED with an optical resonant cavity (RC) formed by two DBRs. The optical cavity is obviously wavelength-selective since only those special modes of the optical cavity that fall into the spontaneous emission spectrum can be supported or excited. The width of the emitted spectrum is significantly narrower, as much as ten times, than a conventional LED. The reason is that the light from the active region can only escape if it is a mode of the resonant cavity. The spectral width of this light is determined by the resonant cavity losses rather than the emission spectrum from the active region.

EXAMPLE 3.13.1 Light extraction from a bare LED chip

As shown in Figure 3.40 (a), due to total internal reflection (TIR) at the semiconductor–air surface, only a fraction of the emitted light can escape from the chip. The critical angle θ_c is determined by $\sin \theta_c = n_a/n_s$ where n_a and n_s are the refractive indices of the ambient (e.g., for air, $n_a = 1$) and the semiconductor, respectively. The light within the **escape cone** defined by θ_c can escape into the ambient without TIR as indicated in Figure 3.40 (a). To find the fraction of light within the escape cone we need to consider solid angles, which leads to $(1/2)(1 - \cos \theta_c)$. Further, suppose that T is the average light transmittance of the n_s – n_a interface for those rays within the escape cone, then for a simple bare chip,²⁵

$$\text{Light extraction ratio} \approx (1/2)(1 - \cos \theta_c) \times T \quad (3.13.1)$$

Light
extraction
ratio

Estimate the extraction ratio for a GaAs chip with $n_s = 3.4$ and air as ambient ($n_a = 1$) and then with an epoxy dome with $n_a = 1.8$.

Solution

First, note that $\theta_c = \arcsin(n_a/n_s) = \arcsin(1/3.4) = 17.1^\circ$. For T we will assume near-normal incidence (somewhat justified since the angle 17.1° is not too large) so that from Chapter 1,

$$T = 4n_s n_a / (n_s + n_a)^2 = 4(3.4)(1)/(3.4 + 1)^2 = 0.702$$

Using Eq. (3.13.1)

$$\begin{aligned} \text{Light extraction ratio} &\approx (1/2)(1 - \cos \theta_c) \times T = (1/2)[1 - \cos(17.1^\circ)] \times 0.702 \\ &\approx 0.0155 \quad \text{or} \quad 1.6\% \end{aligned}$$

It is clear that only 1.6% of the generated light power is extracted from a bare chip, which is disappointingly small. The technological drive is therefore to improve light extraction as much as possible. If we now repeat the calculation for $n_a = 1.8$, we would find, $\theta_c = 32^\circ$, and 6.9% light extraction.

²⁵ An extensive treatise on various LED structures and light extraction techniques can be found in E. F. Schubert, *Light-Emitting Diodes*, 2nd Edition (Cambridge University Press, 2006). Equation (3.13.1) is an approximation inasmuch as we should average the transmittance.

3.14 LED EFFICIENCIES AND LUMINOUS FLUX

The comparison of different LED materials and device structures requires that we define internal and external quantum efficiency. Further, we also need to meaningfully compare the output power and brightness of different light-emitting devices per unit electrical energy input.

The **internal quantum efficiency** (IQE) η_{IQE} gauges what fraction of electron hole recombinations in the forward-biased *pn* junction are radiative and therefore lead to photon emission. Nonradiative transitions are those in which an electron and a hole recombine through a recombination center such as a crystal defect or an impurity and emit phonons (lattice vibrations). Suppose that τ_r is the mean lifetime of a minority carrier before it recombines radiatively and τ_{nr} is the mean lifetime before it recombines via a recombination center (or a defect) without emitting a photon. By definition,

$$\begin{array}{l} \text{Internal} \\ \text{quantum} \\ \text{efficiency} \end{array} \quad \eta_{\text{IQE}} = \frac{\text{Rate of radiative recombination}}{\text{Total rate of recombination (radiative and nonradiative)}} \quad (3.14.1)$$

or

$$\begin{array}{l} \text{Internal} \\ \text{quantum} \\ \text{efficiency} \end{array} \quad \eta_{\text{IQE}} = \frac{\tau_r^{-1}}{\tau_r^{-1} + \tau_{nr}^{-1}} \quad (3.14.2)$$

The rate of nonradiative recombination rate ($1/\tau_{nr}$) would also include the recombination of injected carriers at the interfaces of heterostructures between different crystals: an important loss mechanism that would reduce η_{IQE} .

The total current I is determined by the total rate of recombinations whereas the number of photons emitted per second, the **photon flux** Φ_{ph} , is determined by the rate of radiative recombinations.

$$\begin{array}{l} \text{Internal} \\ \text{quantum} \\ \text{efficiency} \end{array} \quad \eta_{\text{IQE}} = \frac{\text{Photons emitted per second}}{\text{Total carriers lost per second}} = \frac{\Phi_{\text{ph}}}{I/e} = \frac{P_{o(\text{int})}/h\nu}{I/e} \quad (3.14.3)$$

where $P_{o(\text{int})}$ is the *optical* power generated internally (not yet extracted).

The **external quantum efficiency** (EQE) η_{EQE} of an LED represents the efficiency of conversion from electrical quanta, *i.e.*, electrons, that flow into the LED to optical quanta, *i.e.*, photons, that are emitted into the outside world. It incorporates the “internal” efficiency of the radiative recombination process [embedded in Eq. (3.14.1)] and the subsequent efficiency of photon extraction from the device. Suppose that the actual optical power emitted to the ambient, called the **radiant flux**, is P_o . (Φ_e is also used in the literature.) $P_o/h\nu$ is the number of emitted photons per second. Since the number of electrons flowing into the LED is I/e , we have

$$\begin{array}{l} \text{External} \\ \text{quantum} \\ \text{efficiency} \end{array} \quad \eta_{\text{EQE}} = \frac{P_o/h\nu}{I/e} \quad (3.14.4)$$

Thus, we should use external QE to meaningfully compare different LED efficiencies and internal QE in comparing different LED materials. For indirect bandgap semiconductors η_{EQE} are generally less than 1% whereas for direct bandgap semiconductors with the right device structure, η_{EQE} can be substantial, for example, 30–40%.

The light extraction ratio was introduced in Example 3.13.1. We can now define this more formally. The light **extraction ratio**, or the **extraction efficiency** (EE), η_{EE} , is the fraction of light that is extracted to the ambient from the internally generated light, that is,

$$\eta_{\text{EE}} = \frac{\text{Photons emitted externally from the device}}{\text{Photons generated internally by recombination}} \quad (3.14.5)$$

Extraction efficiency

Using Eqs. (3.14.1), (3.14.3) and (3.14.5), the emitted optical output power, the radiant flux, is

$$P_o = \eta_{\text{EE}} P_{o(\text{int})} = h\nu\eta_{\text{EE}}\eta_{\text{IQE}}(I/e) \quad (3.14.6)$$

Emitted optical power

The **power conversion efficiency** (PCE), η_{PCE} , or simply the **power efficiency**, gauges the overall efficiency of conversion from the input of electrical power to the output of optical power, *i.e.*,

$$\eta_{\text{PCE}} = \frac{\text{Optical output power}}{\text{Electrical input power}} = \frac{P_o}{IV} \approx \eta_{\text{EQE}} \left(\frac{E_g}{eV} \right) \quad (3.14.7)$$

Power efficiency

If we wish to compare the brightness of light-emitting devices per unit electrical power input then we need to examine the luminous flux emitted. The visual “brightness” of a source as observed by an average daylight-adapted eye is proportional to the radiation (optical) power emitted, that is, the radiant flux, and the efficiency of the eye to detect the spectrum of the emitted radiation. While the eye can see a red color source, it cannot see an infrared source and the brightness of the infrared source would be zero. The **luminous flux** Φ_v is a measure of *visual brightness*, in lumens (lm), and is defined by

$$\Phi_v = P_o \times (633 \text{ lm W}^{-1}) \times V(\lambda) \quad (3.14.8)$$

Luminous flux

where P_o is the radiant flux or the radiation power emitted (in watts) and $V(\lambda)$ is the **relative luminous efficiency** (or the relative sensitivity) of an average light-adapted (photopic) eye, which depends on the wavelength and hence λ in parenthesis. The function $V(\lambda)$ is also called the **luminosity function** and the **visibility function**. $V(\lambda)$ is a Gaussian-like function with a peak of unity at 555 nm as shown in Figure 3.41. One lumen of luminous flux, or brightness, is obtained from a 1.58 mW light source emitting at a single wavelength of 555 nm (green). A typical 60 W incandescent lamp provides roughly 900 lm (or 15 lm W⁻¹). When we buy a light bulb, we are buying lumens.

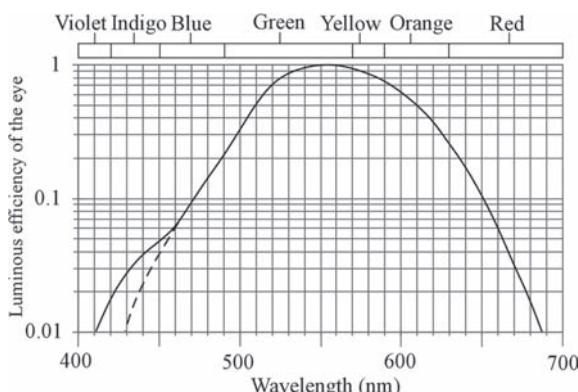


FIGURE 3.41 The luminous efficiency $V(\lambda)$ of the light-adapted (photopic) eye as a function of wavelength. The solid curve is the Judd-Vos modification of the CIE 1924 photopic photosensitivity curve of the eye. The dashed line shows the modified region of the original CIE 1924 curve to account for its deficiency in the blue-violet region. (The vertical axis is logarithmic.)

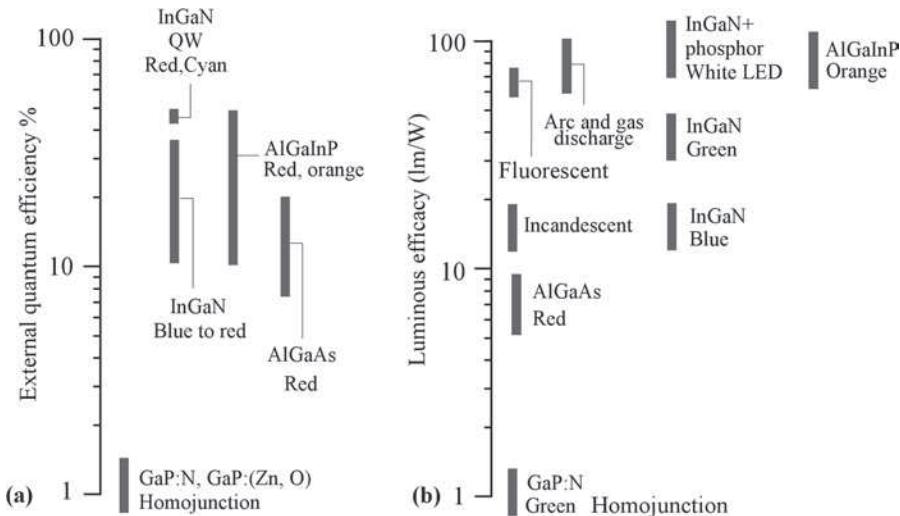


FIGURE 3.42 Typical (a) external quantum efficiency and (b) luminous efficacy of various selected LEDs, and how they stand against other light sources such as the fluorescent tube, arc, and gas discharge lamps and the incandescent lamp.

The **luminous efficacy²⁶ of a light source** (such as a lamp) as commonly used in *lighting applications* is the efficiency with which an electrical light source converts the input electric power (watts) into an emitted luminous flux (lumens).

$$\eta_{LE} = \frac{\Phi_v}{IV} \quad (3.14.9)$$

A 100 W light bulb producing 1700 lumens has an efficacy of 17 lumens watt⁻¹ (lm W⁻¹). Recent technological advances have led to LEDs with efficacies that are comparable to standard fluorescent tubes—efficiencies around 100 lm W⁻¹. LEDs as solid state lamps have much longer lifetimes and much higher reliability, and hence are expected to be more economical than incandescent and fluorescent lamps. Figure 3.42 (a) and (b) show some typical external efficiency and luminous efficacy values of various selected LEDs. The advantages of III-Nitrides (*e.g.*, InGaN) are clearly obvious. It is also important to note that in some cases, as in InGaN based LEDs, the high EQE and hence luminous efficacy cannot be maintained as the current is increased to obtain higher luminous flux (more lumens) because the output flux is not directly proportional to the input current, especially at high currents, as apparent in Figure 3.37 (a).

EXAMPLE 3.14.1 LED efficiencies

A particular 870 nm IR LED for use in optical links and instrumentation has a GaAs chip. Active layer that has been doped *p*-type with $2 \times 10^{17} \text{ cm}^{-3}$ of acceptors and the nonradiative lifetime is about 100 ns. At a forward current of 30 mA, the voltage across it is 1.35 V, and the emitted optical power is 6.5 mW. Calculate the IQE, EQE, and PCE, and estimate the light extraction ratio. For GaAs, $B \approx 2 \times 10^{-16} \text{ m}^2 \text{ s}^{-1}$.

²⁶Some authors use the term *luminous efficiency* but the latter, strictly, needs the output and input quantities to have the same units so that the efficiency can be expressed as a percentage, which is not the case here. *Efficacy* would be a better term.

Solution

The radiative lifetime $\tau_r = 1/BN_a = 1/\left[(2 \times 10^{-16} \text{ m}^3 \text{ s}^{-1})(2 \times 10^{23} \text{ m}^{-23})\right] = 2.5 \times 10^{-8} \text{ s}$ or 25 ns. IQE is,

$$\eta_{\text{IQE}} = \frac{\tau_r^{-1}}{\tau_r^{-1} + \tau_{nr}^{-1}} = \frac{(25 \text{ ns})^{-1}}{(25 \text{ ns})^{-1} + (100 \text{ ns})^{-1}} = 0.80 \quad \text{or} \quad 80\%$$

The emitted photon energy $h\nu = hc/\lambda \approx 1.43 \text{ eV}$. Thus the EQE is

$$\begin{aligned}\eta_{\text{EQE}} &= (P_o/h\nu)/(I/e) \\ &= [(6.5 \times 10^{-3} \text{ W})/(1.43 \text{ eV} \times 1.6 \times 10^{-19} \text{ J eV}^{-1})]/[(30 \times 10^{-3} \text{ A})/(1.6 \times 10^{-19} \text{ C})] \\ &= 0.15 \quad \text{or} \quad 15\%\end{aligned}$$

The PCE is simply P_o/IV or $(6.5 \text{ mW})/[(30 \text{ mA})(1.35 \text{ V})]$, that is, 0.16, *i.e.*, 16%.

From Eq. (3.14.6), using $P_o = h\nu\eta_{\text{EE}}\eta_{\text{IQE}}(I/e)$,

$$6.5 \times 10^{-3} \text{ W} = (1.43 \text{ eV} \times 1.6 \times 10^{-19} \text{ J eV}^{-1})\eta_{\text{EE}}(0.80)(30 \times 10^{-3} \text{ A}/1.6 \times 10^{-19} \text{ C})$$

solving the above gives $\eta_{\text{EE}} = 0.19$ or 19%.

EXAMPLE 3.14.2 LED brightness

Consider two LEDs, one red, with an optical output power (radian flux) of 10 mW, emitting at 650 nm, and the other, a weaker 5 mW green LED, emitting at 532 nm. Find the luminous flux emitted by each LED.

Solution

For the red LED, at $\lambda = 650 \text{ nm}$, Figure 3.41 gives $V \approx 0.10$ so that from Eq. (3.14.8)

$$\Phi_v = P_o \times (633 \text{ lm W}^{-1}) \times V = (10 \times 10^{-3} \text{ W})(633 \text{ lm W}^{-1})(0.10) = 0.63 \text{ lm}$$

For the green LED, $\lambda = 532 \text{ nm}$, Figure 3.41 gives $V \approx 0.87$ so that from Eq. (3.14.8)

$$\Phi_v = P_o \times (633 \text{ lm W}^{-1}) \times V = (5 \times 10^{-3} \text{ W})(633 \text{ lm W}^{-1})(0.87) = 2.8 \text{ lm}$$

Clearly the green LED at half the optical power is 4 times brighter than the red LED.

3.15 BASIC LED CHARACTERISTICS

The I - V characteristics of LEDs depend on the device structure, including the material properties of the semiconductors used in the LED. The DC current generally increases steeply with the voltage, as shown in Figure 3.43 (a). The I - V characteristics do not always follow an exact exponential behavior that is often seen for a simple forward-biased pn junction. Most modern LEDs are heterostructure devices, with several layers of differently doped semiconductors. The current increases sharply over a narrow voltage range, as in Figure 3.43 (a), where it can be seen that there is an apparent **turn-on** or a **cut-in voltage** beyond which the current increases very sharply with voltage. Most LED manufacturers, however, quote the **forward voltage** V_F when the LED is operating fully; for example, the manufacturer of one type of red AlGaInP-based LED quotes $V_F = 2.0 \text{ V}$ at $I = 20 \text{ mA}$, at 40% rated maximum current. V_F depends on both the semiconductor material and the device structure. There is an overall trend in which V_F tends to increase with decreasing wavelength, that is, with increasing photon energy or roughly the bandgap E_g of the active region. However, the device structure is also very important, and can skew this

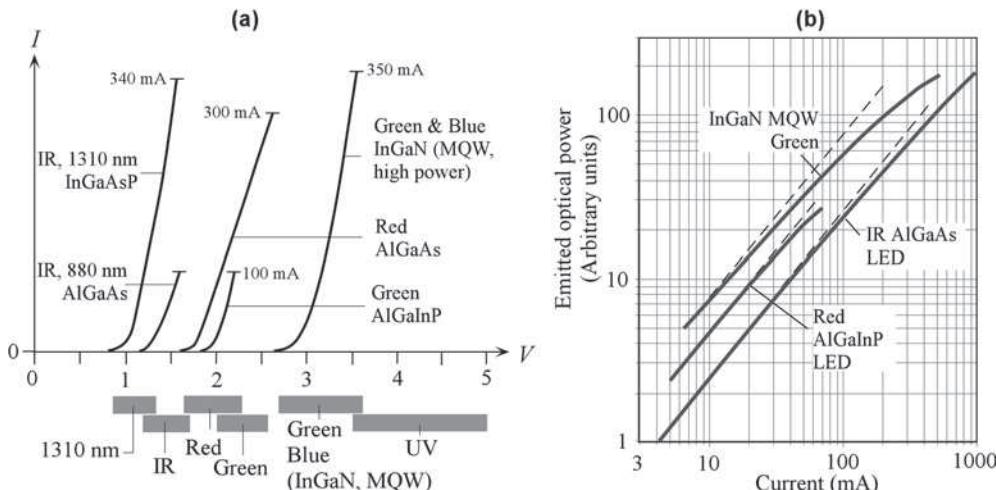


FIGURE 3.43 (a) Current-voltage characteristics of a few LEDs emitting at different wavelengths from the IR to blue. (b) Log-log plot of the emitted optical output power vs. DC current for three commercial devices emitting at IR (890 nm), red and green. The vertical scale is in arbitrary units and the curves have been shifted to show the dependence of P_o on I . The ideal linear behavior $P_o \propto I$ is also shown.

observation. MQW InGaN LEDs have a higher V_F than corresponding AlGaInP heterojunction LEDs, that is, they operate at a higher voltage across the device as apparent for the two green LEDs in Figure 3.43 (a). Indeed, InGaN green to violet LEDs all have very similar V_F .

Typical optical output power (P_o) vs. current (I) characteristics of LEDs are shown in Figure 3.43 (b) on a log-log plot for three cases. For comparison, the expected linear relationship, $P_o \propto I$, has been also shown for each device. In general, at high currents, P_o vs. I relationship curves down from the expected linear, $P_o \propto I$, behavior. The worst case is for InGaN MQW LEDs in which there is significant deviation from the expected linear relationship almost from the start, that is, P_o cannot keep up linearly with the current and droops as the current increases. This droop represents a drop in the quantum efficiency of the MQW InGaN device, and is currently a topical area of research. The P_o - I characteristics for standard AlGaInP and AlGaAs heterojunction LEDs deviate from linearity mainly at high currents, exhibiting an extensive range of reasonable linearity. While a non-linear P_o - I behavior is not a serious problem in digital communications, it can create distortion in analog modulation, especially under large signals.

3.16 LEDs FOR OPTICAL FIBER COMMUNICATIONS

The type of light source suitable for optical communications depends not only on the communication distance but also on the bandwidth requirement. For short-haul applications, for example, local networks, LEDs are preferred as they are simpler to drive, more economic, have a longer lifetime, and provide the necessary output power even though their output spectrum is much wider than that of a laser diode. LEDs are frequently used with graded index fibers inasmuch as typically dispersion in a graded index fiber is primarily due to intermodal dispersion rather than intramodal dispersion. (For example, at 1310 nm communications, the material and hence chromatic dispersion in a graded index fiber is usually much smaller than intermodal dispersion.) For long-haul and wide-bandwidth communications, invariably laser diodes are used because of their narrow linewidth, high output power, and higher signal bandwidth capability.

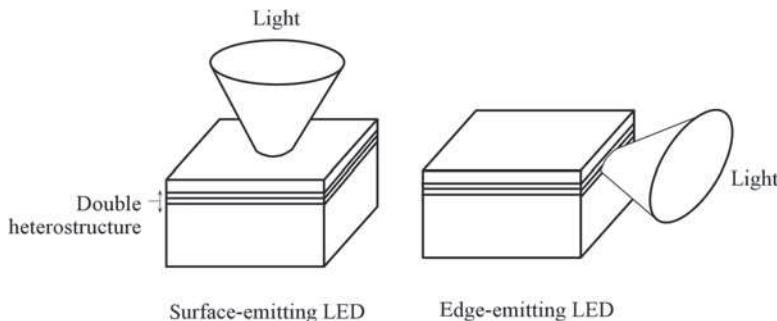


FIGURE 3.44
A surface-emitting and
an edge-emitting LED.

There are essentially two types of LED devices which are illustrated in Figure 3.44. If the emitted radiation emerges from an area in the plane of the recombination layer as in (a) then the device is a **surface-emitting LED** (SLED). If the emitted radiation emerges from an area on an edge of the crystal as in (b), that is, from an area on a crystal face perpendicular to the active layer, then the LED is an **edge-emitting LED** (ELED).

The simplest method of coupling the radiation from a surface-emitting LED into an optical fiber is to etch a well in the planar LED structure and lower the fiber into the well as close as possible to the active region where emission occurs. This type of structure, as shown in Figure 3.45 (a), is called a **Burrus-type device** (after its originator). An epoxy resin is used to bond the fiber and provide refractive index matching between the glass fiber and the LED material to capture as much of the light rays as possible. Note that in the double heterostructure LED used in this way, the photons emitted from the active region, with a smaller bandgap, do not get absorbed by the neighboring layer, which has a wider bandgap. Another method is to use a truncated **spherical lens** (a microlens) with a high refractive index ($n = 1.9 - 2$) to focus the light into the fiber as shown in Figure 3.45 (b). The lens is bonded to the LED with a refractive index-matching cement and, in addition, the fiber can be bonded to the lens with a similar cement.

Edge-emitting LEDs provide a greater intensity light and also a beam that is more collimated than the surface-emitting LEDs. Figure 3.46 shows the structure of a typical edge-emitting LED for operation at $\sim 1.5 \mu\text{m}$. The light is guided to the edge of the crystal by a **dielectric waveguide** formed by wider bandgap semiconductors surrounding a double heterostructure. The recombination of injected carriers occurs in the InGaAs active region which has a bandgap $E_g \approx 0.83 \text{ eV}$. Recombination is confined to this layer because the surrounding InGaAsP layers, **confining layers**, have a wider bandgap ($E_g \approx 1 \text{ eV}$) and the InGaAsP/InGaAs/InGaAsP layers form a

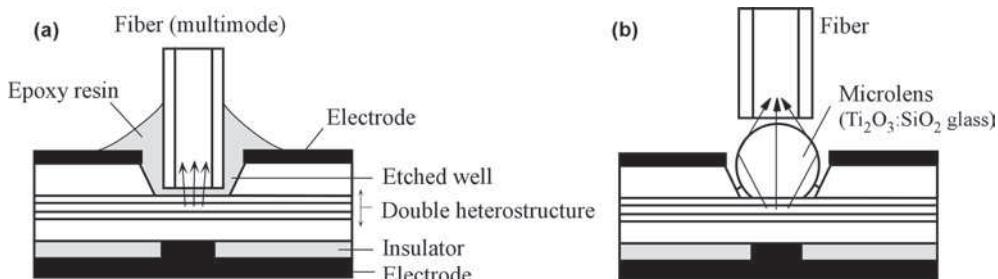


FIGURE 3.45 Coupling of light from surface-emitting LEDs into optical fibers. (a) Light is coupled from a surface-emitting LED into a multimode fiber using an index matching epoxy. The fiber is bonded to the LED structure. (b) A microlens focuses diverging light from a surface-emitting LED into a multimode optical fiber.

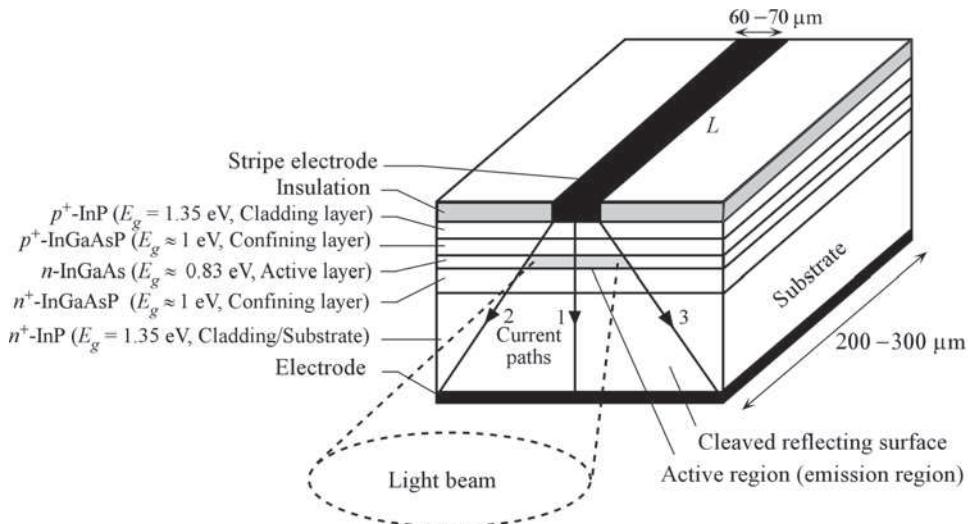


FIGURE 3.46 Schematic illustration of the structure of a double heterojunction stripe contact edge-emitting LED. (Upper case notation for a wider bandgap semiconductor is not used as there are several layers with different bandgaps.)

double heterostructure. The light emitted in the active region (InGaAs) spreads into the neighboring layers (InGaAsP) which guide it along the crystal to the edge. InP has a wider bandgap ($E_g \approx 1.35$ eV) and thus a lower refractive index than InGaAsP. The two InP layers adjoining the InGaAsP layers therefore act as **cladding layers** and thereby confine the light to the DH structure.

Generally some kind of lens system is used to conveniently couple the emitted radiation from an ELED into a fiber. For example, in Figure 3.47 (a), a hemispherical lens attached to the fiber end is used for collimating the beam into the fiber. A **graded index (GRIN) rod lens** is a glass rod that has a parabolic refractive index profile across its cross-section with the maximum index on the rod axis. It is like a large diameter short length graded index “fiber” (typical diameters are 0.5–2 mm). A GRIN rod lens can be used to focus the light from an ELED into a fiber as illustrated in Figure 3.47 (b). This coupling is particularly useful for single mode fibers inasmuch as their core diameters are typically ~ 10 μm .

The output spectra from surface and edge-emitting LEDs using the same semiconductor material are not necessarily the same. The first reason is that the active layers have different doping levels. The second is the self-absorption of some of the photons guided along the active layer as in the ELED. Typically the linewidth of the output spectrum from an ELED is *less* than that from a SLED. For example, in a particular set of experiments on an InGaAsP ELED operating near 1300 nm, the emission linewidth was reported as 75 nm whereas the corresponding SLED at the same wavelength had a linewidth of 125 nm, which is significantly wider.

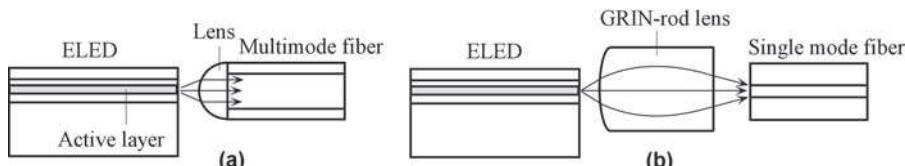


FIGURE 3.47 Light from an edge-emitting LED is coupled into a fiber typically by using a lens or a GRIN rod lens.

3.17 PHOSPHORS AND WHITE LEDs

Photoluminescence is the emission of light by a material, called a **phosphor**, that has been first excited by light of higher frequency; higher energy photons are first absorbed, and then lower energy photons are emitted. Typically the emission of light occurs from certain dopants, impurities, or even defects, called luminescent or **luminescence centers**, purposefully introduced into a **host matrix**, which may be a crystal or glass. The luminescent center is also called an **activator**. Many phosphors are based on activators doped into a host matrix; for example, Eu^{3+} (europium ion) in a Y_2O_3 (yttrium oxide) matrix is a widely used modern phosphor. When excited by UV radiation, it provides an efficient luminescence emission in the red (around 613 nm). It is used as the red-emitting phosphor in color TV tubes and in modern tricolor fluorescent lamps. Another important phosphor is Ce^{3+} in $\text{Y}_3\text{Al}_5\text{O}_{12}$ (YAG), written as $\text{Y}_3\text{Al}_5\text{O}_{12}:\text{Ce}^{3+}$, which is used in white LEDs. YAG:Ce³⁺ can absorb blue radiation and emit yellow light.

In very general terms, we can represent the energy of an activator in a host matrix by the highly simplified energy diagram in Figure 3.48 (a). We can take this figure to very roughly represent the energy of Ce³⁺ in YAG. Although it looks complicated, there are actually two manifolds of energies involved in blue absorption and yellow emission. There is a manifold of energies, labeled E_1 , starting at E_1' , which represents the ground energy of the activator. There is a higher manifold of energies, labeled E_2 , starting at E_2' . A manifold such as E_1 has a number of energy levels that are closely spaced. The energy levels within the E_1 - and E_2 -manifolds represent the vibrational energies of the Ce³⁺-ion in the host (YAG).²⁷ The slight horizontal shifts in the energy levels in a manifold in Figure 3.48 (a) indicate the relative position (or the tiny little displacement) of the Ce³⁺ ion in the host. Radiative transitions (absorption or emission of a photon) occur quickly compared with the time scale of vibrations of the Ce³⁺-ion so that, during the radiative transition, the Ce³⁺ ion is essentially stationary. Thus, these radiative transitions are vertical lines in the activator energy diagram in Figure 3.48 (a); this rule is called the **Franck–Condon principle**.

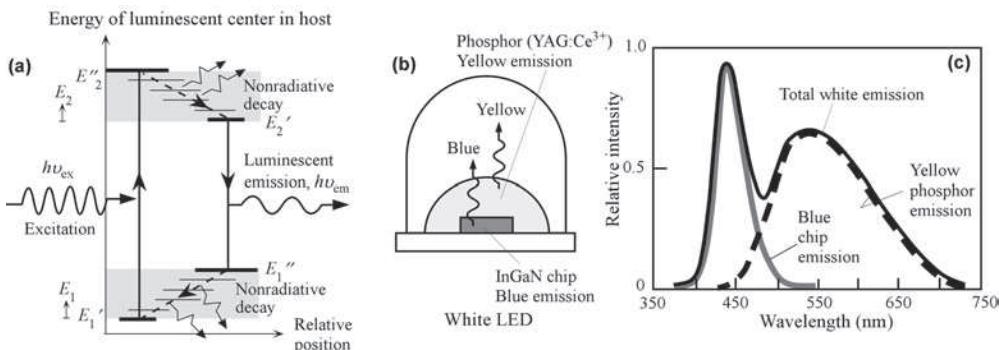


FIGURE 3.48 (a) A simplified energy diagram to explain the principle of photoluminescence. The activator is pumped from E_1' to E_2'' . It decays nonradiatively down to E_2' . The transition from E_2' down to E_1'' emits a photon with a lower energy than the excitation photon. (b) Schematic structure of a blue chip yellow phosphor white LED (c) The spectral distribution of light emitted by a white LED. Blue luminescence is emitted by GaInN chip and “yellow” phosphorescence is produced by phosphor. The combined spectrum looks “white.”

²⁷The energy diagram in Figure 3.48 (a) is still oversimplified. The vibrations of the Ce³⁺ ion is quantized and these energy levels are added to the usual electronic energy levels. However, there is also a dependence on the position of the Ce³⁺ ion, which has been simplified and represented as slightly shifted energy levels along the horizontal direction.

Upon excitation by an incident radiation of suitable energy $h\nu_{\text{ex}}$, the activator becomes excited to E_2'' in the E_2 -manifold; a vertical transition. From this energy level, it decays, or *relaxes*, down relatively quickly (on a time scale of the order of picoseconds) to an energy level E_2' by emitting phonons or lattice vibrations. This type of decay is called **radiationless** or **non-radiative decay**. The Ce^{3+} ion becomes slightly displaced as shown in Figure 3.48 (a). From E_2' , the activator decays down to E_1' in the E_1 -manifold by emitting a photon by spontaneous emission, which is the emitted luminescent radiation. This is a vertical transition (Franck–Condon principle). The emitted photon energy is $h\nu_{\text{em}}$, which is less than the excitation photon energy $h\nu_{\text{ex}}$. The return from E_1' to the ground E_1 state involves phonon emissions. In some activators, the higher levels may form multilevel narrow energy “bands.” In this example, the activator absorbed the incident radiation, and was directly excited, which is known as **activator excitation**. The Ce^{3+} ions in $\text{Y}_2\text{Al}_5\text{O}_{12}:\text{Ce}^{3+}$ can be excited directly by blue light, and would then emit in the yellow region as discussed below. It is apparent from Figure 3.48 (a) that the emitted radiation ($h\nu_{\text{em}}$) has a *longer* wavelength than the exciting radiation ($h\nu_{\text{ex}}$), that is, $h\nu_{\text{em}} < h\nu_{\text{ex}}$. The downshift in the light frequency from absorbed to emitted radiation is called the **Stokes shift**. It should be emphasized that the energy levels of the activator also depend on the host, because the internal electric fields within the host crystal act on the activator, and shift these levels up and down. The emission characteristics depend firstly on the activator, and secondly on the host.

Commercially available and popular white LEDs can provide bright light at a fraction of input electric power compared with conventional tungsten-based incandescent lights. In addition, the white LED has an extremely long lifetime. Most common white LEDs use a chip emitting at a short wavelength (blue or violet) and a *phosphor*, which absorbs some of the light from the blue diode and undergoes secondary luminescence emissions at a longer wavelength (yellow) as illustrated in Figure 3.48 (b). The mixture of blue light from the chip and yellow light from the phosphor results in an overall spectrum, shown in Figure 3.48 (c), that appears white to the eye. The quality and spectral characteristics of the combined emission in Figure 3.48 (c) vary with different white LED designs, for example, chip and phosphor combination.

Typically, a phosphor is composed of an inorganic host substance that contains optically active dopants; the dopants are the atoms that emit the lower energy photons, that is, the activators. Yttrium aluminum garnet (YAG) is a common host material. For white LED applications, it is usually doped with one of the rare-earth elements or a rare-earth compound. Cerium (Ce) is a common dopant element in YAG phosphors designed for white light-emitting diodes in



LUXEON Rebel ES white-emitting LED. (Courtesy of Philips Lumileds.)



Snap LED emitting in the amber for automotive signaling applications. The lamp includes the driver under the LED. (Courtesy of Philips Lumileds.)



The Audi A8 uses LEDs for nearly all its lighting, including headlights. (Used with permission of Audi of America, Inc.)

so-called **blue chip and yellow phosphor white LEDs**. White LEDs are now challenging the existing incandescent sources for general lighting.

In applications requiring a full spectrum of colors from a single point source, red, green, and blue (RGB)-emitting chips in a single package are used, which allows the generation of white light or any of 256 colors by utilizing circuits that drive the three diodes independently.

Additional Topics

3.18 LED ELECTRONICS

The electronics involved in driving LEDs depends on the application, which generally falls into three categories: (a) DC operation, (b) modulation of the LED output light intensity about some DC value in analog applications, and (c) pulsed operation in digital applications, including optical communications. In DC operation, the LED is driven by providing the necessary current for the required optical output. The I - V characteristics is such that large changes in the current correspond to small voltage changes across the LED as in Figure 3.43 (a), and for a given LED, V_F does not vary very much. The simplest circuit is shown in Figure 3.49 (a), where a DC voltage $V (> V_F)$ provides the current, which is determined by $I_F = (V - V_F)/R$. The problem is that keeping I_F constant depends on keeping V constant. Most LEDs have low reverse breakdown voltages and usually another diode, such as D , is used in parallel to shunt the LED when a reverse voltage is accidentally applied to the LED.

A better circuit than that in Figure 3.49 (a) should provide the necessary drive current I_F to the LED, with some convenient way of controlling I_F . Bipolar junction transistors (BJTs) are commonly used in many applications as convenient current sources. Figure 3.49 (b) is shown as an example in which the LED is in the collector circuit of the BJT, which is driven by the output of an op amp A . The resistor R_F in the emitter of the BJT provides feedback, $I_F R_F$, to the negative input of A , which stabilizes the circuit. It is straightforward to show that $I_F \approx V/R_F$, that is, the circuit provides a linear control of I_F by the input voltage V . There are many ICs available that will perform the function of providing the necessary constant current under a wide range of input

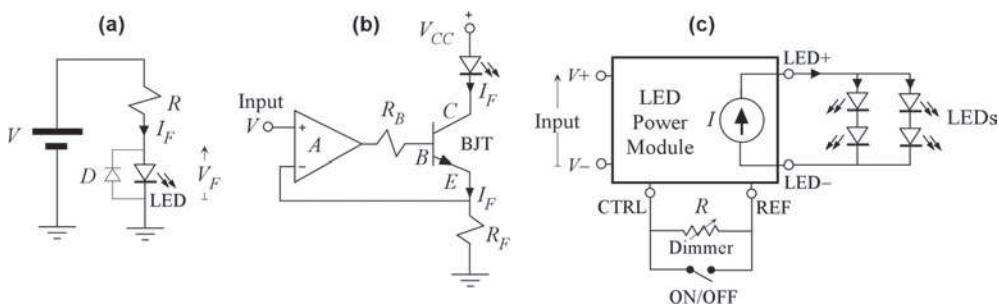


FIGURE 3.49 (a) The simplest circuit to drive an LED involves connecting it to a voltage supply (V) through a resistor R . (b) Bipolar junction transistors are well suited for supplying a constant current. Using an IC and negative feedback, the current is linearly controlled by V . (c) There are various commercial LED driver modules that can be easily configured to drive a number of LEDs in parallel and/or series. The example has a module driving four LEDs, a dimmer (R), and an on/off switch.

voltages, and temperature variations. Without a proper control of the current, the temperature variations can cause substantial changes in the current and hence the light output. Figure 3.49 (c) uses one of several LED driver modules available in the market, which can drive more than one LED in various parallel and series combinations, provided the maximum current rating is satisfied. The current can be externally and easily controlled, for example, by an external variable resistor or a microprocessor.

The LED drive circuit shown in Figure 3.49 (b) can also be used in modulating the output of the LED as well since the current is proportional to the input voltage as in Figure 3.50 (a). One can, of course, design faster drive circuits, but the bottleneck will eventually be in the speed of response of the LED. An AC signal of frequency f on top of a DC voltage is used to drive the LED in Figure 3.50 (a). The frequency dependence of the output light power $P_o(f)$ with respect to its DC value $P_o(0)$ is shown in Figure 3.50 (b). There is a clear roll-off at high frequencies. The **cutoff frequency** f_c is defined as the frequency at which the normalized power drops by a factor of $2^{1/2}$, that is, at $f = f_c$, $P_o(f_c) = 0.707P_o(0)$. Any modulation of the LED current would result in modulating the minority carrier injection profile as shown in Figure 3.24 (a). The small signal equivalent circuit shown in Figure 3.24 (c) is usually sufficient to understand the frequency response of the LED but we have to be careful inasmuch as r_d and C_{diff} also depend on the frequency at high frequencies. First, consider very low frequencies, that is, almost DC. The capacitances are negligible. The diode action embedded in $I = I_o \exp(eV/k_B T)$ and hence the optical power output P_o , depend on the current in r_d . At high frequencies, the capacitances shunt away this current from r_d and decrease the optical output power. The characteristic time constant τ should be $r_d C_{\text{diff}}$ (C_{diff} is greater than the depletion layer capacitance C_{dep}), that is, the net minority carrier recombination time. A more rigorous treatise arrives at the same conclusion and gives the output optical power $P_o(\omega)$ at frequency ω as

LED optical power output

$$P_o(\omega)/P_o(0) = 1/[1 + (\omega\tau)^2]^{1/2} \quad (3.18.1)$$

At the critical frequency $f_c = 1/(2\pi\tau)$ in Eq. (3.18.1), the relative output power is 0.707 as shown in Figure 3.50 (b). The **optical bandwidth** f_{op} is defined as the frequency at which $P_o(f_{\text{op}})/P_o(0) = 1/2$, and occurs when $f_{\text{op}} = \sqrt{3}f_c$, that is,

Optical bandwidth of an LED

$$f_{\text{op}} = \sqrt{3}/(2\pi\tau) \quad (3.18.2)$$

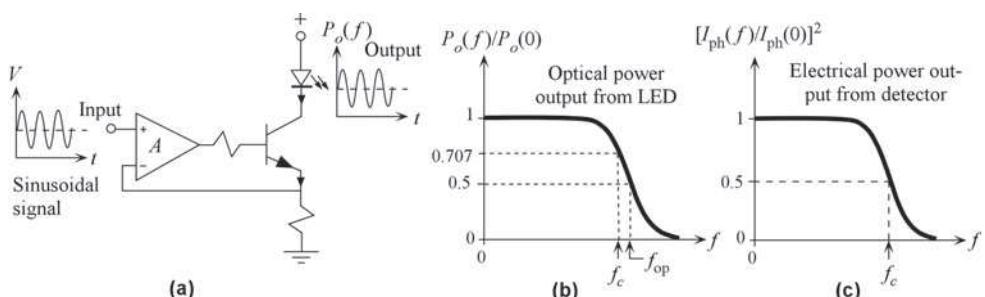


FIGURE 3.50 (a) Sinusoidal modulation of an LED. (b) The frequency response where f_c is the cutoff frequency at which $P_o(f)/P_o(0)$ is 0.707. (c) The electrical power output from the detector as a function of frequency. At f_c , $[I_{\text{ph}}(f)/I_{\text{ph}}(0)]^2$ is 0.5. However, it is 0.707 at a lower frequency than f_c .

where τ is the effective or net recombination time that is, $1/\tau = 1/\tau_r + \tau_{nr}$, where τ_r and τ_{nr} are the radiative and nonradiative recombination lifetimes. Recombination centers in the semiconductor reduce τ_{nr} and hence τ , which would widen the bandwidth. However, the IQE is also reduced, so that the optical output power is diminished as well, which is normally undesirable. If we feed the output from the LED to a photodetector, then the detector photocurrent I_{ph} will be proportional to P_o and the electrical power delivered by the detector will be proportional to I_{ph}^2 or P_o^2 . At the frequency f_c , the detector's electrical output power will be proportional to $[0.707P_o(0)]^2$ or $0.5P_o(0)^2$. Thus, f_c represents a cutoff frequency (a bandwidth) at which the electrical power available from the detector has fallen to half its unmodulated value ($f = 0$ value), as illustrated in Figure 3.50 (c); f_c is called the **electrical bandwidth**.

The turning on and off of an LED in digital applications involves pumping the minority carriers into the active region and then removing them when the LED is turned off. In principle, the minority carrier lifetime τ should also play a key role in this pulsed operation. Figure 3.51 shows a simplified diagram for driving the LED digitally, for example, from a logic gate. The light output pulse should have an exponential rise and fall portions that delay the full turn-on and turn-off of the LED. The characteristic time constant for the rise and fall portions of the pulse should be the net recombination time τ . In practice, exponential rise and fall only serve as approximations to the switching behavior of the output light pulse. The net recombination time τ itself can depend on the current, and hence we should not expect a perfect exponential rise and fall.

Most LED data sheets quote the rise and fall times τ_R and τ_F for LEDs that have been pulsed. τ_R is the rise time from 10% to 90% of the final pulse height as shown in Figure 3.51. Similarly the fall time τ_F is the time it takes for the pulse to drop from its 90% to 10% value before the turn-off was triggered. If we approximate the rise and fall portions as exponentials as mentioned above, then, $\tau_R = \tau_F = 2.2\tau$. It is possible to shape the output pulse by introducing a pulse shaping capacitor C across the resistor R as shown in Figure 3.51. Upon triggering, the capacitor can provide the initial large current to pump (inject) the minority carriers in and hence turn on the device more quickly. Similarly, the capacitor would allow a large turn-off current to take out the minority carriers. Thus τ_R and τ_F can be shortened. The RC time constant must match the net recombination time in the LED for best reduction in the rise and fall times.

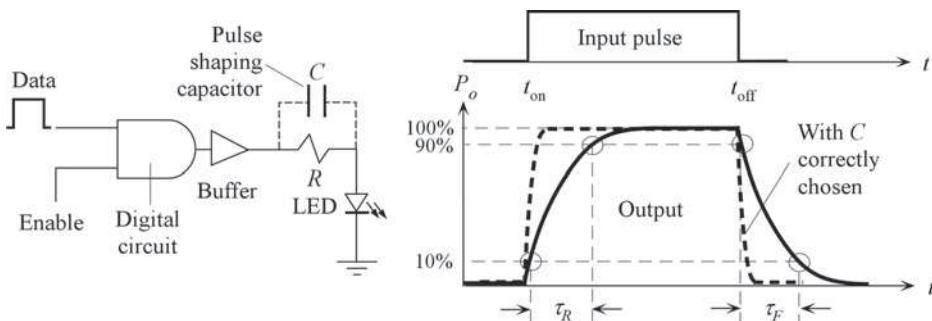


FIGURE 3.51 An LED in a digital circuit is turned on and off by a logic gate, assumed to have a buffered output as shown, to avoid being loaded. A BJT can be used after the logic gate to drive the LED as well (not shown). Definitions of rise and fall times are shown in the light output pulse.

Questions and Problems

3.1 Metals and work function The metal sodium (Na) has an atomic concentration of $2.54 \times 10^{22} \text{ cm}^{-3}$. Each Na atom in isolation has one outer valence electron in an unfilled 3s-subshell. Once the solid is formed, the outer valence electrons from all Na atoms are shared in the crystal by all Na^+ ions, that is, these valence electrons find themselves in an unfilled energy band as in Figure 3.1 (c). Since each Na atom donates one electron to the energy band, the electron concentration is the same as the atomic concentration. Calculate the Fermi energy at 0 K. What is the speed of the electrons at E_F ? What should be the speed at 300 K if the electrons could be treated classically in terms of the kinetic molecular theory as if they were free, similar to the atoms in a gas?

3.2 Photocathode and work function The photocathode of a photomultiplier tube has a multi-alkaline (Sb-Na-K-Cs) metal with a work function (Φ) of 1.55 eV. What is the longest wavelength that will cause photoemission? What is the kinetic energy of a photoemitted electron if the incident light wavelength is 450 nm (blue)? The quantum efficiency (QE) of a photocathode is defined by

$$\text{Quantum efficiency} = \frac{\text{Number of photoemitted electrons}}{\text{Number of incident photons}}$$

The QE is 100% if each incident photon ejects one electron. Suppose that blue light of wavelength 450 nm with an intensity of $1 \mu\text{W cm}^{-2}$ is incident on this photocathode with an area of 50 mm^2 . If the emitted electrons are collected by applying a positive bias voltage to an anode, and the photocathode has a QE of 25%, what will be the photocurrent? (Normally the photoemitted electron is accelerated by a suitable applied field and impacts another electrode, a dynode, where it causes secondary electron emission, and so on, until the current is multiplied by orders of magnitude.)

3.3 Refractive index and bandgap Diamond, silicon, and germanium all have the same diamond unit cell. All three are covalently bonded solids. Their refractive indices (n) and energy bandgaps (E_g) are shown in Table 3.2. (a) Plot n vs. E_g and (b) Plot also n^4 vs. $1/E_g$. What is your conclusion? According to **Moss's rule**, roughly, $n^4 E_g \approx K$, a constant. What is the value of K ?

TABLE 3.2 The refractive index n and the bandgap E_g of diamond, Ge, and Si, all of which have the same crystal structure

Material →	Diamond	Silicon	Germanium
Bandgap, E_g (eV)	5	1.1	0.66
n	2.4	3.46	4.0

3.4 Electrons in the CB of a nondegenerate semiconductor

- (a) Consider the energy distribution of electrons $n_E(E)$ in the conduction band. Assuming that the density of state $g_{\text{CB}}(E) \propto (E - E_c)^{1/2}$ and using Boltzmann statistics $f(E) \approx \exp[-(E - E_F)/k_B T]$, show that the energy distribution of the electrons in the CB can be written as

$$n_x(x) = Cx^{1/2} \exp(-x)$$

where $x = (E - E_c)/k_B T$ is the electron energy in terms of $k_B T$ measured from E_c and C is a constant at a given temperature (independent of E).

- (b) Setting arbitrarily $C = 1$, plot $n_x(x)$ vs. x . Where is the maximum and what is the FWHM (full width at half maximum, that is, between half maximum points)? Is the use of $1.8k_B T$ for the half-maximum width correct?
- (c) Show that the average electron energy in the CB is $(3/2)k_B T$, by using the definition of average,

$$x_{\text{average}} = \int_0^{\infty} xn_x dx / \int_0^{\infty} n_x dx$$

where the integration is from $x = 0$ (E_c) to say $x = 10$ (far away from E_c where $n_x \rightarrow 0$). You need to use a numerical integration.

- (d) Show that the maximum in the energy distribution is at $x = (1/2)$ or at $E_{\text{max}} = (1/2)k_B T$.

- 3.5 Intrinsic and doped GaAs** The properties of GaAs are shown in Table 3.1. Calculate the intrinsic concentration and the intrinsic resistivity at room temperature (take as 300 K). Where is the Fermi level? Assuming the N_c and N_v scale as $T^{3/2}$, what would be the intrinsic concentration at 100°C? If this GaAs crystal is doped with 10^{17} donors cm^{-3} (such as Te), where is the new Fermi level and what is the resistivity of the sample? The drift mobilities in GaAs are shown in Table 3.3.

**TABLE 3.3 Ionized dopant impurities scatter carriers and reduce the drift mobility.
The dependence of μ_e for electrons and μ_h for holes on the total ionized dopant concentration**

Dopant concentration (cm^{-3})	0	10^{14}	10^{15}	10^{16}	10^{17}	10^{18}
GaAs, $\mu_e (\text{cm}^2 \text{V}^{-1} \text{s}^{-1})$	8500	—	8000	7000	5000	2400
GaAs, $\mu_h (\text{cm}^2 \text{V}^{-1} \text{s}^{-1})$	400	—	380	310	250	160
Si, $\mu_e (\text{cm}^2 \text{V}^{-1} \text{s}^{-1})$	1450	1420	1370	1200	730	280
Si, $\mu_h (\text{cm}^2 \text{V}^{-1} \text{s}^{-1})$	490	485	478	444	328	157

- 3.6 Electrons in GaAs** Given that the electron effective mass m_e^* for the GaAs is $0.067m_e$, calculate the thermal velocity of the electrons in the CB of a nondegenerately doped GaAs at room temperature (300 K). If μ_e is the drift mobility of the electrons and τ_e the mean free time between electron scattering events (between electrons and lattice vibrations) and if $\mu_e = e\tau_e/m_e^*$, calculate τ_e , given $\mu_e = 8500 \text{ cm}^2 \text{V}^{-1} \text{s}^{-1}$. Calculate the **drift velocity** $v_d = \mu_e E$ of the CB electrons in an applied field E of 10^5 V m^{-1} . What is your conclusion?
- 3.7 Extrinsic *n*-GaAs** An *n*-type GaAs crystal is doped with 10^{16} donors cm^{-3} (such as Te), what are the electron and hole concentrations, and the conductivity? (See Table 3.3.)
- 3.8 Extrinsic *n*-Si** A Si crystal has been doped *n*-type with $1 \times 10^{17} \text{ cm}^{-3}$ phosphorus (P) donors. The electron drift mobility μ_e depends on the total concentration of ionized dopants N_{dopant} , as in Table 3.3, inasmuch as these ionized dopants scatter the electrons and thereby decrease their drift mobility. What is the conductivity of the sample? Where is the Fermi level with respect to the intrinsic crystal?
- 3.9 Compensation doping in *n*-type Si** An *n*-type Si sample has been doped with 10^{16} phosphorus (P) atoms cm^{-3} . (a) What are the electron and hole concentrations? (b) Calculate the room temperature conductivity of the sample. (c) Where is the Fermi level with respect to E_{Fi} ? (d) If we now dope the crystal with 10^{17} boron acceptors, what will be the electron and hole concentrations? (e) Where is the Fermi level with respect to E_{Fi} ?
- 3.10 Free carrier absorption in semiconductors** The attenuation of light due to the optical field drifting the free carriers is called **free carrier absorption**. As the free electrons in a semiconductor crystal are accelerated by the optical field, they eventually become scattered by lattice vibrations or impurities, and pass the energy absorbed from the radiation to lattice vibrations. In such cases, ε''_r and the AC conductivity σ at the same frequency are related by

$$\varepsilon''_r = \sigma / \varepsilon_o \omega \quad (\text{P3.1})$$

We consider a semiconductor in which the free carriers are electrons (an *n*-type semiconductor). The AC conductivity σ in general is given by

$$\sigma = \sigma_o / (1 + j\omega\tau_e) \quad (\text{P3.2})$$

where σ_o is the DC conductivity, ω is the angular frequency of light, and τ_e is the scattering time of the conduction electrons. σ decreases with frequency in Eq. (P3.2). Consider an *n*-type semiconductor. The free carriers are the electrons in the CB. If the drift mobility of the electrons is μ_e , then $\mu_e = e\tau_e/m_e^*$, where m_e^* is the effective mass of the electrons in the CB of the semiconductor.²⁸ The DC conductivity $\sigma_o = en\mu_e$, where n is the concentration of CB electrons. Show that the absorption coefficient due to free carrier absorption (due to the conductivity) when $\omega > 1/\tau$ is given by

$$\alpha = \left(\frac{\sigma_o}{n c \varepsilon_o \tau_e^2} \right) \left(\frac{1}{\omega^2} \right) = \left(\frac{e^3 n}{4\pi^2 n c^3 \varepsilon_o m_e^{*2} \mu_e} \right) \lambda^2 \quad (\text{P3.3})$$

²⁸ As we know, conduction in an *n*-type semiconductor occurs by the drift of free electrons inside the semiconductor crystal. The dopants donate electrons to the crystal, which are free within the crystal.

where n is the refractive index. What would you expect if you plotted α vs. λ^2 ? Consider a Si crystal doped with 10^{15} cm^{-3} donors. Estimate the free carrier absorption (in m^{-1} and dB m^{-1}) at 1.55 and at 5 μm . What is your conclusion? [Although Eq. (P3.3) is a highly simplified approximation to describing free carrier absorption, it nonetheless provides a rough estimate of its magnitude.]

- 3.11 GaAs pn junction** Consider a GaAs pn junction that has the following properties: $N_a = 10^{16} \text{ cm}^{-3}$ (p -side); $N_d = 10^{18} \text{ cm}^{-3}$ (n -side); $B = 2.0 \times 10^{-16} \text{ m}^3 \text{ s}^{-1}$; cross-sectional area $A = 1.5 \text{ mm} \times 1.5 \text{ mm}$. Assume a long diode. What is the diode current due to diffusion in the neutral regions and recombination in the SCL at 300 K when the forward voltage across the diode is 0.8 V and then 1.1 V? (Use the drift mobilities in Table 3.3 for calculating the diffusion coefficients through the Einstein relation.)
- 3.12 InP pn junction** Consider an InP pn junction that has the following properties: $N_a = 10^{15} \text{ cm}^{-3}$ (p -side); $N_d = 10^{17} \text{ cm}^{-3}$ (n -side); using $B \approx 4 \times 10^{-16} \text{ m}^3 \text{ s}^{-1}$; cross-sectional area $A = 1 \text{ mm} \times 1 \text{ mm}$. Assume a long diode. What is the diode current due to diffusion in the neutral regions and recombination in the SCL at 300 K when the forward voltage across the diode is 0.70 and 0.9 V? The electron mobility in the p -side is about $\sim 6000 \text{ cm}^2 \text{ s}^{-1}$ and the hole mobility on the n -side is roughly $\sim 100 \text{ cm}^2 \text{ s}^{-1}$. (See also Table 3.1 for n_i and ε_r) Comment on the ideality factor of this InP pn junction.
- 3.13 Si pn junction** Consider a long pn junction diode with an acceptor doping N_a of 10^{18} cm^{-3} on the p -side and donor concentration of N_d on the n -side. The diode is forward biased and has a voltage of 0.6 V across it. The diode cross-sectional area is 1 mm^2 . The minority carrier recombination time, τ , depends on the total dopant concentration, $N_{\text{dopant}}(\text{cm}^{-3})$, through the following approximate empirical relation

$$\tau \approx (5 \times 10^{-7}) / (1 + 2 \times 10^{-17} N_{\text{dopant}})$$

where τ is in seconds.

- (a) Suppose that $N_d = 10^{15} \text{ cm}^{-3}$. Then the depletion layer extends essentially into the n -side and we have to consider minority carrier recombination time, τ_h , in this region. Calculate the diffusion and recombination contributions to the total diode current given $N_a = 10^{18} \text{ cm}^{-3}$ and $N_d = 10^{15} \text{ cm}^{-3}$. Use Table 3.3 for μ_e and μ_h . What is your conclusion?
- (b) Suppose that $N_d = N_a$. Then W extends equally to both sides and, further, $\tau_e = \tau_h$. Calculate the diffusion and recombination contributions to the diode current given $N_a = 10^{18} \text{ cm}^{-3}$ and $N_d = 10^{18} \text{ cm}^{-3}$. Use Table 3.3 for μ_e and μ_h . What is your conclusion?
- 3.14 Injected minority carrier charge** Consider a pn junction with heavier doping on the p -side. The injected minority carriers (holes) represent an *injected excess minority carrier charge* Q_h in the neutral region as shown in Figure 3.24 (a). (There is also excess majority carrier charge so the region is neutral.) Show that

$$Q = I\tau_h \text{ for a long diode and } Q = I\tau_t \text{ for a short diode}$$

in which τ_h is the hole lifetime and τ_t is the diffusion time, or the transit time of holes across the width of the neutral n -region, that is, $\tau_t = l_n^2/2D_h$. What is your conclusion?

- 3.15 High injection condition** The Shockley equation for a pn junction under forward bias, as shown in Figure 3.16 (a), was derived by assuming low (weak) injection conditions, that is $p_n(0) \approx \Delta p_n(0) \ll n_{no}$ or N_d on the n -side. Show that when the injection is no longer weak, that is when $p_n(0) \approx n_{no} = N_d$, the applied voltage V reaches V_{SI} (strong injection) given by

$$V_{SI} = V_o - V_{th} \ln(N_a/N_d)$$

where V_{th} is the thermal voltage ($k_B T/e$). Calculate V_o and V_{SI} for a Si pn junction that has $N_a = 10^{18} \text{ cm}^{-3}$ and $N_d = 10^{16} \text{ cm}^{-3}$. Can you use the Shockley equation when $V > V_{SI}$? What happens when $N_a = N_d$? What is your conclusion?

- 3.16 Heterostructure** Consider a Type I heterostructure as shown in Figure 3.27.
- (a) If $E_{g1} < E_{g2}$ and if χ_1 and χ_2 are the electron affinities of each semiconductor, show that
- $$\Delta E_c = \chi_1 - \chi_2 \quad \text{and} \quad \Delta E_v = E_{g2} - E_{g1} - \Delta E_c$$
- (b) Using the data in Table 3.1, draw the energy band diagram of an nP junction between an n -type Ge and P -GaAs. Under forward bias, is it easier to inject electrons or holes?
- (c) Draw the energy band diagram for a pN junction between p -type Ge and N -GaAs. Under forward bias, is it easier to inject electrons or holes?
- 3.17 Heterojunction I – V characteristics** We use some of the data reported by Womac and Rediker (*J. Appl. Phys.*, 43, 4130, 1972) for AlGaAs/GaAs pn heterojunction at 298 K (25°C). Sample A is an N^+p and sample B is a P^+n junction. The I – V data of interest are listed in Table 3.4. By a suitable plot find the ideality factor for each. What is your conclusion?

TABLE 3.4 I - V data on two heterojunctions. First set is N^+p and the second set is a P^+n junction

A: Heterojunction N^+p														
V	0.206 V	0.244	0.290	0.322	0.362	0.412	0.453	0.485	0.537	0.576	0.612	0.662	0.708	
I	1.03 nA	2.07	5.20	10.3	20.7	52.8	105	192	515	1.02 μ A	2.03	4.89	10.1	
B: Heterojunction P^+n														
V	0.310 V	0.364	0.402	0.433	0.485	0.521	0.561	0.608	0.682	0.726	0.764	0.807	0.859	0.885
I	2.01 nA	4.91	9.79	19.0	49.5	96.1	194	466	1.96 μ A	5.02	9.75	19.5	50.6	99.5

- 3.18 AlGaAs LED emitter** An AlGaAs LED emitter for in a local optical fiber network has the output spectrum shown in Figure 3.32 (b). It is designed for peak emission at about 822 nm at 25°C. (a) Why does the peak emission wavelength increase with temperature? (b) What is the bandgap of AlGaAs in this LED? (c) The bandgap, E_g , of the ternary $Al_xGa_{1-x}As$ alloys follows the empirical expression, $E_g(eV) = 1.424 + 1.266x + 0.266x^2$. What is the composition of the $Al_xGa_{1-x}As$ in this LED?

- 3.19 III-V compound semiconductors in optoelectronics** Figure 3.52 represents the bandgap E_g and the lattice parameter a in a quaternary III-V alloy system. A line joining two points represents the changes in E_g and a with composition in a ternary alloy composed of the compounds at the ends of that line. For example, starting at GaAs point, $E_g = 1.42\text{ eV}$ and $a = 0.565\text{ nm}$, E_g decreases and a increases as GaAs is alloyed with InAs, as we move

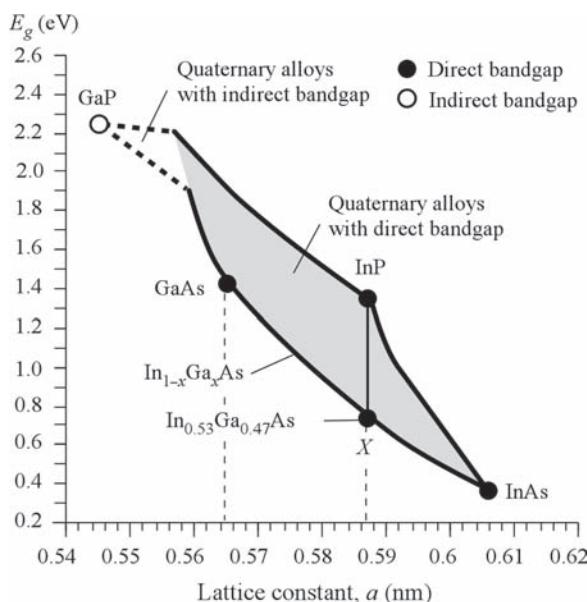
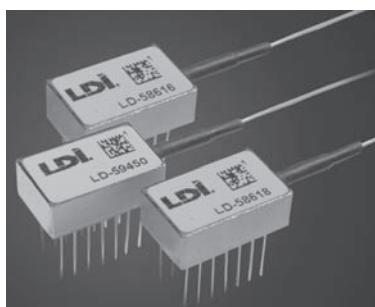


FIGURE 3.52 Bandgap energy E_g and lattice constant a for various III-V alloys of GaP, GaAs, InP, and InAs. A line represents a ternary alloy formed with compounds from the end points of the line. Solid lines are for direct bandgap alloys whereas dashed lines for indirect bandgap alloys. Regions between lines represent quaternary alloys. The line from X to InP represents quaternary alloys $In_{1-x}Ga_xAs_{1-y}P_y$ made from $In_{0.53}Ga_{0.47}As$ and InP, which are lattice-matched to InP.



InGaAsP 1300 nm LED emitters, each pigtailed to an optical fiber for use in optical communication modems and lower speed data/analog transmission systems. (Courtesy of OSI Laser Diode, Inc.)

along the line joining GaAs to InAs. Eventually at InAs, $E_g = 0.35$ eV and $a = 0.606$ nm. Point X in Figure 3.52 is composed of InAs and GaAs and it is the ternary alloy $\text{In}_{1-x}\text{Ga}_x\text{As}$. At X, $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ (often called “in-gas” in telecom) has $E_g = 0.73$ eV and $a = 0.587$ nm, which is the same a as that for InP. $\text{In}_{1-x}\text{Ga}_x\text{As}$ at X is therefore lattice-matched to InP and can hence be grown on an InP substrate without creating defects at the interface.

Further, $\text{In}_{1-x}\text{Ga}_x\text{As}$ at X can be alloyed with InP to obtain a quaternary alloy²⁹ $\text{In}_{1-x}\text{Ga}_x\text{As}_{1-y}\text{P}_y$ whose properties lie on the line joining X and InP and therefore all have the same lattice parameter as InP but different bandgap. Layers of $\text{In}_{1-x}\text{Ga}_x\text{As}_{1-y}\text{P}_y$ with composition between X and InP can be grown epitaxially on an InP substrate by various techniques such as liquid phase epitaxy (LPE) or molecular beam epitaxy (MBE). The grey shaded area between the solid lines represents the possible values of E_g and a for the quaternary III–V alloy system in which the bandgap is direct and hence suitable for direct recombination. The compositions of the quaternary alloy lattice matched to InP follow the line from X to InP.

- (a) Given that the $\text{In}_{1-x}\text{Ga}_x\text{As}$ at X is $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ show that quaternary alloys $\text{In}_{1-x}\text{Ga}_x\text{As}_{1-y}\text{P}_y$ are lattice matched to InP when $y = 1 - 2.13x$.
- (b) The bandgap energy E_g , in eV for $\text{In}_{1-x}\text{Ga}_x\text{As}_{1-y}\text{P}_y$ lattice-matched to InP is given by the empirical relation, $E_g(\text{eV}) = 0.75 + 0.46y + 0.14y^2$. Find the composition of the quaternary alloy suitable for an LED emitter operating at $1.30 \mu\text{m}$.

- 3.20 Varshni equation and the change in the bandgap with temperature** The Varshni equation describes the change in the energy bandgap E_g of a semiconductor with temperature T as given by Eq. (3.11.2) that is

$$E_g = E_{go} - AT^2/(B + T)$$

where E_{go} is E_g at 0 K, and A and B are constants. Show that

$$\frac{dE_g}{dT} = -\frac{AT(T + 2B)}{(B + T)^2} = -\frac{(E_{go} - E_g)}{T} \left(\frac{T + 2B}{T + B} \right)$$

For GaAs, $E_{go} = 1.519$ eV, $A = 5.41 \times 10^{-4}$ eV K $^{-1}$, $B = 204$ K. What is dE_g/dT for GaAs? Find the shift in the emitted wavelength from a GaAs LED per 1°C change at room temperature (300 K). Find the emission wavelength at 27°C and –30°C.

- 3.21 Dependence on the emission peak and linewidth on temperature** Using the Varshni equation find the peak emission wavelength and the linewidth of the emission spectrum from an $\text{In}_{0.47}\text{Ga}_{0.53}\text{As}$ LED when it is cooled from 25°C to –25°C. You can use Eqs. (3.3.1) and (3.3.2). The Varshni constants for $\text{In}_{0.47}\text{Ga}_{0.53}\text{As}$ are $E_{go} = 0.814$ eV, $A = 4.906 \times 10^{-4}$ eV K $^{-1}$, $B = 301$ K.

- 3.22 LED Output Spectrum** Given that the width of the relative light intensity vs. photon energy spectrum of an LED is typically around $\sim 3 k_B T$, calculate the spectral width in wavelength (nm) of LED emitters operating at 850 nm, 1310 nm, and 1550 nm?

- 3.23 Linewidth of LEDs** Experiments carried out on various direct bandgap semiconductor LEDs give the output spectral linewidth (between half intensity points as listed in Table 3.5). From Figure 3.31 we know that a spread in the wavelength is related to a spread in the photon energy, $\Delta\lambda \approx (hc/E_{ph}^2)\Delta E_{ph}$ where $E_{ph} = h\nu$ is the photon energy. Suppose that we write $E_{ph} = hc/\lambda$ and $\Delta E_{ph} = \Delta(h\nu) \approx mk_B T$ where m is a numerical constant. Therefore,

$$\Delta\lambda \approx (mk_B T/hc)\lambda^2 \quad (\text{P3.4})$$

By appropriately plotting the data in Table 3.5, and assuming $T = 300$ K, find m .

TABLE 3.5 Linewidth $\Delta\lambda_{1/2}$ between half points in the output spectrum (spectral intensity vs. wavelength) of eight LEDs using direct bandgap semiconductors

Material (Direct E_g)	AlGaAs	AlGaAs	AlGaAs	GaAs	GaAs	InGaAsP	InGaAsP	InGaAsP
Peak wavelength of emission (λ) nm	650	810	820	890	950	1150	1270	1500
$\Delta\lambda_{1/2}$ nm	22	36	40	50	55	90	110	150

²⁹Some books have other formats for the chemical composition for example, $\text{Ga}_x\text{In}_{1-x}\text{As}_y\text{P}_{1-y}$. The present notation $\text{In}_{1-x}\text{Ga}_x\text{As}_{1-y}\text{P}_y$ was chosen to reflect the common vernacular for InGaAs (pronounced “in-gas”).

Table 3.6 gives the linewidth $\Delta\lambda_{1/2}$ for various visible LEDs based on GaAsP. Radiative recombination is obtained by appropriately doping the material. Using $m = 3.0$ in Eq. (P3.4), $T = 300\text{K}$, calculate the expected spectral width for each and compare with the experimental value. What is your conclusion?

TABLE 3.6 Linewidth $\Delta\lambda_{1/2}$ between half points in the output spectrum (intensity vs. wavelength) of various visible LEDs using GaAsP

Peak wavelength of emission (λ) nm	565	583	600	635
$\Delta\lambda_{1/2}$ nm	28	36	40	40
Color	Green	Yellow	Orange	Red
Material	GaP(N)	GaAsP(N)	GaAs (N)	GaAsP

- 3.24 Quantum wells** We will consider a quantum well in a semiconductor heterostructure, and assume, for simplicity, an infinite potential energy quantum well with dimensions d along x , and D_y and D_z along y and z directions. The energy of an electron with respect to the bottom of the well is then given by

$$\Delta E = \frac{\hbar^2 n^2}{8m_e^* d^2} + \frac{\hbar^2 n_y^2}{8m_e^* D_y^2} + \frac{\hbar^2 n_z^2}{8m_e^* D_z^2} \quad (\text{P3.5})$$

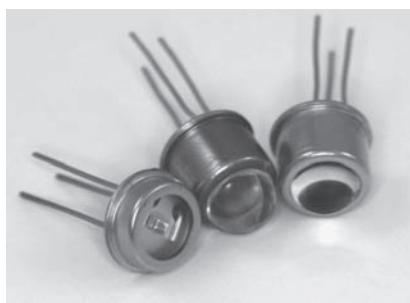
Energy
in a
quantum
well

where n , n_y , and n_z are quantum numbers having the values 1, 2, 3, ... Consider a well that has $d = 10\text{nm}$, $D_x = D_y = 2\text{\mu m}$, and $m_e^* = 0.067m_e$. (a) Calculate the minimum energy. Which is the most significant contributing term in Eq. (P3.5) to the lowest energy? (b) What would n_y need to be to get the same energy as the first term? (c) What is the separation δE between the energy levels for motion in the y and z plane? What is your conclusion?

- 3.25 Energy levels in a quantum well** Consider a GaAs QW sandwiched between two $\text{Al}_{0.40}\text{Ga}_{0.60}\text{As}$ layers. The barrier height ΔE_c is approximately 0.30 eV, the electron effective mass in the well is $0.067m_e$, and the width of the QW (d) is 8 nm. Calculate the energy levels ΔE_1 and ΔE_2 from the bottom of E_c , assuming an infinite PE well. Compare these with the calculations for a finite PE well that give $\Delta E_1 = 0.050\text{ eV}$ and $\Delta E_2 = 0.197\text{ eV}$. What is your conclusion?
- 3.26 QW LED Linewidth** Consider the QW-based InGaN (green to blue) and AlGaN (UV) LEDs whose peak emission wavelengths (λ_o) and linewidths (between half intensity points, $\Delta\lambda$) are listed in Table 3.7. For each LED, apply the linewidth equation $\Delta\lambda \approx (mk_B T/hc)\lambda_o^2$ at 300 K, and find m . Do you think this formula can be used for QW LEDs?

TABLE 3.7 Avago ASMT 1W series of InGaN green, cyan, blue, and royal blue LEDs and SETi AlGaN UV LED

LED	Green	Cyan	Blue	Royal blue	UV
λ_o	519	497	454	450	270
$\Delta\lambda$	31.25	29	22.5	20	12



UV emitting LEDs (Courtesy of Sensor Electronic)

3.27 External conversion efficiency The total light output power, called the radiant flux, from a number of commercial LEDs is listed in Table 3.8. The table also lists the emitted peak wavelength and the measured linewidth for each LED. For each device calculate the external quantum efficiency (%), EQE, and m in $\Delta\lambda \approx (mk_B T/hc)\lambda^2$. What is your conclusion?

TABLE 3.8 Characteristics of selected commercial LEDs covering the IR, blue, and UV

IR, Hamamatsu L10823 (InGaAs)	IR, Hamamatsu L10822 (InGaAs)	IR, Osram SFH406 (GaAs)	IR, Vishay TSAL6100 (GaAlAs)	IR, Hamamatsu L1909 (AlGaAs)	Super Blue, Cree (InGaN)	UV, UVTOP 315, SETi (InGaN)
V_F (V)	0.9	1.1	1.25	1.35	1.4	3.8
I_F (mA)	50	50	40	100	50	30
P_o (mW)	2	3.1	7	35	10	1.15
λ_o (nm)	1650	1300	950	940	870	466
$\Delta\lambda$ (nm)	120	90	55	50	80	11

3.28 Light extraction efficiency of LEDs Extraction efficiency or ratio depends on the LED structure and the means that have been used to manipulate light to escape from the chip.

- (a) Figure 3.53 (a) shows a bare LED chip in which the light from the active region is incident on the semiconductor-air interface. The refractive indices of the semiconductor and ambient are n_s and n_a respectively. Assume a GaN crystal for the semiconductor with a refractive index of 2.5. For normal incidence, what is the transmittance T from the semiconductor to the air? What is the critical angle θ_c at this interface? The fraction of output light from the active region within the emission cone defined by the critical angle θ_c has been evaluated to be $(1/2)[1 - \cos \theta_c]$. Taking into account T as well, estimate the extraction efficiency (EE) if the ambient is air, $n_a = 1$.

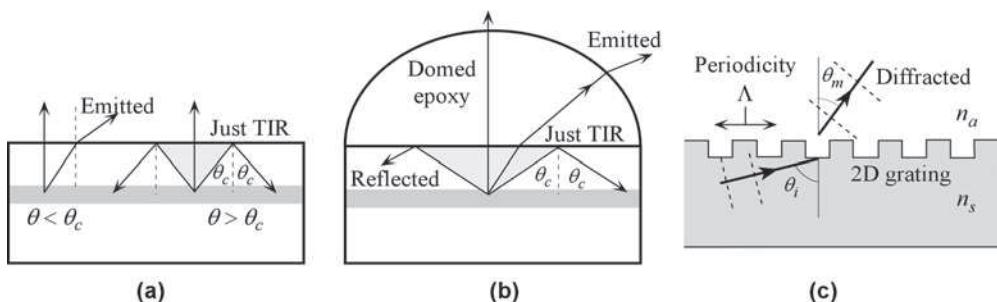


FIGURE 3.53 (a) Total internal reflection results in poor light extraction (b) Improvement in the light extraction by the use of a domed epoxy. (c) A two-dimensional diffraction grating (a photonic crystal) based surface enhances the extraction of light through Bragg diffraction.

- (b) Figure 3.53 (b) shows an epoxy dome placed on top of the semiconductor. Suppose that the refractive index of the epoxy is 1.8. Calculate the new critical angle at the semiconductor-epoxy interface and estimate the new EE. What is the critical angle for light rays at the epoxy-air interface? How critical is it to have an exactly hemispherical surface?
- (c) Figure 3.53 (c) shows an LED chip whose surface has been patterned to form a two-dimensional (2D) *photonic crystal*, which is a periodic modulation of the refractive index. With the right periodicity Λ , the photonic crystal will leak out the light trapped in the crystal. We will take a highly simplified view and treat the periodic variation as a grating. If θ_i is the angle of incidence, there is a diffracted beam at an angle θ_m , which satisfies the Bragg diffraction condition that involves two different media that is,

$$\Lambda(n_s \sin \theta_i - n_a \sin \theta_m) = m\lambda$$

where λ is the free-space wavelength, m is the diffraction order. Suppose that $n_s \approx 2.5$ (GaN), $n_a = 1$ (air), $\lambda = 500$ nm, $\Lambda = 300$ nm, find the critical angle and the first-order diffracted beams when $\theta_i = 60^\circ$. What is your conclusion?

- 3.29 LED efficiencies** A particular 890 nm IR LED for use in instrumentation has a AlGaAs chip. The active region has been doped p -type with $4 \times 10^{17} \text{ cm}^{-3}$ of acceptors and the nonradiative lifetime is about 60 ns. At a forward current of 50 mA, the voltage across it is 1.4 V, and the emitted optical power is 10 mW. Calculate the PCE, IQE, and EQE, and estimate the light extraction ratio. For AlGaAs, $B \approx 1 \times 10^{-16} \text{ m}^3 \text{ s}^{-1}$.

3.30 LED luminous flux

- (a) Consider a particular green LED based on InGaN MQW active region. The emission wavelength is 528 nm. At an LED current of 350 mA, the forward voltage is 3.4 V. The emitted luminous flux is 92 lm. Find the power conversion efficiency, external quantum efficiency, luminous efficacy, and the emitted optical power (radian flux)? (Data for Osram LT CPDP.)
- (b) A red LED emits 320 mW of optical power at 656 nm when the current is 400 mA and the forward voltage is 2.15 V. Calculate the power conversion efficiency, external efficiency and the luminous efficacy. (Data for thin film InGaAlP Osram LH W5AM LED.)
- (c) A deep blue LED emits at an optical power of 710 mW at 455 nm when the current is 350 mA and the forward voltage is 3.2 V. Calculate the power conversion efficiency, external efficiency, and luminous efficacy. (Data for a GaN Osram LD W5AM LED.)

- 3.31 SLEDs and ELEDs** Experiments carried out on an AlGaAs SLED (surface-emitting LED) and an ELED (edge-emitting LED) give the light output power vs. current data in Table 3.9. (a) Show that the output light power vs. current characteristics are not linear. (b) By plotting the optical power output (P_o) vs. current (I) data on a log-log plot show that $P_o \propto I^n$. Find n for each LED.

TABLE 3.9 Light output power vs. DC current for surface- and edge-emitting LEDs

SLED I (mA)	25	50	75	100	150	200	250	300
SLED Light output power, P_o (mW)	1.04	2.07	3.1	4.06	5.8	7.6	9.0	10.2
ELED I (mA)	25	50	75	100	150	200	250	300
ELED Light output power (mW) P_o	0.46	0.88	1.28	1.66	2.32	2.87	3.39	3.84

3.32 LED-Fiber coupling Efficiency

- (a) It is found that approximately 200 μW is coupled into a multimode step index fiber from a surface-emitting LED when the current is 75 mA and the voltage across the LED is about 1.5 V. What is the overall efficiency of operation?
- (b) Experiments are carried out on coupling light from a 1310 nm ELED (edge-emitting LED) into multimode and single-mode fibers. (i) At room temperature, when the ELED current is 120 mA, the voltage is 1.3 V and light power coupled into a 50 μm multimode fiber with NA (numerical aperture) of 0.2 is 48 μW . What is the overall efficiency? (ii) At room temperature, when the ELED current is 120 mA, the voltage is 1.3 V and light power coupled into a 9 μm single mode fiber is 7 μW . What is the overall efficiency?

- 3.33 Internal optical power** For a particular AlGaAs LED emitting at 850 nm, the nonradiative recombination lifetime is $\tau_{nr} = 50$ ns. The recombination occurs in the active region, which has been doped with acceptors of concentration 10^{17} cm^{-3} , and the direct recombination coefficient B is $2 \times 10^{-16} \text{ cm}^3 \text{ s}^{-1}$. What is the internal optical power generated at a current of 100 mA?

- 3.34 Internal quantum efficiency (IQE) and bandwidth product** Show that IQE and the bandwidth f_c product is

$$\eta_{\text{IQE}} f_c = \frac{1}{2\pi\tau_r}$$

What is your conclusion? Consider an LED in which the radiative and nonradiative lifetimes are approximately 10 ns and 50 ns, respectively. What are the IQE and the possible bandwidth f_c ? What happens if defects cause τ_{nr} to be 25 ns? What other factors do you think would reduce the bandwidth?

3.35 Bandwidth and rise and fall time

- (a) Suppose the rise τ_R for an LED is given, and the rise and fall times are the same. Show that its AC modulation bandwidth f_c (the cutoff frequency) is given by

$$f_c = 0.35/\tau_R$$

- (b) The specifications for a particular 1300 nm InGaAs ELED (PE13W series) pigtailed to a fiber has rise and fall times given by 2.5 ns. Calculate the bandwidth f_c . How does that compare with the quoted value 150 MHz?
- (c) At large currents, that is, when the minority injection is strong, the radiative recombination time τ_r is no longer constant, but given by Eq. (3.8.7), $\tau_r = 1/B\Delta n$, where Δn is the injected concentration of carriers. The injected carriers (Δn) into the active region are brought in by the current (J) and then they recombine and emit photons, that is, Δn is proportional to the current density J . How would this affect the bandwidth?

3.36 LED electronic drive circuits Figures 3.49 (b) shows one possible way to drive an LED. Assume that V_{CC} is large and that the base-emitter voltage of the BJT is 0.7 V, show that $I_F \approx V / R_F$. How large should V_{CC} be? Consider a red LED that has $V_F = 2$ V at $I_F = 20$ mA and $I_{max} = 50$ mA. If the input control voltage range is from 0 to 5V, what would you recommend for R_F , R_B , and V_{CC} ?

3.37 LED electronic drive circuit design Figures 3.54 (a)–(c) shows three possible circuits that can be used to drive a high-current LED from CMOS or similar low-current logic circuits. A bipolar junction transistor (Q_1) is commonly used for this task.

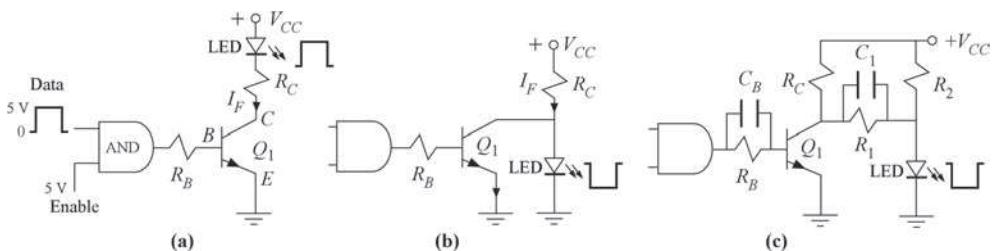
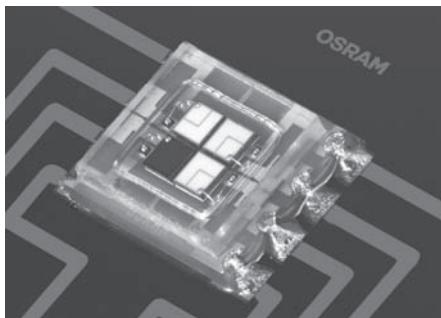


FIGURE 3.54 (a) A CMOS logic circuit (or a gate) drives an LED through a BJT Q_1 . (b) The logic gate drives Q_1 , which shunts the current from the LED and switches it off. (c) A combination of (a) and (b) with speeding capacitors C_B and C_1 to turn the LED on and off more quickly. In all three cases, the inputs to the logic circuit are the same.

- (a) Consider Figure 3.54 (a), if the input is high (5 V), Q_1 is fully turned on and $V_{CE} = 0.25$ V (saturated CE voltage), and $V_{BE} = 0.7$ V. What is the current I_F through the LED? If the LED needs $I_F = 50$ mA and $V_F = 2$ V, and $V_{CC} = 5$ V, what is R_C ? Suppose that Q_1 has a minimum β (or h_{FE}) of 70, what is R_B ? Is there an advantage in using higher V_{CC} voltages?
- (b) Consider Figure 3.54 (b). Show that the LED is turned off when the input is high (5 V) [as opposed to (a) above]. For the same LED as in (a) with $I_F = 50$ mA, $V_F = 2$ V, and $V_{CC} = 5$ V, find R_C and R_B .
- (c) The circuit Figure 3.54 (c) has been *designed* to drive an LED that has $I_F = 25$ mA and $V_F = 2$ V. The resistors are $R_B = 5.6$ k Ω , $R_C = 100$ Ω , $R_1 = 22$ Ω , and $R_2 = 820$ Ω . Explain how the circuit works, assuming first Q_1 is off and then it is on ($V_{CE} = 0.25$ V). Is I_F close to 25 mA when Q_1 is off? The capacitors C_B and C_1 across R_B and R_1 speed up the circuit. What do you think $C_B R_B$ and $C_1 R_1$ should be? What are the requirement on Q_1 ?



This multichip LED from Osram is used in microprojectors. It is based on thin-film GaN technology. (Courtesy of OSRAM Opto Semiconductors, Germany.)



A handheld microprojector MPro120 from 3M for projecting photos and videos on a wall uses LEDs for its projection light. (Courtesy of 3M.)



LEDs have now widely replaced the incandescent light bulbs in traffic lights.



This LED-based light bulb for use in chandeliers uses a few watts of electric power (2.5W) to generate 135 lumens of light. There are 18 LED chips each with a yellow phosphor to generate the white light.

*We thought it [the laser] might have some communications and scientific uses,
but we had no application in mind. If we had, it might have hampered us and
not worked out as well.*
—Arthur Schawlow¹



Zhores Alferov (on the right) with Valery Kuzmin (technician) in 1971 at the Ioffe Physical Technical Institute, discussing their experiments on heterostructures. Zhores Alferov carried out some of the early pioneering work on heterostructure semiconductor devices that led to the development of a number of important optoelectronic devices, including the heterostructure laser. Zhores Alferov and Herbert Kroemer shared the Nobel Prize in Physics (2000) with Jack Kilby. Their Nobel citation is "for developing semiconductor heterostructures used in high-speed- and opto-electronics." (*Courtesy of Professor Zhores Alferov, Ioffe Physical Technical Institute.*)

¹ Arthur Schawlow (1921–1999; Nobel Laureate, 1981) talking about the invention of the laser.

4

STIMULATED EMISSION DEVICES: OPTICAL AMPLIFIERS AND LASERS

4.1 STIMULATED EMISSION, PHOTON AMPLIFICATION, AND LASERS

A. Stimulated Emission and Population Inversion

An electron in an atom can be excited from an energy level E_1 to a higher energy level E_2 by the absorption of a photon of energy $h\nu = E_2 - E_1$ as shown in Figure 4.1 (a). When an electron at a higher energy level transits down in energy to an unoccupied energy level, it emits a photon. There are essentially two possibilities for the emission process. The electron can undergo the downward transition by itself quite *spontaneously*, or it can be *induced* to do so by another photon.

In **spontaneous emission**, the electron falls down in energy from level E_2 to E_1 and emits a photon of energy $h\nu = E_2 - E_1$ in a random direction as indicated in Figure 4.1 (b). The transition is spontaneous provided that the state with energy E_1 is empty, *i.e.*, not already occupied by another electron. In classical physics, when a charge accelerates and decelerates as in an oscillatory motion with a frequency ν it emits an electromagnetic radiation also of frequency ν . The emission process during the transition of the electron from E_2 to E_1 can be classically thought of as if the electron is oscillating with a frequency ν .

In **stimulated emission**², an incoming photon of energy $h\nu = E_2 - E_1$ stimulates the whole emission process by inducing the electron at E_2 to transit down to E_1 . The emitted photon has the *same energy*, since $h\nu = E_2 - E_1$, it is *in phase* with the incoming photon, it is in the *same direction*, and it has the *same polarization*, that is, the optical field oscillations in the emitted and in the incoming photon are in the same direction, as illustrated in Figure 4.1 (c). To get a feel of what is happening during stimulated emission one can think of the electric field of the incoming photon coupling to the electron and thereby driving it with the same frequency as the incoming photon. The forced oscillation of the electron at a frequency $\nu = (E_2 - E_1)/h$ causes it to emit electromagnetic radiation whose electric field is in total phase with that of the stimulating radiation. When the incoming photon leaves the site, the electron has returned to E_1 because it has emitted a photon of energy $h\nu = E_2 - E_1$. Although we considered the

²Also called *induced emission*.

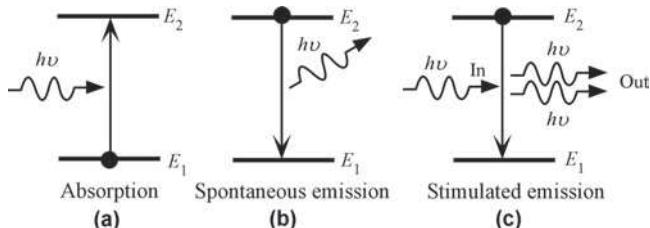


FIGURE 4.1 Absorption, spontaneous (random photon) emission, and stimulated emission.

transitions of an electron in an atom, we could have just well described photon absorption, and spontaneous and stimulated emission in Figure 4.1 in terms of energy transitions of the atom itself, in which case E_1 and E_2 represent the energy levels of the atom.

Stimulated emission is the basis for obtaining photon amplification since one incoming photon results in two outgoing photons which are in phase. How does one achieve a practical light amplifying device based on this phenomenon? It should be quite apparent from Figure 4.1 (c) that to obtain stimulated emission, the incoming photon should *not* be absorbed by another atom at E_1 . When we are considering a collection of atoms to amplify light, we must therefore have the majority of the atoms at the energy level E_2 . If this were not the case, the incoming photons would be absorbed by the atoms at E_1 . When there are more atoms at E_2 than at E_1 , we have what is called a **population inversion**. Under normal equilibrium conditions, as a result of Boltzmann statistics, most of the atoms would be at E_1 , and very few at E_2 . We therefore need to *excite* the atoms, cause population inversion, to obtain stimulated emission. It should be apparent that with *only two* energy levels, we can never achieve population at E_2 greater than that at E_1 because, in the steady state, the incoming photon flux will cause as many upward excitations as downward stimulated emissions. We need at least three energy levels as shown below in Section B.

B. Photon Amplification and Laser Principles

Let us consider the three-energy-level system shown in Figure 4.2 (a), with levels at E_1 , E_2 , and E_3 . There are, of course, other energy levels but they are not involved in the following processes. The present example will roughly represent what happens in a ruby laser. The energy levels

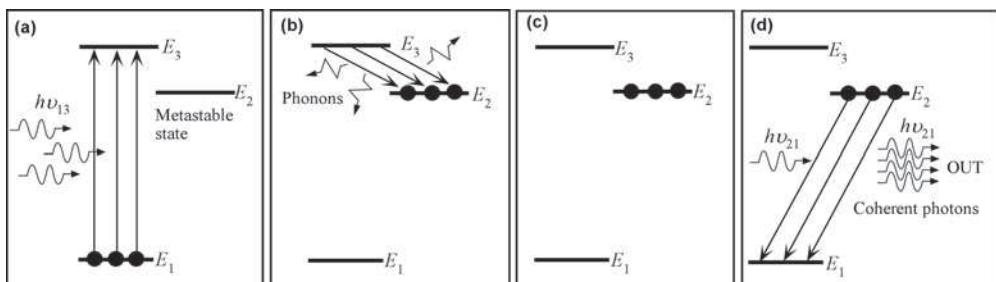


FIGURE 4.2 The principle of the laser, using a ruby laser as an example. (a) The ions (Cr^{3+}) in the ground state are pumped up to the energy level E_3 by photons from an optical excitation source. (b) Ions at E_3 rapidly decay to the long-lived state at the energy level E_2 by emitting lattice vibrations (phonons). (c) As the states at E_2 are long-lived, they quickly become populated and there is a population inversion between E_2 and E_1 . (d) A random photon (from spontaneous decay) of energy $h\nu_{21} = E_2 - E_1$ can initiate stimulated emission. Photons from this stimulated emission can themselves further stimulate emissions, leading to an avalanche of stimulated emissions and hence coherent photons being emitted.

E_1 , E_2 , and E_3 correspond to the energies of the chromium ions (Cr^{3+}) in a crystal of Al_2O_3 (a corundum crystal); E_1 is the ground energy of the Cr^{3+} ion. Suppose that an external optical excitation causes the Cr^{3+} ions in the crystal to become excited to the energy level E_3 . This is called the pump energy level, and the process of exciting the Cr^{3+} ions to E_3 is called **pumping**. In the present case, **optical pumping** is used, and we need a pump source with a photon energy of $h\nu_{13} = E_3 - E_1$.

From E_3 , the Cr^{3+} ions decay rapidly to an energy level E_2 by emitting phonons,³ that is, lattice vibrations (heat), as shown in Figure 4.2 (b). Thus, the transition from E_3 to E_2 is a radiationless transition. The energy level E_2 happens to correspond to a state that does not rapidly and spontaneously decay to the lower-energy state E_1 . In other words, the state at E_2 is a **long-lived state**.⁴ (The characteristic transition time constant from E_2 to E_1 is about 3 ms, which is very long in the atomic world.) Since the Cr^{3+} ions cannot decay rapidly from E_2 to E_1 , they accumulate at this energy level. With ample pumping, we can accumulate sufficient Cr^{3+} ions at E_2 to cause a population inversion between E_2 and E_1 , as pumping takes more and more Cr^{3+} ions to E_3 and hence E_2 . Population inversion is shown in Figure 4.2 (c).

When a Cr^{3+} ion at E_2 decays spontaneously, it emits a photon (a “random photon”) which can go on to a neighboring Cr^{3+} ion and cause that to execute stimulated emission. The photons from the latter can go on to the next Cr^{3+} ion at E_2 and cause that to emit by stimulated emission and so on. The result is an avalanche of stimulated emission processes with all the photons in phase, in the same direction, and with the same polarization, so that the light output is a large collection of *coherent photons* as in Figure 4.2 (d). This is the basic principle of the ruby laser, and hence the three-level laser. At the end of the avalanche of stimulated emission processes, the Cr^{3+} ions at E_2 would have dropped to E_1 and can be pumped again to repeat the stimulated emission cycle. The emission from E_2 to E_1 is called the **lasing emission**.⁵ The system we have just described for photon amplification is a laser, an acronym for **Light Amplification by Stimulated Emission of Radiation**.⁶

A more realistic energy diagram for the Cr^{3+} ion in the ruby crystal is shown in Figure 4.3 (a), which is typical of three-level laser systems that involve ions in crystals or glasses. The main difference from the simple schematic in Figure 4.2 is that the higher energy level E_3 is actually a band of energy levels. There is another band of energies around E'_3 . The Cr^{3+} ion can be excited by green light to E_3 and by blue light to E'_3 . (The ruby crystal is “pink” because it absorbs green and blue light.⁷) Pumping is achieved by using an intense xenon flash light. The ions become excited to the E_3 and E'_3 bands and then decay quickly from these bands by emitting phonons. They populate the E_2 -level as described above. Once more than half the ions at E_1 have been pumped, there is population inversion between E_1 and E_2 and consequently the pumped medium (the ruby crystal) exhibits **optical gain**, or **photon amplification**.

³ A phonon is a quantum of lattice vibrational energy, just as a photon is a quantum of electromagnetic energy.

⁴ We will not examine what causes certain states to be long-lived, but simply accept that these states do not decay rapidly by spontaneous emission to lower-energy states. Put differently, the decay time is very long (typically in milliseconds) on the atomic scale.

⁵ Arthur Schawlow, one of the coinventors of the laser, was well-known for his humor and has, apparently, said that “Anything will lase if you hit it hard enough.” In 1971, Schawlow and Ted Hänsch were able to develop the first edible laser made from Jell-O. (T. Hansch, M. Pernier, A. Schawlow, “Laser Action of Dyes in Gelation,” *IEEE J. Quantum Elect.*, 7, 45, 1971.)

⁶ Although it is indeed an acronym, it is now commonly used as a noun that implies a device that emits coherent radiation.

⁷ The ruby crystal appears red if the Cr^{3+} ions concentration is large. For the Cr^{3+} concentrations used in the ruby laser, the crystal appears pink.

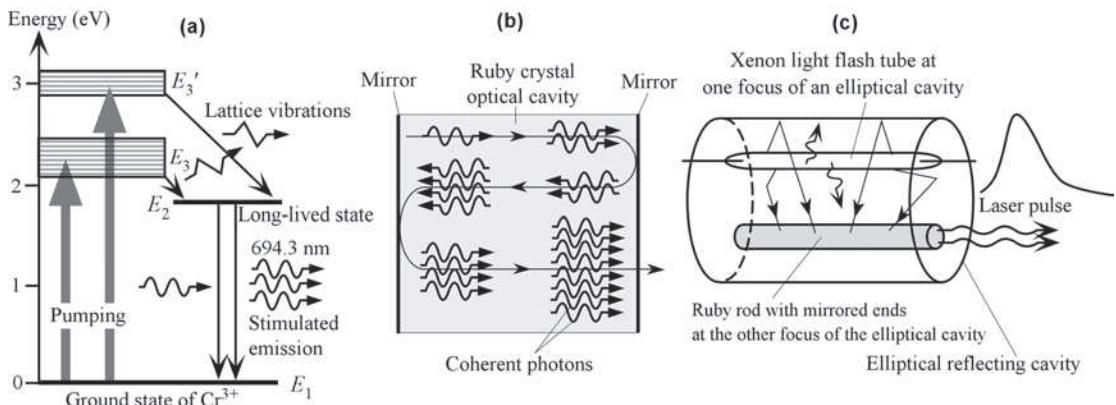
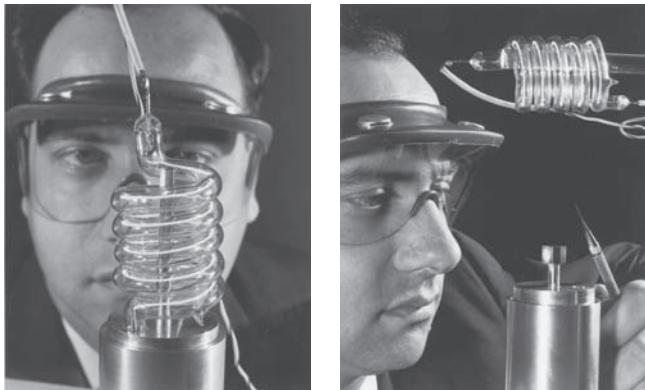


FIGURE 4.3 (a) A more realistic energy diagram for the Cr^{3+} ion in the ruby crystal (Al_2O_3), showing the optical pumping levels and the stimulated emission. (b) The laser action needs an optical cavity to reflect the stimulated radiation back and forth to build up the total radiation within the cavity, which encourages further stimulated emissions. (c) A typical construction for a ruby laser, which uses an elliptical reflector, and has the ruby crystal at one focus and the pump light at the other focus.

To get coherent lasing radiation out from the ruby laser, we need to do more, not just amplify the radiation along the ruby crystal. We can increase stimulated emissions by increasing the number of photons, *i.e.*, the radiation intensity, within the crystal inasmuch as more photons cause more stimulated emissions. The ends of the ruby crystal, which is normally a rod, are silvered to reflect back and forward the stimulated radiation, that is, to form an **optical cavity** with mirrors at the ends, as shown Figure 4.3 (b). As the stimulated photons are reflected back into the crystal, the radiation intensity buildup inside the crystal, in much the same way we build up voltage oscillations in an electrical oscillator circuit by feedback. The buildup of coherent radiation in the cavity encourages further stimulated emissions, until a large avalanche of stimulated transitions occur and takes most of the ions at E_2 down to E_1 . One of the mirrors is partially silvered to allow some of this radiation to be tapped out. What comes out is a pulse of highly coherent radiation that has a high intensity as illustrated in Figure 4.3 (b).

The coherency and the well-defined wavelength of the emitted radiation from a laser are attributes that make it distinctly different from a random stream of different wavelength photons emitted from a tungsten bulb, or randomly phased photons from an LED. The reader might have noticed that the photon energy emitted from the laser system is less than the photon energy we used to pump it, that is, excite it; $h\nu_{21} < h\nu_{13}$. However, we only needed incoherent radiation to pump the system, and we obtained a fully coherent radiation as output.

Practical ruby lasers need efficient optical pumping, which can be obtained by using an elliptical reflector with the ruby crystal rod at one focus, and the pump light, a xenon flash, at the other focus as shown in Figure 4.3 (c). The early ruby lasers used a helical xenon flash tube surrounding the ruby rod. The lasing emission from the ruby laser is a light pulse, whose duration and intensity depend on the laser construction, and the xenon flash. Ruby lasers are frequently used in interferometry, holography, and hair and tattoo removal, among other applications.



Theodore Harold Maiman was born in 1927 in Los Angeles, son of an electrical engineer. He studied engineering physics at Colorado University, while repairing electrical appliances to pay for college, and then obtained a Ph.D. from Stanford. Theodore Maiman constructed this first laser in 1960 while working at Hughes Research Laboratories (T.H. Maiman, "Stimulated optical radiation in ruby lasers", *Nature*, **187**, 493, 1960). There is a vertical chromium ion doped ruby rod in the center of a helical xenon flash tube. The ruby rod has mirrored ends. The xenon flash provides optical pumping of the chromium ions in the ruby rod. The output is a pulse of red laser light. (*Courtesy of HRL Laboratories, LLC, Malibu, California.*)

C. Four-Level Laser System

The problem with the three-level laser system in Figure 4.2 (a) is that to achieve population inversion, we must pump at least half the Cr^{3+} ions at the ground level E_1 to E_2 . It would be much better if we could achieve population inversion as soon as we begin exciting the ions at E_1 to E_3 to E_2 . We can quickly and easily achieve population inversion if we use a four-level system as in Figure 4.4, where the ground energy is E_0 . Figure 4.4 roughly represents the energy levels of Nd^{3+} ions in a $\text{Y}_3\text{Al}_5\text{O}_{11}$ (yttrium aluminate garnet, YAG) crystal, usually written as $\text{Nd}^{3+}:\text{YAG}$. In equilibrium, the population of the ions at these energy levels follow the Boltzmann distribution; and nearly all the ions are at E_0 whereas E_1 , E_2 , and E_3 are mostly empty.

The ions at E_0 are excited by optical pumping to the energy level E_3 . In many cases, this is a band of energy levels and is called the **pump band**. From E_3 , the ions decay rapidly by phonon emission to E_2 , which is a long-lived state with a lifetime in the milliseconds, *i.e.*, the spontaneous decay from E_2 to E_1 is slow. Clearly, as soon as ions begin to populate E_2 we have population inversion between the level E_1 and E_2 inasmuch as E_1 is initially nearly empty. The lasing emission takes place by stimulated emission as ions drop from E_2 to E_1 . From E_1 , the ions return back to E_0 by the emission of phonons (or spontaneous photons in some cases). We also need the ions at E_1 to decay quickly back to E_0 so that a buildup of ions at E_1 is avoided, and the ions are quickly returned back to E_0 for repumping. In cases where E_1 is within only a few $k_B T$ to E_0 , the four-level system behaves almost like the three-level system in Figure 4.2.

The four-level laser system in Figure 4.4 serves as an excellent way to explain the majority of practical lasers. Indeed, most modern lasers are based on a four-level system, or its equivalent. For example, in the $\text{Nd}^{3+}:\text{YAG}$ laser, the Nd^{3+} ions are excited by using a flash tube or a laser diode (LD) to E_3 and hence to E_2 , which is the *upper* laser level. The lifetime of the ions at E_2 is about 230 μs . The *lower* laser level is E_1 . The lasing emission from E_2 to E_1 is in the infrared at 1064 nm. The $\text{Nd}^{3+}:\text{YAG}$ can be operated in the pulsed mode as well as in CW mode. In contrast, the ruby laser is operated in the pulsed mode because of the large power required for pumping the ruby crystal.

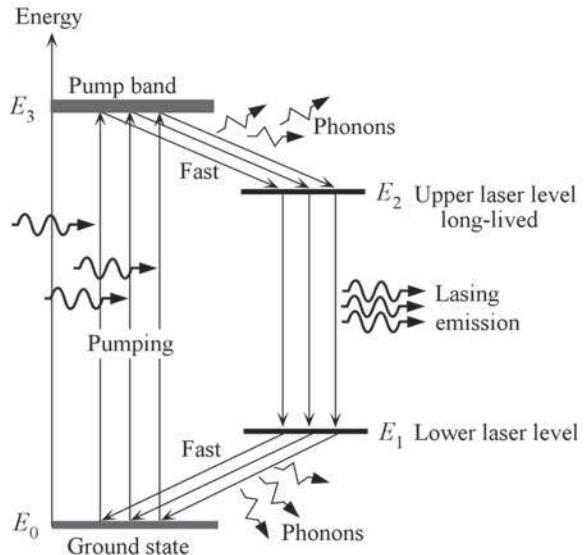


FIGURE 4.4 A four-energy-level laser system. Highly simplified representation of Nd³⁺:YAG laser.

4.2 STIMULATED EMISSION RATE AND EMISSION CROSS-SECTION

A. Stimulated Emission and Einstein Coefficients

A useful laser medium must have a higher efficiency of stimulated emission compared with the efficiencies of spontaneous emission and absorption. We need to determine the controlling factors for the rates of stimulated emission, spontaneous emission, and absorption. Consider a medium as in Figure 4.5 that has N_1 atoms per unit volume at the energy E_1 and N_2 atoms per unit volume at the energy E_2 . Then the rate of upward transitions from E_1 to E_2 by photon absorption will be proportional to the concentration of atoms N_1 , and also to the number of photons per unit volume with energy $h\nu = E_2 - E_1$. Put differently, this rate will depend on the energy density in the radiation. Thus, the upward transition rate R_{12} is

Upward transitions

$$R_{12} = B_{12}N_1\rho(v) \quad (4.2.1)$$

where B_{12} is a proportionality constant termed the **Einstein B_{12} coefficient**, and $\rho(v)$ is the photon energy density per unit frequency,⁸ that is, the energy in the radiation per unit volume

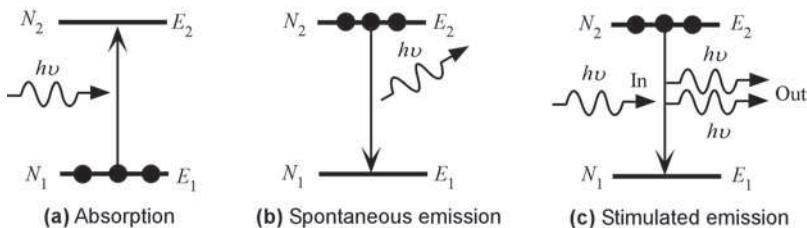


FIGURE 4.5 Absorption, spontaneous emission, and stimulated emission.

⁸Using $\rho(v)$ defined in this way simplifies the evaluation of the proportionality constants in this section. Further, in the simplified treatise here, we neglected the degeneracy of the energy levels, which introduces numerical factors that multiply N_1 and N_2 .

per unit frequency due to photons with energy $h\nu = E_2 - E_1$. The rate of downward transitions from E_2 to E_1 involves spontaneous and stimulated emission. The spontaneous emission rate depends on the concentration N_2 of atoms at E_2 . The stimulated emission rate depends on both N_2 and the photon concentration $\rho(\nu)$ with energy $h\nu (= E_2 - E_1)$. Thus, the total downward transition rate is

$$R_{21} = A_{21}N_2 + B_{21}N_2\rho(\nu) \quad (4.2.2) \quad \text{Downward transitions}$$

where the first term is due to spontaneous emission [does not depend on the photon energy density $\rho(\nu)$ to drive it] and the second term is due to stimulated emission, which requires photons to drive it. A_{21} and B_{21} are the proportionality constants termed the **Einstein coefficients** for spontaneous and stimulated emissions, respectively.

To find the coefficients A_{21} , B_{12} , and B_{21} , we consider the events in equilibrium, that is, the medium in thermal equilibrium (no external excitation). There is no net change with time in the populations at E_1 and E_2 which means

$$R_{12} = R_{21} \quad (4.2.3)$$

Furthermore, in thermal equilibrium, Boltzmann statistics demands that

$$N_2/N_1 = \exp[-(E_2 - E_1)/k_B T] \quad (4.2.4) \quad \text{Boltzmann statistics in equilibrium}$$

where k_B is the Boltzmann constant and T is the absolute temperature.

Now in *thermal equilibrium*, in the collection of atoms we are considering, radiation from the atoms must give rise to an equilibrium photon energy density, $\rho_{\text{eq}}(\nu)$, that is given by *Planck's black body radiation distribution law*⁹

$$\rho_{\text{eq}}(\nu) = \frac{8\pi h\nu^3}{c^3 \left[\exp\left(\frac{h\nu}{k_B T}\right) - 1 \right]} \quad (4.2.5) \quad \text{Planck's black body radiation in equilibrium}$$

It is important to emphasize that Planck's law in Eq. (4.2.5) applies only in thermal equilibrium; we are using this equilibrium condition to determine the Einstein coefficients. During the laser operation, of course, $\rho(\nu)$ is not described by Eq. (4.2.5); in fact it is much larger. From Eqs. (4.2.1) to (4.2.5) we can readily show that

$$B_{12} = B_{21} \quad (4.2.6) \quad \text{Absorption and stimulated emission coefficients}$$

and

$$A_{21}/B_{21} = 8\pi h\nu^3/c^3 \quad (4.2.7) \quad \text{Spontaneous and stimulated emission coefficients}$$

Now consider the ratio of stimulated to spontaneous emission

$$\frac{R_{21}(\text{stim})}{R_{21}(\text{spon})} = \frac{B_{21}N_2\rho(\nu)}{A_{21}N_2} = \frac{B_{21}\rho(\nu)}{A_{21}} \quad (4.2.8)$$

⁹See, for example, any modern physics textbook.

which, by Eq. (4.2.7), can be written as

Stimulated to spontaneous emission

$$\frac{R_{21}(\text{stim})}{R_{21}(\text{spon})} = \frac{c^3}{8\pi h\nu^3} \rho(v) \quad (4.2.9)$$

In addition, the ratio of stimulated emission to absorption is

Stimulated to absorption

$$\frac{R_{21}(\text{stim})}{R_{12}(\text{absorp})} = \frac{N_2}{N_1} \quad (4.2.10)$$

There are two important conclusions. For stimulated photon emission to exceed photon absorption, by Eq. (4.2.10), we need to achieve **population inversion**, that is, $N_2 > N_1$. For stimulated emission to far exceed spontaneous emission, by Eq. (4.2.9) we must have a large photon concentration, which is achieved by building an **optical cavity** to contain the photons.

The rate R_{12} represents the rate at which N_1 is decreasing by absorption, thus, $R_{12} = -dN_1/dt$. The rate R_{21} represents the rate at which N_2 is decreasing by spontaneous and stimulated emission, thus, $R_{21} = -dN_2/dt$. Further, the rate at which N_2 changes by spontaneous emission can be written as

Spontaneous decay lifetime

$$dN_2/dt = -A_{21}N_2 = -N_2/\tau_{\text{sp}} \quad (4.2.11)$$

where, by definition, $\tau_{\text{sp}} = 1/A_{21}$, and is termed the **spontaneous decay time**, or more commonly called the **lifetime** of level E_2 .

It is important to point out that the population inversion requirement $N_2 > N_1$ means that we depart from thermal equilibrium. According to Boltzmann statistics in Eq. (4.2.4), $N_2 > N_1$ implies a *negative temperature!* The laser principle is based on **non-thermal equilibrium**.¹⁰

The number of states an ion or an atom has at a given energy level is called its **degeneracy** g . We assumed that the degeneracy g_1 and g_2 of E_1 and E_2 are the same. We can, of course, redo the treatment above by including g_1 and g_2 . The final result is equivalent to replacing N_1 by $N_1(g_2/g_1)$.

EXAMPLE 4.2.1 Minimum pumping power for three-level laser systems

Consider the three-level system as shown in Figure 4.2 (a). Assuming that the transitions from E_3 to E_2 are fast, and the spontaneous decay time from E_2 to E_1 is τ_{sp} , show that the *minimum* pumping power P_{pmin} that must be absorbed by the laser medium per unit volume for population inversion ($N_2 > N_1$) is

Minimum pumping for population inversion for three-level laser

$$P_{\text{pmin}}/V = (N_0/2)h\nu_{13}/\tau_{\text{sp}} \quad (4.2.12)$$

where V is the volume, N_0 is the concentration of ions in the medium, and hence at E_0 before pumping. Consider a ruby laser in which the concentration of Cr^{3+} ions is 10^{19} cm^{-3} , the ruby crystal rod is 10 cm long and 1 cm in diameter. The lifetime of Cr^{3+} at E_2 is 3 ms. Assume the pump takes the Cr^{3+} ions to the E_3 -band in Figure 4.3 (a), which is about 2.2 eV above E_0 . Estimate the minimum power that must be provided to this ruby laser to achieve population inversion.

¹⁰“But I thought, now wait a minute! The second law of thermodynamics assumes thermal equilibrium. We don’t have that!” Charles D. Townes (born 1915; Nobel Laureate, 1964). The laser idea occurred to Charles Townes, apparently, while he was taking a walk one early morning in Franklin Park in Washington, DC, while attending a scientific committee meeting. Non-thermal equilibrium (population inversion) is critical to the principle of the laser.

Solution

Consider the three-level system in Figure 4.2 (a). To achieve population inversion we need to get half the ions at E_1 to level E_2 so that $N_2 = N_1 = N_0/2$, since N_0 is the total concentration of Cr^{3+} ions all initially at E_1 . We will need $[(N_0/2)h\nu_{13} \times \text{volume}]$ amount of energy to pump to the E_3 -band. The ions decay quickly from E_3 to E_2 . We must provide this pump energy before the ions decay from E_2 to E_1 , that is, before τ_{sp} . Thus, the *minimum* power the ruby needs to absorb is

$$P_{\text{pmin}} = V(N_0/2)h\nu_{13}/\tau_{\text{sp}}$$

which is Eq. (4.2.12). For the ruby laser

$$P_{\text{pmin}} = [\pi(0.5 \text{ cm})^2(10 \text{ cm})][(10^{19} \text{ cm}^{-3})/2][(2.2 \text{ eV})(1.6 \times 10^{-19} \text{ J/eV})]/(0.003 \text{ s}) = 4.6 \text{ kW}$$

The total pump energy that must be provided in less than 3 ms is 13.8 J.

B. Emission and Absorption Cross-Sections

In many photonic applications, we need to describe absorption and emission in terms of cross-sections. Consider a medium that has been doped with ions, perhaps Nd^{3+} ions in YAG or glass, and consider two energy levels E_1 and E_2 , as in Figure 4.6 (a). Suppose that there is a flux of photons Φ_{ph} passing through the medium along a particular direction (x), and the photon energy $h\nu = E_2 - E_1$. The flux is the number of photons flowing per unit area along a certain direction so that $\Phi_{\text{ph}} = I/h\nu$, where I is the intensity of the beam¹¹ (radiation energy flow per unit area per unit time). The absorption of radiation, that is, photons, in the medium depends on two factors. First is the number of ions per unit volume, N_1 ; higher concentration N_1 will mean more absorption. Second is the nature of the interaction of the photon with an individual ion, and is related to the properties of the ion in the medium. The interaction of a photon with an individual ion can be intuitively visualized as the ion having a certain *cross-sectional area*, σ_{ab} , and if the photon

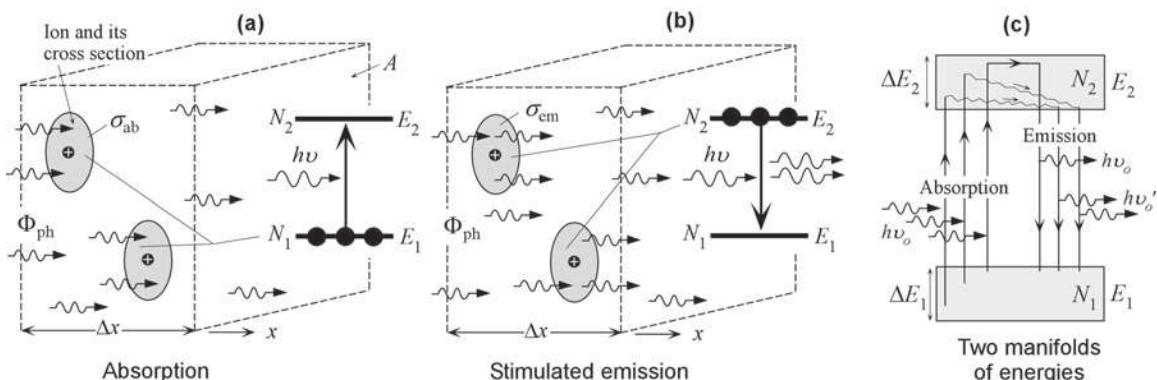


FIGURE 4.6 (a) Absorption of photons by ions in a medium, and the illustration of the absorption cross-section σ_{ab} . (b) Stimulated emission of photons by the same ions in the medium, and the illustration of the emission cross-section σ_{em} . (c) Upper and lower manifolds of energy. Each manifold has numerous closely spaced sublevels. Absorption and emission can involve the same or different levels.

¹¹ I (san-serif font) is used for the light intensity (optical power flow per unit area, W m^{-2}) and I for the current in this chapter.

crosses this area, it becomes absorbed, as illustrated in Figure 4.6 (a). The optical power absorbed by a single ion depends on the light intensity incident *and* the ion's absorption cross-section, *i.e.*,

Definition of absorption cross-section

$$\begin{aligned} \text{Optical power absorbed by an ion} &= \text{Light intensity} \\ &\times \text{Absorption cross-section of ion} = I\sigma_{ab} \end{aligned} \quad (4.2.13)$$

which defines the **absorption cross-section** σ_{ab} . For the collection of N_1 ions per unit volume, the total power absorbed per unit volume is $I\sigma_{ab}N_1$.

Consider a small volume $A\Delta x$ of the medium, as in Figure 4.6 (a), where A is the cross-sectional area and the radiation propagates along x . The total optical power ΔP_o absorbed in this volume is then $(I\sigma_{ab}N_1)(A\Delta x)$. Suppose that ΔI is the *change* (negative if absorption) in the intensity of the propagating radiation over the distance Δx . Then

$$A\Delta I = -\text{Total power absorbed in } A\Delta x = -(I\sigma_{ab}N_1)(A\Delta x)$$

i.e.,

Absorption cross-section σ_{ab} and coefficient α

$$-\frac{\Delta I}{\Delta x} = \sigma_{ab}N_1 = \alpha \quad (4.2.14)$$

where the left-hand side is the fractional change in the light intensity per unit distance, and therefore represents α , the **absorption coefficient** of the medium, due to the absorption of radiation by ions at E_1 , which reach E_2 . Clearly, $\alpha = \sigma_{ab}N_1$. We could have easily used the latter as the definition of σ_{ab} and derived Eq. (4.2.13) instead.

The stimulated emission case, shown in Figure 4.6 (b), is similar. We can define an **emission cross-section** σ_{em} such that if an incident photon crosses the area σ_{em} , it stimulates the ion to emit a photon as shown in Figure 4.6 (b). In a similar way to Eq. (4.2.13)

Definition of emission cross-section

$$\begin{aligned} \text{Stimulated optical power emitted by an ion} &= \text{Light intensity} \\ &\times \text{Emission cross-section of ion} = I\sigma_{em} \end{aligned} \quad (4.2.15)$$

As the radiation propagates along x , its intensity I increases due to stimulated emissions from E_2 to E_1 . Since only ions at E_2 with a concentration N_2 can be stimulated to emit, the equivalent of Eq. (4.2.14) for the fractional increase in the intensity per unit distance is

Emission cross-section σ_{em} and optical gain

$$\frac{\Delta I}{\Delta x} = \sigma_{em}N_2 \quad (4.2.16)$$

The net fractional increase in the intensity I per unit distance is defined as the **gain coefficient** g of the medium, so that, in the presence of stimulated emission and absorption

Net optical gain coefficient

$$g = \left[\frac{\Delta I}{\Delta x} \right]_{\text{net}} = \sigma_{em}N_2 - \sigma_{ab}N_1 \quad (4.2.17)$$

In the presence of just two discrete energy levels, as in Figure 4.6 (a) and (b), we know that $B_{12} = B_{21}$, and similarly, $\sigma_{ab} = \sigma_{em}$.

However, in many practical cases, the two energy levels are actually two manifolds of energies, that is, each manifold consists of narrowly spaced sublevels, as illustrated in Figure 4.6 (c). The lower E_1 -manifold has sublevels that cover the range ΔE_1 , as in Figure 4.6 (c), and the higher E_2 -manifold has sublevels that are within ΔE_2 . Suppose we excite the ions with photons of energy $h\nu_o$ that are absorbed and take some of ions from the E_1 - to the E_2 -manifold. Three of the

possibilities for absorption are shown in Figure 4.6 (c) that take ions from different sublevels in ΔE_1 and put them at various sublevels in ΔE_2 in which $h\nu_o$ is the same for all. However, the downward transitions are more selective because the end level in the E_1 -manifold must be empty, and only the higher levels in this manifold (top of ΔE_1) are less populated. Thus, there are less chances for the downward transitions that emit the same $h\nu_o$ as shown in Figure 4.6 (c). In addition, the excited ions in the E_2 -manifold decay rapidly by emitting lattice vibrations (they thermalize) to the lower levels in E_2 and then emit photons with $h\nu'_o < h\nu_o$. Clearly, the net effect is that emission at ν_o is less likely than absorption at the same frequency. Consequently, σ_{em} and σ_{ab} are then *not* the same at ν_o ; and σ_{ab} is larger than σ_{em} in this example.

We can consider other photon energies besides $h\nu_o$ in Figure 4.6 (c), and hence consider σ_{ab} and σ_{em} at other frequencies. We must therefore generalize Eq. (4.2.17) and write σ_{ab} and σ_{em} as frequency dependent, *i.e.*,

$$g(v) = \sigma_{\text{em}}(v)N_2 - \sigma_{\text{ab}}(v)N_1 \quad (4.2.18)$$

Optical gain coefficient

Although Eq. (4.2.18) has been derived in terms of v as a variable, it could have just as well be written in terms of λ .

When the net gain in Eq. (4.2.18) is zero, the medium becomes **transparent** to the incoming radiation, which is neither absorbed nor amplified. The **transparency condition** is

$$\sigma_{\text{em}}(v)N_2 - \sigma_{\text{ab}}(v)N_1 = 0 \quad (4.2.19)$$

Transparency condition

To find the overall optical power (or intensity) gain G through a medium that has been pumped or excited, we need to integrate Eq. (4.2.17) along the length of the medium, from $x = 0$ to $x = L$. If N_2 and N_1 remain the same over L and hence g is constant along L , then $G = I(L)/I(0)$ is given by

$$G = \exp(gL) \quad (4.2.20)$$

Net optical power gain

Henceforth we will use the above equations in describing optical amplifiers.

EXAMPLE 4.2.2 Gain coefficient in a Nd³⁺-doped glass fiber

Consider a Nd³⁺-doped silica-based glass fiber. Suppose we want to use this fiber as an optical amplifier. The Nd³⁺ concentration in the fiber core is $1 \times 10^{19} \text{ cm}^{-3}$. The Nd³⁺ ions can be pumped from E_0 to E_2 in Figure 4.4 from which they decay rapidly down to E_1 , and the population inversion between E_2 and E_1 gives optical amplification. The E_2 to E_1 emission is approximately at $1.05 \mu\text{m}$, and the emission cross-section is about $3 \times 10^{-20} \text{ cm}^2$. Estimate the maximum gain coefficient with this amount of doping if we could pump all the Nd³⁺ ions. What is gain G in dB if the fiber is 20 cm long?

Solution

The maximum gain will be achieved when all Nd³⁺ ions have been pumped to E_2 . From Eq. (4.2.17)

$$g = \sigma_{\text{em}}N_2 = (3 \times 10^{-20} \text{ cm}^2)(1 \times 10^{19} \text{ cm}^{-3}) = 0.3 \text{ cm}^{-1} \quad \text{or} \quad 30 \text{ m}^{-1}$$

This gain increases with the Nd³⁺ concentration. The maximum theoretical gain G_{max} for a 20 cm fiber is

$$G = \exp(gL) = \exp[(0.3 \text{ cm}^{-1}) \times (20 \text{ cm})] = 403.4 \quad \text{or} \quad 26.1 \text{ dB}$$

4.3 ERBIUM-DOPED FIBER AMPLIFIERS

A. Principle of Operation and Amplifier Configurations

A light signal that is traveling along an optical fiber over a long distance suffers marked attenuation. It becomes necessary to regenerate the light signal at certain intervals for long-haul communications over several thousand kilometers. Instead of regenerating the optical signal by photodetection, conversion to an electrical signal, amplification, and then conversion back from electrical to light energy by a laser diode, it becomes practical to amplify the signal directly by using an optical amplifier.

One practical **optical amplifier** is based on the **erbium (Er^{3+} ion)-doped fiber amplifier (EDFA)**.¹² The core region of an optical fiber is doped with Er^{3+} ions. The host fiber core material is a glass based on silica-aluminate ($\text{SiO}_3\text{-Al}_2\text{O}_3$) or silica-germania ($\text{SiO}_3\text{-GeO}_2$), or both; other glass hosts have also been used. It is easily fused to a single-mode long-distance optical fiber by a technique called *splicing*.

When the Er^{3+} ion is implanted in the host glass material it has the energy levels indicated in Figure 4.7 (a), which appears as though it is a three-level system, but actually mimics a pseudo four-level system.¹³ The energies of the Er^{3+} ion in the glass actually fall into *manifolds of energy levels*, which are shown as 1, 2, and 3. Each manifold has several closely spaced energy levels, which are themselves broadened. When an isolated Er^{3+} ion is placed in a glass structure, it finds itself interacting with neighboring host ions. The electric field from these neighboring ions splits the Er^{3+} energy levels in much the same way an applied magnetic field splits the energy levels in the H-atom. In the present case, the energy splitting is due to the local electric field from the ions

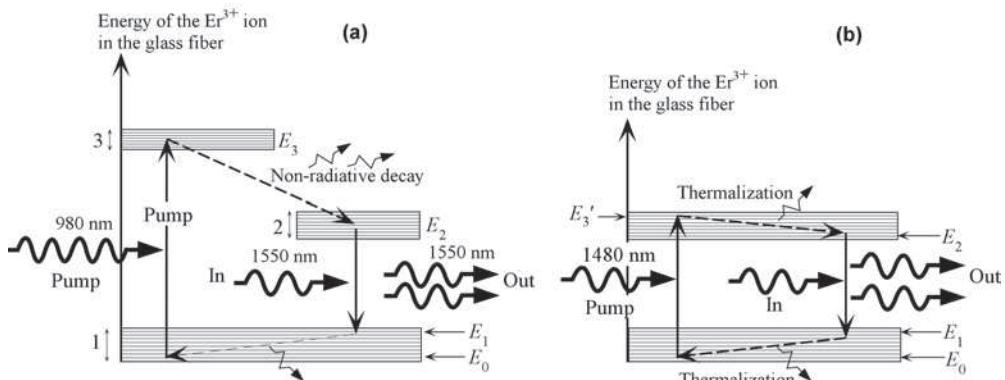


FIGURE 4.7 (a) Energy diagram for the Er^{3+} ion in the glass fiber medium and light amplification by stimulated emission from E_2 to E_1 (features are highly exaggerated). Dashed arrows indicate radiationless transitions (energy emission by lattice vibrations). The pump is a 980-nm laser diode. (b) EDFA can also be pumped with a 1480-nm laser diode.

¹²EDFA was first reported in 1987 by E. Desurvire, J. R. Simpson, and P. C. Becker, and in 1994 AT&T began deploying EDFA repeaters in long-haul fiber communications.

¹³The operation of the Er^{3+} -doped glass fiber laser can be easily explained by using the three-level laser system in Figure 4.2 (a) since the spread between E_1 and E_0 is comparable to $k_B T$; but a somewhat better explanation would be to treat it as a quasi four-level system as in Figure 4.4. The nonradiative transitions within a manifold are very fast; these are thermalization processes that take typically a few picoseconds.

surrounding the Er^{3+} , and the energy splitting is called the **Stark effect**. Further, the environment of each Er^{3+} ion in the glass is not identical since the glass, unlike a crystal, does not have a well-defined periodic structure; there are local variations in composition and density. The Stark energy levels are therefore broadened as well. The positions and spreads of actual energy levels within a manifold depend on the glass host.

The E_1 -manifold is the lowest manifold of energies and is labeled 1 in Figure 4.7 (a); it spreads from E_0 to E_1 , and represents the lowest energy levels possible for the Er^{3+} ion.¹⁴ Although most Er^{3+} ions are at E_0 , from Boltzmann statistics, there are also Er^{3+} ions at higher levels in this manifold since the manifold is quite narrow, typically about 0.03–0.04 eV (comparable to $k_B T$ at room temperature); the population ratio depends on the spread $E_1 - E_0$ but is very roughly 1 to 4.

There is a higher energy manifold, shown as 3, of narrowly spaced energy levels labeled E_3 , as shown in Figure 4.7 (a), that is approximately 1.27 eV above the ground E_1 -manifold. The Er^{3+} ions are optically pumped, usually from a laser diode, to excite them from the ground energy manifold 1 to 3. The wavelength for this pumping is about 980 nm. The Er^{3+} ions then rapidly and nonradiatively decay from E_3 to a long-lived energy level E_2 within the manifold 2 of energies in Figure 4.7 (a). The decay from E_3 to E_2 involves energy losses by the emission of phonons, and is very rapid. The Er^{3+} ions at E_2 have a long lifetime ~ 10 ms. Thus, more and more Er^{3+} ions accumulate at E_2 , which is roughly 0.80 eV above the energy level E_1 at the top of the manifold 1. The accumulation of Er^{3+} ions at E_2 eventually leads to a population inversion between E_2 and E_1 inasmuch as E_1 is sparsely populated. Signal photons around 1550 nm have an energy of 0.80 eV, or $E_2 - E_1$, and give rise to *stimulated transitions* of Er^{3+} ions from E_2 to E_1 . The transition from E_1 to E_0 is by phonon emission and occurs rapidly (over picoseconds); it corresponds to the thermalization of Er^{3+} ions in this manifold which is narrow. Without pumping, there will be some Er^{3+} ions at E_1 that will *absorb* the incoming 1550 nm photons and reach E_2 . To achieve light amplification we must therefore have stimulated emission exceeding absorption. This is only possible if there are more Er^{3+} ions at E_2 than at E_1 ; that is, if we have population inversion. (As explained below we also need to consider absorption and emission cross-sections.)

It is also possible to pump the Er^{3+} ions to the top of the manifold 2 in Figure 4.7 (a), which is shown in Figure 4.7 (b). Notice that the top of this manifold now acts like a E_3 -level, and is labeled E'_3 . The pump wavelength is 1480 nm. The Er^{3+} ions decays rapidly by emitting phonons to E_2 , and then by stimulated emission down to E_1 . Pumping at 980 nm is more efficient than pumping at 1480 nm.

To understand the basic operation of the EDFA we need to consider the absorption and emission cross-sections σ_{ab} and σ_{em} and their spectral dependence, which are shown in Figure 4.8 (a).¹⁵ The net gain $g(\lambda)$ at λ in Eq. (4.2.18) depends on $\sigma_{em}N_2 - \sigma_{ab}N_1$. Under strong pumping, there will be more Er^{3+} in the E_2 -manifold than in the E_1 -manifold, the emission term $\sigma_{em}N_2$ will dominate over $\sigma_{ab}N_1$ ($N_2 \gg N_1$) and hence the net gain coefficient will be $g(\lambda) \approx \sigma_{em}(\lambda)N_2 \approx \sigma_{em}(\lambda)N_0$, where N_0 is the Er^{3+} concentration in the core of the fiber. Thus, σ_{em} vs. λ also represents the

¹⁴The valence electrons of the Er^{3+} ion are arranged to satisfy the Pauli exclusion principle and Hund's rule, and this arrangement has the energies lying between E_0 and E_1 .

¹⁵If this were a simple two energy level system, $\sigma_{em} = \sigma_{ab}$ inasmuch as $B_{12} = B_{21}$. It would be impossible to optically pump from E_1 to E_2 and achieve net gain. However, the E_1 and E_2 are manifolds of energy levels. In each manifold we have many levels that have been broadened and overlap.

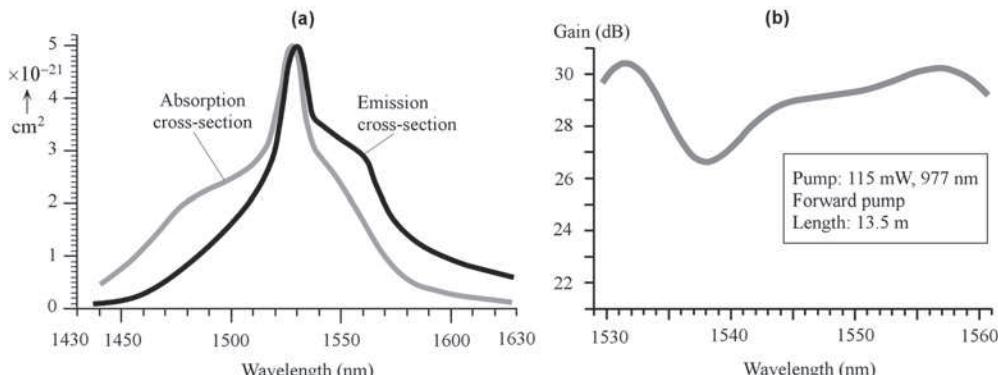


FIGURE 4.8 (a) Typical absorption and emission cross-sections, σ_{ab} and σ_{em} , respectively, for Er^{3+} in a silica glass fiber doped with alumina ($\text{SiO}_2\text{-Al}_2\text{O}_3$). (Cross-section values for the plots were extracted from B. Pedersen *et al.*, *J. Light Wave Technol.*, 9, 1105, 1991.) (b) The spectral characteristics of gain, G in dB, for a typical commercial EDF, available from Fibercore as IsoGain™ fiber. Forward pumped at 115 mW and at 977 nm. The insertion losses are 0.45 dB for the isolator, 0.9 dB for the pump coupler and splices.

spectral dependence of the gain coefficient under strong pumping. Clearly, the EDFA response is not “flat”; there is a peak around 1530 nm. The absolute maximum gain with nearly full inversion would be $G = \exp(\sigma_{em}N_0L)$. We can quickly estimate the maximum possible gain at 1550 nm. For an Er^{3+} concentration (N_0) of 10^{19} cm^{-3} , at 1550 nm, using $\sigma_{em} \approx 3.2 \times 10^{-21} \text{ cm}^2$ from Figure 4.8 (a), $g \approx \sigma_{em}N_0 = 3.2 \text{ m}^{-1}$, and the optical gain G for a fiber length of 1 m is 24.5 or 13.9 dB.

If, on the other hand, the EDFA is left unpumped, the Er^{3+} ions would be in the ground state, and the signal will be absorbed with an attenuation coefficient $\alpha = \sigma_{ab}N_0$ as in Eq. (4.2.14). From Figure 4.8 (a), $\sigma_{ab} \approx 2.4 \times 10^{-21} \text{ cm}^2$ and hence $\alpha = \sigma_{ab}N_0 = 2.4 \text{ m}^{-1}$ or 10.4 dB m^{-1} . Over the length of the EDFA, $\sim 10 \text{ m}$, this would significantly extinguish the signal. Thus, the EDFA must be pumped at all times.

The spectral shapes of σ_{ab} and σ_{em} in Figure 4.8 (a) depend on the host glass composition and vary significantly between host glasses. It is possible to broaden the bandwidth of the gain curve (*i.e.*, σ_{em} vs. λ) by appropriately choosing the glass composition. Figure 4.8 (b) shows typical spectral characteristics of gain, G in dB, for a commercial erbium-doped fiber that has been pumped with a laser diode at 980 nm with 115 mW. The spectrum is not flat and is normally flattened to obtain uniform gain over the band of wavelengths used in multi channel communications; this is described below.

In practice, the erbium-doped fiber is inserted into the fiber communications line by splicing as shown in the simplified schematic diagram in Figure 4.9, and it is pumped from a laser diode through a coupling fiber arrangement which allows only the pumping wavelength to be coupled; it is a **wavelength-selective coupler**. Some of the Er^{3+} ions at E_2 will decay spontaneously from E_2 to E_1 , which will give rise to unwanted noise in the amplified light signal. Further, if the EDFA is not pumped at any time, it will act as an attenuator as the 1550 nm photons will be absorbed by Er^{3+} ions, which will become excited from the ground energy manifold 1 to the E_2 -manifold. In returning back to ground energy by spontaneous emission, they will emit 1550 nm photons randomly, which generates noise in the optical link. **Optical isolators** inserted at the entry and exit end of the amplifier allow only the optical signals at 1550 nm to pass in one direction and prevent the 980 pump light from propagating back or forward into the communication system.

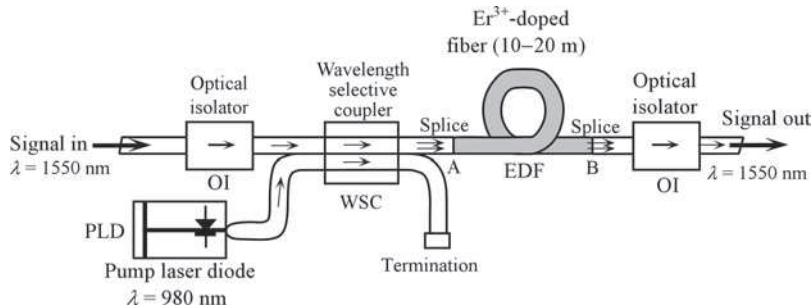
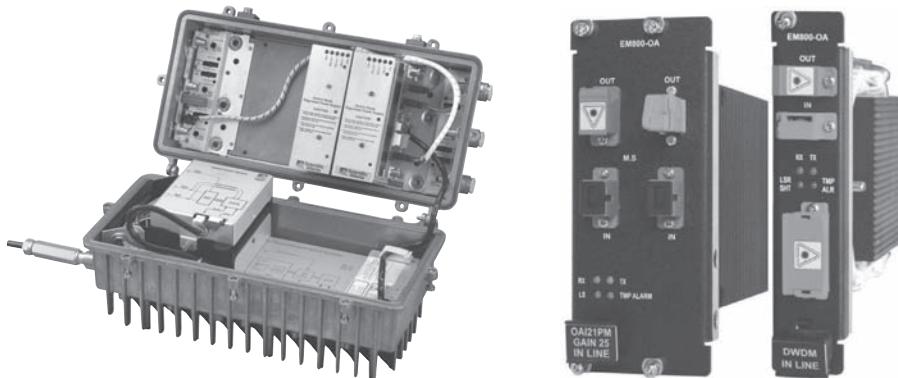


FIGURE 4.9 A simplified schematic illustration of an EDFA (optical amplifier). The Er^{3+} -doped fiber is pumped by feeding the light from a laser pump diode, through a coupler, into the erbium ion-doped fiber. Forward or codirectional pump configuration.



EDFA (Strand Mounted Optical Amplifier, Prisma 1550) for optical amplification at 1550 nm. This model can be used underground to extend the reach of networks, and operates over -40°C to $+65^\circ\text{C}$. The output can be as high as 24 dBm. (*Courtesy of Cisco.*)

EDFAs (LambdaDriver[®] Optical Amplifier Modules) with low noise figure and flat gain (to within $\pm 1 \text{ dB}$) for use in DWDM over 1528–1563 nm. These amplifiers can be used for booster, in-line, and preamplifier applications. (*Courtesy of MRV Communications, Inc.*)

The pumping configuration shown in Figure 4.9 feeds the pump light at A. A simplified fiber optic diagram of this configuration is shown in Figure 4.10 (a). Since the pump light and the signal travel in the same direction, this particular configuration is called **codirectional** or **forward pumping**. Alternatively, one can feed the pump light at B in Figure 4.9, as in Figure 4.10 (b), in which the pump light propagates in the opposite direction to the signal. This configuration is

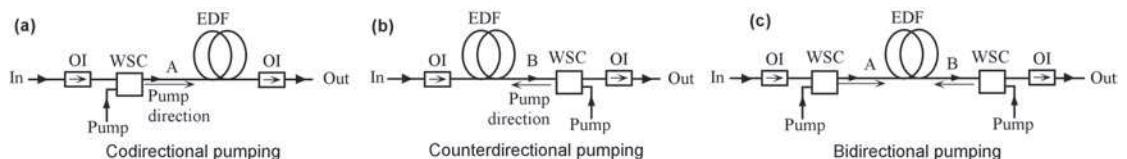


FIGURE 4.10 (a) Codirectional or forward pumping. (b) Counterdirectional or backward pumping. (c) Bidirectional pumping. OI, optical isolator; WSC, wavelength-selective coupler.

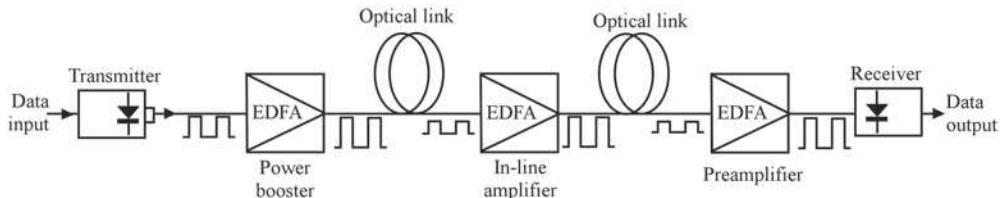


FIGURE 4.11 EDFA s are widely used in telecommunications as a power booster, after the transmitter, in-line amplifier within the optical link, and preamplifier before the receiver. The symbol shown for the EDFA is widely used.

called **counterdirectional** or **backward pumping**. In **bidiirectional pumping**, also called **dual pumping**, there are two pump diodes coupled to the EDF at A and B as shown in Figure 4.10 (c). In addition, there is usually a photodetector system coupled to monitor the pump power or the EDFA output power. These are not shown in Figure 4.9 nor in Figure 4.10. Typically EDFA s are used in three types of applications in optical communications. As a **power booster**, the EDFA is placed after the transmitter to boost the signal power fed into the optical network as shown in Figure 4.11. Quite often the signal from a transmitter needs to be boosted if the signal is to be split into N channels through an optical splitter. Booster EDFA s are designed so that the output signal does not easily saturate for large input signals. As an **in-line amplifier**, the EDFA is placed within the optical transmission link, as shown in Figure 4.11, where the signal has become weak and needs to be amplified. The in-line EDFA should have a large gain and add very little noise into the fiber. The noise in the EDFA comes from the amplification of spontaneously (and randomly) emitted photons in the direction of the fiber. As a **preamplifier**, it is placed before the receiver to amplify the signal fed into the detector as illustrated in Figure 4.11. The preamplifier EDFA is designed to provide good gain with low noise.

B. EDFA Characteristics, Efficiency, and Gain Saturation

The gain of an EDFA depends on the launched pump power from the pump diode. The gain coefficient g depends on $\sigma_{\text{em}}N_2 - \sigma_{\text{ab}}N_1$, and as more pump power is fed into the core of the EDFA, N_2 increases and N_1 decreases given that $N_1 + N_2 = N_0$, the Er^{3+} concentration in the fiber. Figure 4.12 (a) shows typical characteristics of EDFA gain, G , in dB vs. launched pump

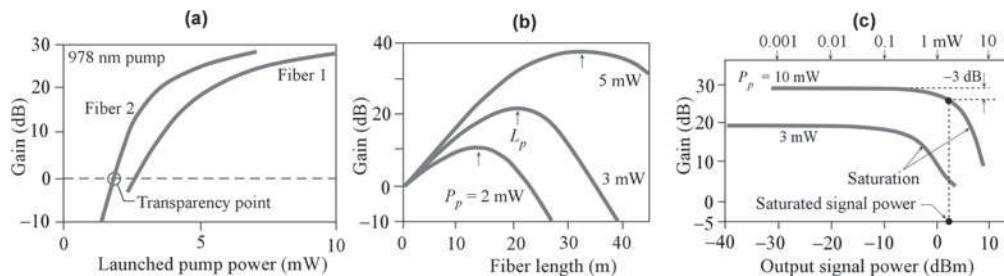


FIGURE 4.12 (a) Typical characteristics of EDFA small signal gain in dB vs. launched pump power for two different types of fibers pumped at 980 nm. The fibers have different core compositions and core diameter, and different lengths ($L_1 = 19.9$ m and $L_2 = 13.6$ m). (b) Typical dependence of small signal gain G on the fiber length L at different launched pump powers. There is an optimum fiber length L_p . (c) Typical dependence of gain on the output signal strength for different launched pump powers. At high output powers, the output signal saturates, i.e., the gain drops. (Source: Figures were constructed by using typical data from C. R. Jiles *et al.*, *IEEE Photon. Technol. Lett.*, 3, 363, 1991; and C. R. Jiles *et al.*, *J. Light Wave Technol.*, 9, 271, 1991.)

power P_p for two different types of fibers. As expected, G first increases sharply with P_p . At high pump powers, further increases in N_2 and hence G become constrained due to the diminishing population N_1 with increasing pump power. The gain G shown in Figure 4.12 (a) is for **small signals**, that is, for signal optical powers that do not themselves significantly depopulate N_2 by encouraging stimulated emissions, and hence decrease G .

Since the gain G is given by $\exp(gL)$, G increases with the fiber length L as shown in Figure 4.12 (b) for a given fiber at different pump power levels. However, we cannot indefinitely increase the gain by using a longer fiber. As shown in Figure 4.12 (b), initially G increases with L but then drops with L for long fiber lengths. Put differently, there is an **optimum fiber length or pump length** that maximizes the gain at a given launched pump power. Beyond this optimum length, the pump power in the core decreases with length so much that it is unable to maintain the required gain coefficient g . Moreover, further along the fiber, the pump becomes too weak, the gain disappears and the fiber attenuates the signal, which is highly undesirable. Thus, it is important to design the EDFA with a length that is not longer than the optimum length. The optimum fiber length is longer for higher launched pump powers as can be seen from Figure 4.12 (b).

One very important characteristic of an EDFA is the **saturation** of its output signal power, that is, the fall in the gain, under large signals. As the signal is amplified, the optical power at 1550 nm in the fiber core increases, which itself encourages further stimulated emission and thereby depopulates the population N_2 ; and hence the gain drops. The gain shown in Figures 4.12 (a) and (b) is for small signals, that is, the signal power is much less than the pump power in the core. Normally manufacturers quote the output power that saturates the amplifier, which is the output signal at which the gain drops by 3 dB below its constant small signal value as shown in Figure 4.12 (c). It is simply called the maximum output power and is quoted in dBm. In commercial high power booster EDFA, the saturated output power is typically around 22–25 dBm (160–300 mW) but at the expense of gain; the corresponding small signal gain is around 10–15 dB (or 10–30). In contrast, small in-line EDFA typically have a saturation power of 15–20 dBm (30–100 mW) and a small signal gain of 20–30 dB or higher.

The **gain efficiency** of an EDFA is the maximum optical gain achievable per unit optical pumping power and is quoted in dB mW^{-1} . Typical gain efficiencies are around 8–10 dB mW^{-1} at 980 nm pumping. A 30 dB or 10^3 gain is easily attainable with a few tens of milliwatts of pumping at 980 nm, depending on the fiber core doping and the coupling of the pump into the fiber.

The power conversion efficiency (PCE), η_{PCE} , of an EDFA represents the efficiency with which the amplifier is able to convert power from the pump to the signal. Suppose that P_{pin} is the pump power into the EDFA as shown in Figure 4.13 (a), where P_{sin} and P_{sout} are the input and output powers, respectively. Then

$$\eta_{\text{PCE}} = \frac{P_{\text{sout}} - P_{\text{sin}}}{P_{\text{pin}}} \approx \frac{P_{\text{sout}}}{P_{\text{pin}}} \quad (4.3.1)$$

Power conversion efficiency

Suppose that the flux of photons from the pump is Φ_{pin} and the signal input and output photon fluxes are Φ_{sin} and Φ_{sout} . Since optical power is proportional to photon flux $\times hc/\lambda$ where hc/λ is the photon energy, we can write Eq. (4.3.1) as

$$\eta_{\text{PCE}} \approx \frac{\Phi_{\text{sout}}}{\Phi_{\text{pin}}} \times \frac{\lambda_p}{\lambda_s} \quad (4.3.2)$$

Power conversion efficiency

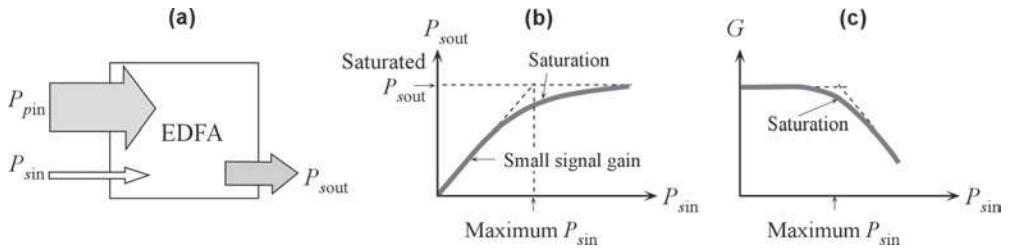


FIGURE 4.13 (a) The EDFA is pumped by P_{pin} . The power from the pump is converted into the output signal power P_{sout} . (b) The P_{sout} vs. P_{sin} behavior is initially linear but saturates at high input signal powers. (c) The output becomes saturated. The gain G drops in the saturation region.

where λ_p and λ_s are the pump and signal wavelengths. Equation (4.3.2) shows that η_{PCE} has a maximum value of λ_p/λ_s , which corresponds to each pump photon yielding one additional signal output photon. Thus, maximum PCE is 63% for 980 nm pumping.

The gain can be defined and written as

$$\boxed{\text{EDFA gain} \quad G = \frac{P_{\text{sout}}}{P_{\text{sin}}} = 1 + \eta_{\text{PCE}} \left(\frac{P_{\text{pin}}}{P_{\text{sin}}} \right)} \quad (4.3.3)$$

in which we used Eq. (4.3.1) to substitute for P_{sout} . The maximum gain G_{max} is for $\eta_{\text{PCE}} = \lambda_p/\lambda_s$ so that $G < G_{\text{max}}$ implies

$$G < 1 + (\lambda_p/\lambda_s)(P_{\text{pin}}/P_{\text{sin}})$$

or, rearranging, the input signal is

EDFA gain and input signal power

$$P_{\text{sin}} < (\lambda_p/\lambda_s)P_{\text{pin}}/(G - 1) \quad (4.3.4)$$

The importance of this equation is that it specifies the range of input signals, that is, the maximum input signal power, beyond which the output signal power begins to saturate as indicated in Figure 4.13 (b). Consequently, the gain drops with the input signal power as in Figure 4.13 (c). This phenomenon is called **gain saturation or compression**. For example, for an EDFA that has a small signal gain of 30 dB (or 10^3), $P_{\text{pin}} = 100 \text{ mW}$, 980 nm pumping, we can use Eq. (4.3.4) to find P_{sin} must be less than $60 \mu\text{W}$, or the output must be less than 60 mW . Thus, for outputs beyond 60 mW , the EDFA will saturate.

For a given pump level P_p , there is a maximum fiber length L_p beyond which the gain is lost rapidly with length, and the fiber attenuates the signal. To find the length of fiber L_p required to absorb the pump radiation, we note that under strong pumping, N_1 is very small. There are very few ions in the E_1 -manifold. The absorbed pump power is primarily taken by those excited ions that have spontaneously decayed from E_2 to E_1 , because these have to be pumped back up to E_3 and hence E_2 . The characteristic time of this natural decay from the E_2 to E_1 is the **spontaneous emission lifetime** τ_{sp} ($= 1/A_{21}$). Thus,

$$\begin{aligned} \text{Absorbed pump power} &= \text{Absorbed energy per unit time} \\ &\approx (\text{Volume})(N_2)(\text{Pump photon energy})/(\text{Decay time}) \\ &= (A L_p)(N_2)(h\nu_p)/\tau_{\text{sp}} \end{aligned}$$

where A is the cross-sectional area of the fiber core that has the Er^{3+} dopants, and L_p is the length of fiber needed to absorb the necessary pump radiation, v_p is the pump frequency. Thus, the absorbed pump power P_p is

$$P_p \approx AN_0hv_pL_p/\tau_{\text{sp}} \quad (4.3.5)$$

where $N_2 \approx N_0$, the Er^{3+} concentration in the core. Equation (4.3.5) essentially determines (very approximately) the maximum fiber length allowed for a given amount of pumping. The actual fiber length L needs to be shorter than L_p for the following reason. If L is longer than L_p , then beyond L_p , the gain will be lost and the signal will become absorbed since this region will have unpumped Er^{3+} ions. We assumed a small signal in our derivation that did not significantly upset the population of the upper level. If the pumping is not strong and/or the signal is not small, then N_1 cannot be neglected. Hence, the actual L is usually less than L_p in Eq. (4.3.5). Further, not all the pump radiation will be confined to the fiber core, whereas we assumed it was. Only some fraction Γ , called the **confinement factor**, of the pump radiation will be guided within the core and hence pump the Er^{3+} ions. Γ is typically 0.6–0.8. Thus, we need to use ΓP_p instead of P_p in Eq. (4.3.5), *i.e.*,

$$\Gamma P_p \approx AN_0hv_pL_p/\tau_{\text{sp}} \quad (4.3.6)$$

Given the fiber length $L (< L_p)$, the **small signal gain** G is easily found from

$$G = \exp(gL) \quad (4.3.7)$$

which is maximum for a fiber length $L = L_p$.

Strong absorption and pump length of EDFA

Strong absorption and pump length of EDFA

Small signal gain

EXAMPLE 4.3.1 An erbium-doped fiber amplifier

Consider a 3-m EDFA that has a core diameter of 5 μm , Er^{3+} doping concentration of $1 \times 10^{19} \text{ cm}^{-3}$, and τ_{sp} (the spontaneous decay time from E_2 to E_1) is 10 ms. The fiber is pumped at 980 nm from a laser diode. The pump power coupled into the EDFA fiber is 25 mW. Assuming that the confinement factor Γ is 70%, what is the fiber length that will absorb the pump radiation? Find the small signal gain at 1550 nm for two cases corresponding to full population inversion and 90% inversion.

Solution

The pump photon energy $h\nu = hc/\lambda = (6.626 \times 10^{-34})(3 \times 10^8)/(980 \times 10^{-9}) = 2.03 \times 10^{-19} \text{ J}$ (or 1.27 eV).

Rearranging Eq. (4.3.6), we get

$$L_p \approx \Gamma P_p \tau_{\text{sp}} / A N h \nu_p$$

i.e.,

$$L_p \approx (0.70)(25 \times 10^{-3} \text{ W})(10 \times 10^{-3} \text{ s}) / [\pi(2.5 \times 10^{-4} \text{ cm})^2(1 \times 10^{19} \text{ cm}^{-3})(2.03 \times 10^{-19} \text{ J})] = 4.4 \text{ m}$$

which is the maximum allowed length. The small signal gain can be rewritten as

$$g = \sigma_{\text{em}}N_2 - \sigma_{\text{ab}}N_1 = [\sigma_{\text{em}}(N_2/N_0) - \sigma_{\text{ab}}(N_1/N_0)]N_0$$

where $N_1 + N_2 = N_0$ is the total Er^{3+} concentration. Let $x = N_2/N_0$, then $1 - x = N_1/N_0$ where x represents the extent of pumping from 0 to 1, 1 being 100%. Thus, the above equation becomes

$$g = [\sigma_{\text{em}}x - \sigma_{\text{ab}}(1 - x)]N_0$$

For 100% pumping, $x = 1$,

$$g = [(3.2 \times 10^{-21} \text{ cm}^2)(1) - 0](1 \times 10^{19} \text{ cm}^{-3}) = 3.2 \text{ m}^{-1}$$

and

$$G = \exp(gL) = \exp[(3.2 \text{ m}^{-1})(3 \text{ m})] = 14,765 \quad \text{or} \quad 41.7 \text{ dB}$$

For $x = 0.9$ (90% pumping), we have

$$g = [(3.2 \times 10^{-21} \text{ cm}^2)(0.9) - (2.4 \times 10^{-21} \text{ cm}^2)(0.1)](1 \times 10^{19} \text{ cm}^{-3}) = 2.64 \text{ m}^{-1}$$

and

$$G = \exp(gL) = \exp[(2.64 \text{ m}^{-1})(3 \text{ m})] = 2751 \quad \text{or} \quad 34.4 \text{ dB}$$

Even at 90% pumping the gain is significantly reduced. At 70% pumping, the gain is 19.8 dB. In actual operation, it is unlikely that 100% population inversion can be achieved; 41.7 dB is a good indicator of the upper ceiling to the gain.

C. Gain-Flattened EDFA and Noise Figure

The gain spectrum of an EDFA refers to the dependence of the small signal gain G on the wavelength within the communications band of interest. Figure 4.8 (b) shows how the gain varies as a function of wavelength for an EDFA that has not had its gain flattened. In wavelength division multiplexing, it is essential to have the gain as flat as possible over all the channels, *i.e.*, the signal wavelengths, that are used. Many gain-flattened commercial EDFA for DWDM have their gain variations within ± 0.5 dB. Figure 4.14 (a) shows typical gain spectra for a commercial EDFA for use in DWDM in the C-band. For small input signals (which do not cause any saturation) the gain is reasonably flat. The gain decreases at high input signals and, in addition, the gain becomes “less” flat.

The simplest way to flatten the gain spectrum is to use a filter (or a series of filters) after the EDFA that has an attenuation or a transmission spectrum such that it attenuates

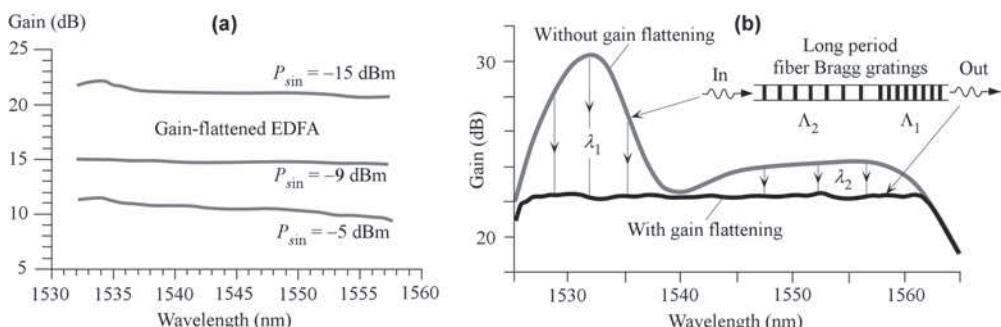


FIGURE 4.14 (a) The gain spectrum of one type of commercial gain-flattened EDFA. The gain variation is very small over the spectrum, but the gain decreases as the input signal power increases due to saturation effects. (*Note:* The corresponding input signal power levels are 0.031, 0.13, and 0.32 mW.) (b) Schematic illustration of gain equalization by using long period fiber Bragg grating filters in series that attenuate the high gain regions to equalize the gain over the spectrum. (An idealized example.)

the higher gain region of the spectrum and not the lower gain region. The high gain regions are attenuated by the right amount to bring these regions down to the same level as the low gain region, as shown for an idealized case in Figure 4.14 (b). The filter **equalizes** the gain. The combination of this filter with the EDFA will then result in a gain spectrum that has been **equalized**. The effect is to attenuate the high gain region more than the low gain region, and hence achieve gain equalization. Typically **long period gratings** (LPGs) are used for the filter part. If the period Λ of a fiber Bragg grating (FBG) is much longer (*e.g.*, by $100\times$) than the wavelength, then the diffraction occurs into the cladding for those wavelengths that satisfy the Bragg diffraction condition. Thus, by choosing the period Λ correctly we can diffract some of the signal in the high gain region into the cladding and hence suppress the signal in this gain region, *i.e.*, equalize the gain. Normally at least two or more properly designed LPGs are needed in series to usefully equalize the gain, as should be apparent from Figure 4.14 (b) where one would bring the gain in the λ_1 -region down and the other lowers the gain in the λ_2 -region. LPGs have a periodicity (Λ) that is of the order of $100\text{ }\mu\text{m}$ and hence are easier to manufacture than regular FBGs. In addition, the diffracted wave is not simply reflected but is fed into the cladding where it is attenuated and lost; hence the back reflection is low. Commercially available LPGs can equalize the gain down to about $\pm 0.5\text{ dB}$. Good gain equalization requires the filter transmission spectrum to have a shape that is the **inverse** of the shape of the gain spectrum of the EDFA. Some manufacturers also provide the transmission spectrum of the filter in their data sheets.

It is also possible to obtain gain equalization by using a regular FBG in which the periodicity Λ is not fixed, but adjusted along the fiber length, *e.g.*, becomes shorter along the fiber, so that the transmission spectrum can be tailored to the need, and can be made to be an almost exact inverse of the EDFA's gain spectrum to yield a flat net gain spectrum. Such FBGs in which the period changes along the fiber are called **chirped FBGs**. Most commercial chirped FBGs can provide a gain flatness down to $\pm 0.25\text{ dB}$, which is better than $\pm 0.5\text{ dB}$ for LPGs; research shows that one can, in fact, do better, down to $\pm 0.1\text{ dB}$. Chirped FBGs suffer from high back reflections (attenuated signals are actually back reflected) but they are more stable to temperature variations compared to LPGs.

Most commercial gain-flattened EDFAs have more than just a series of filters placed after the EDFA. Usually a multistage EDFA design is used. A simple broadband gain-flattened two-stage EDFA configuration is shown in Figure 4.15 (a) as reported by Lucent researchers.¹⁶ A 980-nm laser diode is used to forward pump the EDFA1, whose output is passed through a well-designed long period grating for gain flattening. There is a second stage, EDFA2, that is forward pumped from a 1480-nm laser diode. The small signal, that is unsaturated, gain spectrum from such a two-stage gain-flattened EDFA is shown in Figure 4.15 (b), where it can be seen that over a broadband (40 nm), the gain variation is less than 1 dB; or put differently, less than $\pm 0.5\text{ dB}$ about the mean. By using more stages and filters, excellent gain flattening can be easily obtained. Table 4.1 summarizes the properties of a particular commercial EDFA that has been gain flattened and designed for DWDM applications.

Like all amplifiers, EDFA also generate noise in addition to the amplification they provide. The noise in an EDFA arises from **amplified spontaneous emission** (ASE), which is the

¹⁶See P. F. Wysocki *et al.*, *IEEE Photon. Technol. Lett.*, 9, 1343, 1997; Y. Sun *et al.*, *Bell Labs Tech. J.*, 4, 187, 1999; and R. P. Espindola and P. F. Wysocki, U.S. Patent 5,920,424, June 6, 1999.

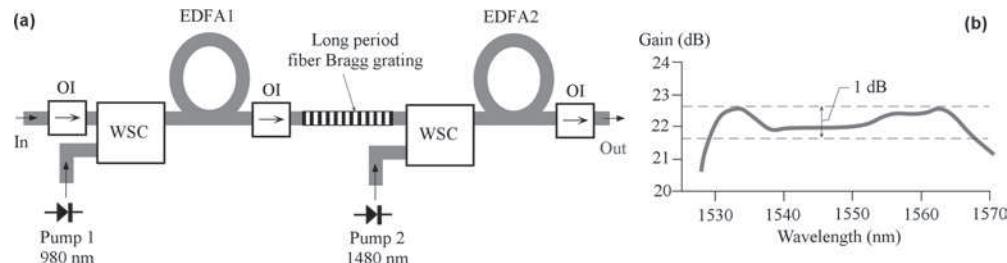


FIGURE 4.15 (a) A gain-flattened EDFA reported by Lucent Technologies¹⁷ uses two EDFA stages and a long period grating between the two stages. (A simplified diagram.) The two EDFA stages are pumped at 980 nm and 1480 nm. (b) The resulting gain spectrum for small signals is flat to better than 1 dB over a broad spectrum, 40 nm. The length of EDFA1 is 14 m, and that of EDFA2 is 15 m. Pump 1 (980 nm) and Pump 2 (1480 nm) diodes were operated at most at power levels 76 mW and 74.5 mW, respectively. EDFA2 can also be pumped counterdirectionally.

amplification of randomly emitted photons from the upper E_2 -manifold to the lower E_1 -manifold. This spontaneous emission generates random photons, and as these photons travel along the fiber, they become amplified. Thus, the output from an EDFA not only has the signal but also noise from the ASE as shown in Figure 4.16 (a).

The ratio of the signal power (P_s) to the noise power (P_n) present at the input or at the output of an amplifier is called the **signal-to-noise ratio** (SNR). At the input and the output, the SNR would be designated by SNR_{in} and SNR_{out} , respectively. An ideal amplifier should not add any noise while amplifying the signal so that the SNR_{out} should be the same as SNR_{in} because both the signal and noise at the input would be amplified by the same amount, gain G (and no additional noise would be added). Noise characteristics of amplifiers are generally quantified by their **noise figure** (NF), defined by

Noise
figure
definition

$$\text{NF} = \frac{\text{SNR}_{\text{in}}}{\text{SNR}_{\text{out}}} \quad \text{or} \quad \text{NF(dB)} = 10 \log \left(\frac{\text{SNR}_{\text{in}}}{\text{SNR}_{\text{out}}} \right) \quad (4.3.8)$$

where the second term gives NF in the units of dB, the most commonly used units for NF. For a noiseless amplifier, $\text{NF} = 0 \text{ dB}$ (or 1). Table 4.1 shows a value of 5 dB for the NF for the particular EDFA listed in the table. This value is quite typical for EDFA in the unsaturated region, but increases in the saturation region, *i.e.*, high input (or output) power levels.

TABLE 4.1 Specifications for a typical gain-flattened EDFA for DWDM (dense wavelength division multiplexing) applications

Operating wavelength (nm)	Output power (dBm)	Number of pump lasers (nm)	Input signal for flat gain (dBm)	Channels	Max gain (dB channel $^{-1}$)	Flatness (dB)	NF (dB)
1529–1561	18–24	2 × 980 2 × 1480	−15 to −7	64	21	±0.5	5

Note: Optilab EDFA-GI series EDFA.

¹⁷ See footnote number 15.

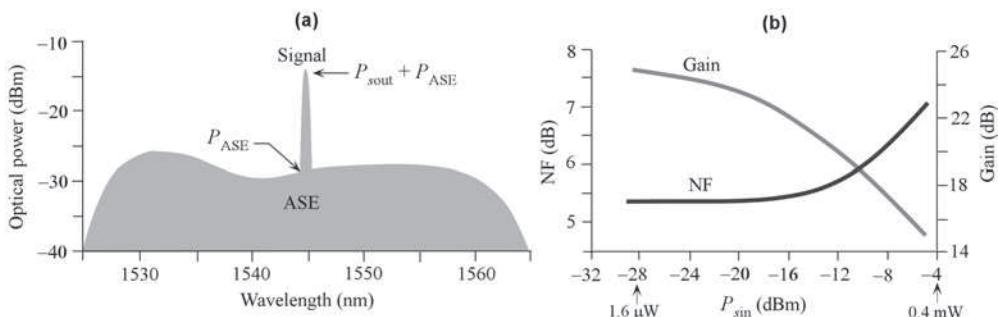


FIGURE 4.16 (a) Amplified spontaneous emission (ASE) noise in the output spectrum and the amplified signal. (b) The dependence of NF and gain (G) on the input signal power level (P_{sin}) for an EDFA under forward (codirectional) pumping. (Source: Data for the plots were selectively extracted from G. R. Walker *et al.*, *J. Light Wave Technol.*, 9, 182, 1991.)

As shown in Figure 4.16 (b), the NF is initially independent of the input signal power P_{sin} level, but at sufficiently high P_{sin} , it increases with P_{sin} ; or put differently, NF increases with falling gain.

4.4 GAS LASERS: THE He-Ne LASER

With the He-Ne laser one has to confess that the actual explanation is by no means simple since we have to know such things as the energy states of the whole atom. We will consider only the lasing emission at 632.8 nm which gives the well-known red color to the He-Ne laser light. The actual stimulated emission occurs from the Ne atoms. He atoms are used to excite the Ne atoms by atomic collisions. Overall, the He-Ne gas laser is like a four-level laser system (Figure 4.4).

Ne is an inert gas with a ground state ($1s^2 2s^2 2p^6$) which will be represented as ($2p^6$) by ignoring the inner closed $1s$ and $2s$ subshells. If one of the electrons from the $2p$ orbital is excited to a $5s$ -orbital, then the excited configuration ($2p^5 5s^1$) is a state of the Ne atom that has higher energy. Similarly, He is also an inert gas which has the ground state configuration of ($1s^2$). The state of He when one electron is excited to a $2s$ -orbital can be represented as ($1s^1 2s^1$) and has higher energy than the ground state.

The He-Ne laser consists of a gaseous mixture of He and Ne atoms in a gas discharge tube, as sketched schematically in Figure 4.17. The ends of the tube are mirrored to reflect the stimulated radiation and build up intensity within the cavity. In other words, an *optical cavity* is

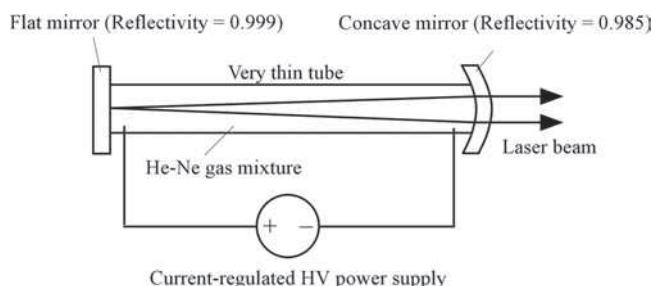
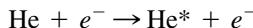


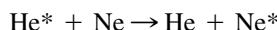
FIGURE 4.17 A schematic illustration of the principle of the He-Ne laser.

formed by the end mirrors so that reflection of photons back into the lasing medium builds up the photon concentration in the cavity, a requirement of an efficient stimulated emission process as discussed above. By using DC or RF high voltage, an electrical discharge is obtained within the tube which causes the He atoms to become excited by collisions with those energetic electrons that have been accelerated by the field. Thus,



where He^* is an excited He atom.

The excitation of the He atom by an electron collision puts the second electron in He into a $2s$ state, so that the excited He atom, He^* , has the configuration $(1s^12s^1)$, which is *metastable* (that is, a long-lasting state) with respect to the $(1s^2)$ state as shown schematically in Figure 4.18. He^* cannot spontaneously emit a photon and decay down to the $(1s^2)$ ground state because the *orbital quantum number* ℓ of the electron must change by ± 1 , *i.e.*, $\Delta\ell$ must be ± 1 for any photon emission or absorption process. Thus a large number of He^* atoms build up during the electrical discharge because they are not allowed to simply decay back to ground state. When an excited He atom collides with a Ne atom, it transfers its energy to the Ne atom by resonance energy exchange because, by good fortune, Ne happens to have an empty energy level, corresponding to the $(2p^55s^1)$ configuration, matching that of $(1s^12s^1)$ of He^* . Thus the collision process excites the Ne atom and de-excites He^* down to its ground energy, *i.e.*,



With many He^* -Ne collisions in the gaseous discharge, we end up with a large number of Ne^* atoms and a population inversion between $(2p^55s^1)$ and $(2p^53p^1)$ states of the Ne atom as indicated in Figure 4.18. A spontaneous emission of a photon from one Ne^* atom falling from $5s$ to $3p$ gives rise to an avalanche of stimulated emission processes which leads to a lasing emission with a wavelength of 632.8 nm in the red.

There are a few interesting facts about the He-Ne laser, some of which are quite subtle. First, the $(2p^55s^1)$ and $(2p^53p^1)$ electronic configurations of the Ne atom actually have a spread of energies. For example, for $\text{Ne}(2p^55s^1)$, there are four closely spaced energy levels. Similarly for $\text{Ne}(2p^53p^1)$ there are ten closely separated energies. We see that we can achieve population

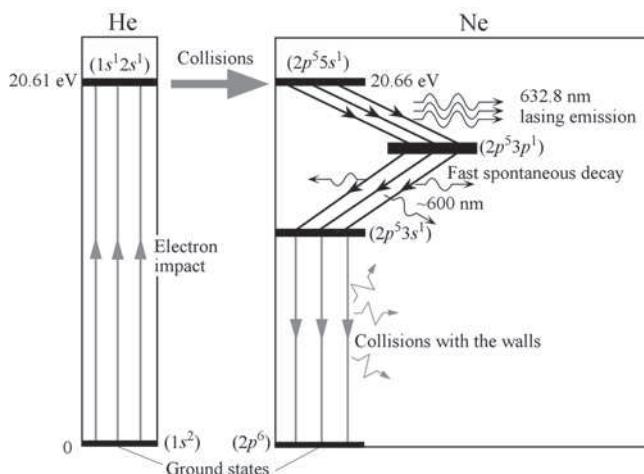


FIGURE 4.18 The principle of operation of the He-Ne laser and the He-Ne laser energy levels involved for 632.8 nm emission.

inversion with respect to a number of energy levels and, as a result, the lasing emissions from the He-Ne laser contain a variety of wavelengths. The two lasing emissions in the visible spectrum at 632.8 nm and 543 nm can be used to build a red or a green He-Ne laser. Further, we should note that the energy of the state $\text{Ne}(2p^54p^1)$ (not shown) is above $\text{Ne}(2p^53p^1)$ but below $\text{Ne}(2p^55s^1)$. There will therefore also be stimulated transitions from $\text{Ne}(2p^55s^1)$ to $\text{Ne}(2p^54p^1)$ and hence a lasing emission at a wavelength of $\sim 3.39 \mu\text{m}$ (infrared). To suppress lasing emissions at the unwanted wavelengths and to obtain lasing only at the wavelength of interest, the reflecting mirrors can be made wavelength selective. This way, the optical cavity builds up optical oscillations at the selected wavelength.

From the $(2p^53p^1)$ energy levels, the Ne atoms decay rapidly to the $(2p^53s^1)$ energy levels by spontaneous emission. Most of Ne atoms with the $(2p^53s^1)$ configuration, however, cannot simply return to the ground state $2p^6$ by photon emission because the return of the electron in $3s$ requires that its spin is flipped to close the $2p$ -subshell. An electromagnetic radiation cannot change the electron spin. Thus, the $\text{Ne}(2p^53s^1)$ energy levels are *metastable* states. The only possible return to the ground state (and become pumped again) is by collisions with the walls of the laser tube. We cannot therefore increase the power obtainable from a He-Ne laser by simply increasing the laser tube diameter because that will accumulate more Ne atoms at the metastable $(2p^53s^1)$ states.

A typical He-Ne laser, as illustrated in Figure 4.17, consists of a narrow glass tube which contains the He and Ne gas mixture—typically He to Ne ratio of 10 to 1 and a pressure of ~ 1 torr. (The optimum pressure depends on the tube diameter.) The lasing emission intensity increases with the tube length since there are more Ne atoms used in stimulated emissions. The intensity decreases with increasing tube diameter since Ne atoms in the $(2p^53s^1)$ states can only return to the ground state by collisions with the walls of the tube. The ends of the tube are generally sealed with a flat mirror (99.9% reflecting) at one end and, for easy alignment, a concave mirror (99% reflecting) at the other end to obtain an **optical cavity** within the tube. The outer surface of the concave mirror is shaped like a convergent lens to compensate for the divergence in the beam arising from reflections from the concave mirror. The output radiation from the tube is typically a beam of diameter 0.5–1 mm and a divergence of a few milliradians at a power of few milliwatts. In high power He-Ne lasers, the mirrors are external to the tube. In addition, *Brewster windows* are typically used at the ends of the laser tube to allow only polarized light to be transmitted and amplified within the cavity so that the output radiation is polarized (has electric field oscillations in one plane).

Even though we can try to get as parallel a beam as possible by lining up the mirrors perfectly, we will still be faced with diffraction effects at the output. When the output laser beam hits the end of the laser tube it becomes diffracted so that the emerging beam is necessarily divergent. Simple diffraction theory can readily predict the divergence angle. Further, typically one or both of the reflecting mirrors in many gas lasers are made concave for a more efficient containment of the lasing radiation within the **active medium** and for easier alignment. The beam within the cavity and hence the emerging radiation is approximately a *Gaussian beam*. As mentioned in Chapter 1, a Gaussian beam diverges as it propagates in free space. Optical cavity engineering is an important part of laser design and there are various advanced texts on the subject.

Due to their relatively simple construction, He-Ne lasers are widely used in numerous applications such as interferometry, accurately measuring distances or flatness of an object, laser printing, holography, and various pointing and alignment applications (such as in civil engineering).



Ali Javan and his associates, William Bennett Jr. and Donald Herriott at Bell Labs, were the first to successfully demonstrate a continuous wave (CW) helium-neon laser operation (1960–1962). (*Reprinted with permission of Alcatel-Lucent USA Inc.*)



A modern He-Ne laser with its power supply. This unit provides a linearly polarized TE_{00} output at 633 nm (red) at a power of 10 mW. The beam diameter is 0.68 mm and the divergence is 1.2 mrd. The longitudinal mode separation is 320 MHz. (*Courtesy of Thorlabs.*)

EXAMPLE 4.4.1 Efficiency of the He-Ne laser

A typical low-power, 5-mW He-Ne laser tube operates at a DC voltage of 2000 V and carries a current of 7 mA. What is the efficiency of the laser?

Solution

From the definition of efficiency,

$$\begin{aligned}\text{Efficiency} &= \frac{\text{Output Light Power}}{\text{Input Electrical Power}} = \frac{5 \times 10^{-3} \text{ W}}{(7 \times 10^{-3} \text{ A})(2000 \text{ V})} \\ &= 0.036\%\end{aligned}$$

Typically He-Ne efficiencies are less than 0.1%. What is important is the highly coherent output radiation. Note that 5 mW over a beam diameter of 1 mm is 6.4 kW m^{-2} .

4.5 THE OUTPUT SPECTRUM OF A GAS LASER

The output radiation from a gas laser is not actually at one single well-defined wavelength corresponding to the lasing transition, but covers a spectrum of wavelengths with a central peak. This is not a simple consequence of the Heisenberg uncertainty principle, but a direct result of the broadening of the emitted spectrum by the **Doppler effect**. We recall from the kinetic molecular theory that gas atoms are in random motion with an average kinetic energy of $(3/2)k_B T$. Suppose that these gas atoms emit radiation of frequency ν_o , which we label as the source frequency.

Then, due to the Doppler effect, when a gas atom is moving *away* from an observer, the latter detects a lower frequency ν_1 given by

$$\nu_1 = \nu_o \left(1 - \frac{v_x}{c} \right) \quad (4.5.1)$$

where v_x is the relative velocity of the atom along the laser tube (x -axis) with respect to the observer and c is the speed of light. When the atom is moving *toward* the observer, the detected frequency ν_2 is higher and corresponds to

$$\nu_2 = \nu_o \left(1 + \frac{v_x}{c} \right) \quad (4.5.2)$$

Since the atoms are in random motion the observer will detect a range of frequencies due to this Doppler effect. As a result, the frequency or wavelength of the output radiation from a gas laser will have a “linewidth” $\Delta\nu = \nu_2 - \nu_1$. This is what we mean by a **Doppler broadened linewidth** of a laser radiation. There are other mechanisms that also broaden the output spectrum but we will ignore these in the present case of gas lasers.

From the kinetic molecular theory we know that the velocities of gas atoms obey the Maxwell distribution. Consequently, the stimulated emission wavelengths in the lasing medium must exhibit a distribution about a central wavelength $\lambda_o = c/v_o$. Stated differently, the lasing medium therefore has an **optical gain** (or a photon gain) that has a distribution around $\lambda_o = c/v_o$ as shown in Figure 4.19 (a). The variation in the optical gain with the wavelength is called the **optical gain lineshape**. For the Doppler broadening case, this lineshape turns out to be a Gaussian function. For many gas lasers this spread in the frequencies from ν_1 to ν_2 is 1–5 GHz (for the He-Ne laser the corresponding wavelength spread is ~ 0.002 nm).

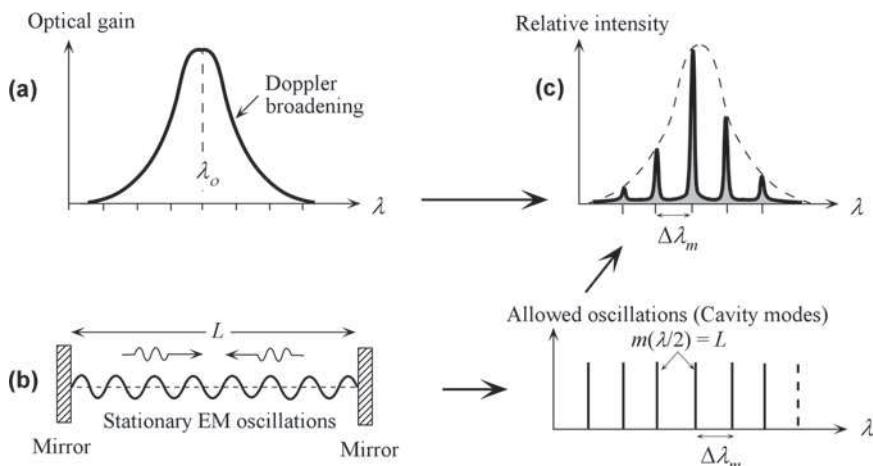


FIGURE 4.19 (a) Optical gain vs. wavelength characteristics (called the optical gain curve) of the lasing medium. (b) Allowed modes and their wavelengths due to stationary EM waves within the optical cavity. (c) The output spectrum (relative intensity vs. wavelength) is determined by satisfying (a) and (b) simultaneously, assuming no cavity losses.

When we consider the Maxwell velocity distribution of the gas atoms in the laser tube, we find that the linewidth $\Delta\nu_{1/2}$ between the half-intensity points (full width at half maximum—FWHM) in the output intensity vs. frequency spectrum is given by

Frequency linewidth (FWHM)

$$\Delta\nu_{1/2} = 2\nu_o \sqrt{\frac{2k_B T \ln(2)}{Mc^2}} \quad (4.5.3)$$

where M is the mass of the lasing atom or molecule (Ne in the He-Ne laser). The FWHM width $\Delta\nu_{1/2}$ is about 18% different than simply taking the difference $\nu_2 - \nu_1$ from Eqs. (4.5.1) and (4.5.2) and using a root-mean-square effective velocity along x , that is, using ν_x in $\frac{1}{2} M\nu_x^2 = \frac{1}{2} k_B T$. Equation (4.5.3) can be taken to be the FWHM width $\Delta\nu_{1/2}$ of the optical gain curve of most gas lasers. It does not apply to solid state lasers in which other broadening mechanisms operate.

Suppose that, for simplicity, we consider an optical cavity of length L with parallel end mirrors as shown in Figure 4.19 (b). Such an optical cavity is called a **Fabry-Perot optical resonator**. The reflections from the end mirrors of a laser give rise to traveling waves in opposite directions within the cavity. These oppositely traveling waves interfere constructively to set up a standing wave, that is, stationary electromagnetic (EM) oscillations as in Figure 4.19 (b). Some of the energy in these oscillations is tapped out by the 99% reflecting mirror to get an output, the same way we tap out the energy from an oscillating field in an LC circuit by attaching an antenna to it. Only standing waves with certain wavelengths, however, can be maintained within the optical cavity just as only certain acoustic wavelengths can be obtained from musical instruments. Any standing wave in the cavity must have an integer number of half-wavelengths, $\lambda/2$, that fit into the cavity length L

Laser cavity modes in a gas laser

$$m\left(\frac{\lambda}{2}\right) = L \quad (4.5.4)$$

where m is an integer that is called the **longitudinal mode number** of the standing wave. The wavelength λ in Eq. (4.5.4) is that within the cavity medium, but for gas lasers the refractive index is nearly unity and λ is the same as the free-space wavelength. Only certain wavelengths will satisfy Eq. (4.5.4), and for each m there will be an EM wave with a certain wavelength λ_m that satisfies Eq. (4.5.4).

Each possible standing wave within the laser tube (cavity) satisfying Eq. (4.5.4) is called a **cavity mode**. The cavity modes, as determined by Eq. (4.5.4), are shown in Figure 4.19 (b). Modes that exist along the cavity axis are called **axial** or **longitudinal modes**. Other types of modes, that is, stationary EM oscillations, are possible when the end mirrors are not flat. An example of an optical cavity formed by confocal spherical mirrors is shown in Figure 1.54 (Chapter 1). The EM radiation within such a cavity is a *Gaussian beam*.

The laser output thus has a broad spectrum with peaks at certain wavelengths corresponding to various cavity modes existing within the Doppler broadened optical gain curve as indicated in Figure 4.19 (c). At wavelengths satisfying Eq. (4.5.4), that is, representing certain cavity modes, we have spikes of intensity in the output. The net envelope of the output spectrum is a Gaussian distribution which is essentially due to the Doppler broadened linewidth. Notice that there is a finite width to the individual intensity spikes within the spectrum which is primarily due to nonidealities of the optical cavity, such as acoustic and thermal fluctuations of the cavity length L and nonideal end mirrors (less than 100% reflection). Typically, the frequency width of an individual spike in a He-Ne gas laser is ~ 1 MHz.

It is important to realize that even if the laser medium has an optical gain, the optical cavity will always have some losses inasmuch as some radiation will be transmitted through the mirrors, and there will be various losses such as scattering within the cavity. Only those modes that have an optical gain that can make up for the radiation losses from the cavity can exist in the output spectrum (as discussed later in Section 4.6).

EXAMPLE 4.5.1 Doppler broadened linewidth

Calculate the Doppler broadened linewidths $\Delta\nu$ and $\Delta\lambda$ (end-to-end of spectrum) for the He-Ne laser transition for $\lambda_o = 632.8\text{ nm}$ if the gas discharge temperature is about 127°C . The atomic mass of Ne is $20.2\text{ (g mol}^{-1}\text{)}$. The laser tube length is 40 cm. What is the linewidth in the output wavelength spectrum? What is mode number m of the central wavelength, the separation between two consecutive modes, and how many modes do you expect within the linewidth $\Delta\lambda_{1/2}$ of the optical gain curve?

Solution

Due to the Doppler effect arising from the random motions of the gas atoms, the laser radiation from gas lasers is broadened around a central frequency ν_o . The central ν_o corresponds to the source frequency. Higher frequencies detected will be due to radiations emitted from atoms moving toward the observer, whereas lower frequency will be result of the emissions from atoms moving away from the observer. We will first calculate the frequency width using two approaches, one approximate and the other more accurate. Suppose that v_x is the root-mean-square (rms) velocity along the x -direction. We can intuitively expect the frequency width $\Delta\nu_{\text{rms}}$ between rms points of the Gaussian output frequency spectrum to be¹⁸

$$\Delta\nu_{\text{rms}} = \nu_o \left(1 + \frac{v_x}{c} \right) - \nu_o \left(1 - \frac{v_x}{c} \right) = \frac{2\nu_o v_x}{c} \quad (4.5.5)$$

We need to know the rms velocity v_x along x which is given by the kinetic molecular theory as $\frac{1}{2}Mv_x^2 = \frac{1}{2}k_B T$, where M is the mass of the atom. We can therefore calculate v_x . For the He-Ne laser, it is the Ne atoms that lase, so $M = (20.2 \times 10^{-3}\text{ kg mol}^{-1})/(6.02 \times 10^{23}\text{ mol}^{-1}) = 3.35 \times 10^{-26}\text{ kg}$. Thus,

$$v_x = [(1.38 \times 10^{-23}\text{ J K}^{-1})(127 + 273\text{ K})/(3.35 \times 10^{-26}\text{ kg})]^{1/2} = 405.9\text{ m s}^{-1}$$

The central frequency is

$$\nu_o = c/\lambda_o = (3 \times 10^8\text{ m s}^{-1})/(632.8 \times 10^{-9}\text{ m}) = 4.74 \times 10^{14}\text{ s}^{-1}$$

The rms frequency linewidth is approximately

$$\begin{aligned} \Delta\nu_{\text{rms}} &\approx (2\nu_o v_x)/c \\ &= 2(4.74 \times 10^{14}\text{ s}^{-1})(405.9\text{ m s}^{-1})/(3 \times 10^8\text{ m s}^{-1}) = 1.28\text{ GHz} \end{aligned}$$

The FWHM width $\Delta\nu_{1/2}$ of the output frequency spectrum will be given by Eq. (4.4.3)

$$\begin{aligned} \Delta\nu_{1/2} &= 2\nu_o \sqrt{\frac{2k_B T \ln(2)}{Mc^2}} = 2(4.74 \times 10^{14}) \sqrt{\frac{2(1.38 \times 10^{-23})(400) \ln(2)}{(3.35 \times 10^{-26})(3 \times 10^8)^2}} \\ &= 1.51\text{ GHz} \end{aligned}$$

which is about 18% wider than the estimate from Eq. (4.5.5).

¹⁸The fact that this is the width between the rms points of the Gaussian output spectrum can be shown from detailed mathematics.

To get FWHM wavelength width $\Delta\lambda_{1/2}$, differentiate $\lambda = c/v$

$$\frac{d\lambda}{dv} = -\frac{c}{v^2} = -\frac{\lambda}{v} \quad (4.5.6)$$

so that

$$\Delta\lambda_{1/2} \approx \Delta\nu_{1/2}|\lambda/v| = (1.51 \times 10^9 \text{ Hz})(632.8 \times 10^{-9} \text{ m})/(4.74 \times 10^{14} \text{ s}^{-1})$$

or

$$\Delta\lambda_{1/2} \approx 2.02 \times 10^{-12} \text{ m} \quad \text{or} \quad 2.02 \text{ pm}$$

This width is between the half-points of the spectrum. The rms linewidth would be 0.0017 nm . Each mode in the cavity satisfies $m(\lambda/2) = L$ and since L is some 4.7×10^5 times greater than λ , the mode number m must be very large. For $\lambda = \lambda_o = 632.8 \text{ nm}$, the corresponding mode number m_o is

$$m_o = 2L/\lambda_o = (2 \times 0.4 \text{ m})/(632.8 \times 10^{-9} \text{ m}) = 1,264,222.5$$

and actual m_o has to be the closest integer value, that is, 1,264,222 or 1,264,223.

The separation $\Delta\lambda_m$ between two consecutive modes (m and $m + 1$) is

$$\Delta\lambda_m = \lambda_m - \lambda_{m+1} = \frac{2L}{m} - \frac{2L}{m+1} \approx \frac{2L}{m^2}$$

or

$$\Delta\lambda_m \approx \frac{\lambda_o^2}{2L} \quad (4.5.7)$$

Substituting the values, we find $\Delta\lambda_m = (632.8 \times 10^{-9})^2/(2 \times 0.4) = 5.01 \times 10^{-13} \text{ m}$ or 0.501 pm .

We can also find the separation of the modes by noting that Eq. (4.5.4), in which $\lambda = c/v$, is equivalent to

$$\nu = \frac{mc}{2L} \quad (4.5.8)$$

so that the separation of modes in frequency, $\Delta\nu_m$, is simply

$$\Delta\nu_m = \frac{c}{2L} \quad (4.5.9)$$

Substituting $L = 0.40 \text{ m}$ in Eq. (4.5.9), we find $\Delta\nu_m = 375 \text{ MHz}$. (A typical value for a He-Ne laser.)

The number of modes, *i.e.*, the number of m values, within the linewidth, that is, between the half-intensity points, will depend on how the cavity modes and the optical gain curve coincide, for example, whether there is a cavity mode right at the peak of the optical gain curve as illustrated in Figure 4.20. Suppose that we try to estimate the number of modes by using,

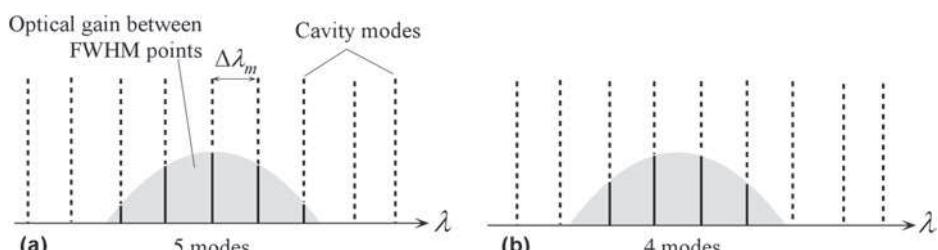


FIGURE 4.20 Number of laser modes depends on how the cavity modes intersect the optical gain curve. In this case we are looking at modes within the linewidth $\Delta\lambda_{1/2}$.

$$\text{Modes} = \frac{\text{Linewidth of spectrum}}{\text{Separation of two modes}} \approx \frac{\Delta\lambda_{1/2}}{\Delta\lambda_m} = \frac{2.02 \text{ pm}}{0.501 \text{ pm}} = 4.03$$

We should expect at most 4 to 5 modes within the linewidth of the output as shown in Figure 4.20. We neglected the cavity losses.

4.6 LASER OSCILLATIONS: THRESHOLD GAIN COEFFICIENT AND GAIN BANDWIDTH

A. Optical Gain Coefficient g

Consider a general laser medium that has an optical gain for coherent radiation along some direction x as shown in Figure 4.21 (a). This means that the medium is appropriately pumped. Consider an electromagnetic wave propagating in the medium along the x -direction. As it propagates, its power (energy flow per unit time) increases due to greater stimulated emissions over spontaneous emissions and absorption across the same two energy levels $E_2 - E_1$ as in Figure 4.21 (a). If the light intensity were decreasing, we would have used a factor $\exp(-\alpha x)$, where α is the attenuation coefficient, to represent the power loss along the distance x . Similarly, we represent the power increase along x by using a factor $\exp(gx)$, where g is a coefficient called the **optical gain coefficient** of the medium. The gain coefficient g is defined as the fractional increase in the light power (or intensity) per unit distance along the direction of propagation. We have already encountered g in Eq. (4.2.17). Optical power along x at any point is proportional to the concentration of coherent photons N_{ph} and the photon energy $h\nu$. These coherent photons travel with a velocity c/n , where n is the refractive index.¹⁹ Thus, in time δt they travel a distance $\delta x = (c/n)\delta t$ in the tube. The optical power P is proportional to N_{ph} so that the fractional increase in P per unit distance, which is the gain coefficient g , is given by

$$g = \frac{\delta P}{P\delta x} = \frac{\delta N_{\text{ph}}}{N_{\text{ph}}\delta x} = \frac{n}{cN_{\text{ph}}} \frac{\delta N_{\text{ph}}}{\delta t} \quad (4.6.1)$$

Optical
gain
coefficient

The gain coefficient g describes the increase in intensity of the lasing radiation in the cavity per unit length due to stimulated emission transitions from E_2 to E_1 exceeding photon

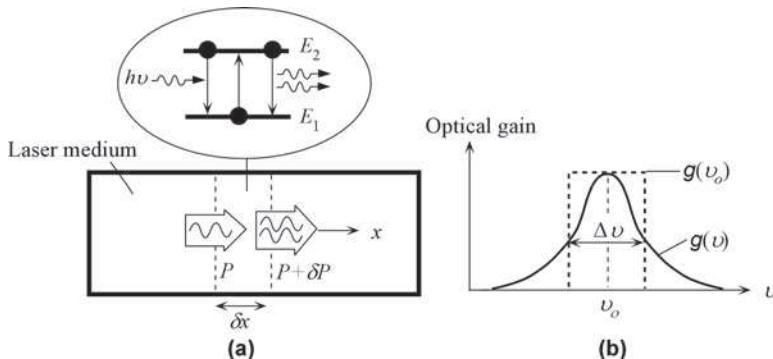


FIGURE 4.21 (a) A laser medium with an optical gain. (b) The optical gain curve of the medium. The dashed line is the approximate derivation in the text.

¹⁹ In semiconductor-related chapters, n is the electron concentration and n is the refractive index; E is the energy and E is the electric field.

absorption across the same two energy levels. We know that the difference between stimulated emission and absorption rates [see Eqs. (4.2.1) and (4.2.2)] gives the *net* rate of change in the coherent photon concentration, that is,

$$\begin{aligned}\frac{dN_{\text{ph}}}{dt} &= \text{Net rate of stimulated photon emission} \\ &= N_2 B_{21}\rho(v) - N_1 B_{12}\rho(v) = (N_2 - N_1)B_{21}\rho(v)\end{aligned}\quad (4.6.2)$$

It is now straightforward to obtain the optical gain by using Eq. (4.6.2) in Eq. (4.6.1) with certain assumptions. As we are interested in the amplification of coherent waves traveling along a defined direction (x) in Figure 4.21 (a), we can neglect spontaneous emissions which are in random directions and do not, on average, contribute to the directional wave. The reader should have noticed the similarity between Eq. (4.6.2) and Eq. (4.2.18).

Normally, the emission and absorption processes occur not at a discrete photon energy $h\nu$, but they would be distributed in photon energy or frequency over some frequency interval $\Delta\nu$. The spread $\Delta\nu$, for example, can be due to Doppler broadening or broadening of the energy levels E_2 and E_1 . There may be other broadening processes. In any event, this means that the optical gain will reflect this distribution, that is, $g = g(\nu)$ as illustrated in Figure 4.21 (b). The spectral shape of the gain curve is called the **lineshape function**.

We can express $\rho(\nu)$ in terms of N_{ph} by noting that $\rho(\nu)$ is defined as the radiation energy density per unit frequency so that at ν_o

$$\rho(\nu_o) \approx \frac{N_{\text{ph}}h\nu_o}{\Delta\nu} \quad (4.6.3)$$

We can now substitute for dN_{ph}/dt in Eq. (4.6.1) from Eq. (4.6.2) and use Eq. (4.6.3) to obtain the *optical gain coefficient*

General
optical
gain
coefficient

$$g(\nu_o) \approx (N_2 - N_1) \frac{B_{21}n h \nu_o}{c \Delta \nu} \quad (4.6.4)$$

Equation (4.6.4) gives the optical gain of the medium at the center frequency ν_o . A more rigorous derivation would have found the optical gain curve as a function of frequency, shown as $g(\nu)$ in Figure 4.21 (b), and would derive $g(\nu_o)$ from this lineshape.²⁰

B. Threshold Gain Coefficient g_{th} and Output Power

Consider an optical cavity with mirrors at the ends, such as the Fabry–Perot (FP) optical cavity shown in Figure 4.22. The cavity contains a laser medium so that lasing emissions build up to a steady state, that is, there is continuous operation. We effectively assume that we have stationary electromagnetic (EM) oscillations in the cavity and that we have reached steady state. The optical cavity acts as an optical resonator. Consider an EM wave with an *initial* optical power P_i starting at some point in the cavity and traveling toward the cavity face 1 as shown in Figure 4.22. It will travel the length of the cavity, become reflected at face 1, travel back the length of the cavity to face 2, become reflected at 2, and arrive at the starting point with a *final* optical power P_f . Under steady state conditions, oscillations do not build up and do not

²⁰Commonly known as Füchtbauer–Ladenburg relation, and described in more advanced textbooks.

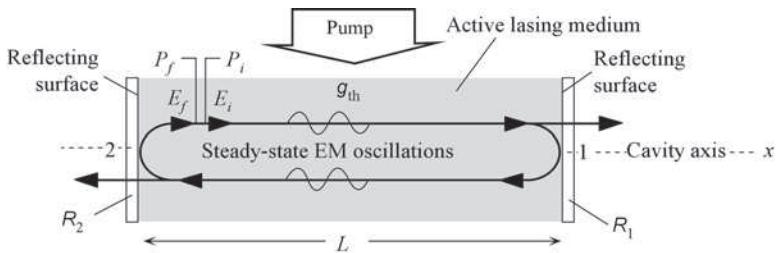


FIGURE 4.22 An optical cavity resonator, and the derivation of the threshold gain condition.

die out, which means that P_f must be the same as P_i . Thus, there should be no optical power loss in the round trip, which means that the **net round-trip optical gain** G_{op} must be unity,

$$G_{\text{op}} = \frac{P_f}{P_i} = 1 \quad (4.6.5)$$

Power condition for maintaining oscillations

Reflections at the faces 1 and 2 reduce the optical power in the cavity by the reflectances R_1 and R_2 of the faces. There are other losses such as some absorption and scattering during propagation in the medium. All these losses have to be made up by stimulated emissions in the optical cavity which effectively provides an optical gain in the medium. As the wave propagates, its power increases as $\exp(gx)$. However, there are a number of losses in the cavity medium acting against the stimulated emission gain such as light *scattering* at defects and inhomogeneities, absorption by impurities, absorption by free carriers (important in semiconductors) and other loss phenomena. These internal cavity losses decrease the power as $\exp(-\alpha_s x)$ where α_s is the **attenuation or loss coefficient of the medium**, also called the **internal cavity loss coefficient**. α_s represents all losses in the cavity and its walls, *except* light transmission losses through the end mirrors and absorption across the energy levels involved in stimulated emissions (E_1 and E_2), which is incorporated into g .²¹

The power P_f of the EM radiation after one round trip of path length $2L$ (Figure 4.22) is given by

$$P_f = P_i R_1 R_2 \exp[g(2L)] \exp[-\alpha_s(2L)] \quad (4.6.6)$$

For steady state oscillations, Eq. (4.6.5) must be satisfied, and the value of the gain coefficient g that makes $P_f/P_i = 1$ is called the **threshold gain coefficient** g_{th} . From Eq. (4.6.6)

$$g_{\text{th}} = \alpha_s + \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right) = \alpha_t \quad (4.6.7)$$

Threshold optical gain coefficient

where α_t is the **total loss coefficient** that represents all the losses in the second term. Equation (4.6.7) gives the optical gain needed in the medium to achieve a continuous wave lasing emission. The right-hand side of Eq. (4.6.7) represents all the losses, that is, the internal cavity losses, α_s , and end losses in $(1/2L) \ln(R_1 R_2)^{-1}$. Threshold gain coefficient g_{th} is that value of the optical gain coefficient that just overcomes all losses, $g_{\text{th}} = \alpha_t$.

²¹ α_s should *not* be confused with the natural absorption coefficient α . The latter represents the absorption from E_1 to E_2 and is already incorporated into g . Note also that some authors use γ for g and some use γ for α_s .

The necessary g_{th} as required by Eq. (4.6.3), that is, the threshold gain, has to be obtained by suitably pumping the medium so that N_2 is sufficiently greater than N_1 . This corresponds to a **threshold population inversion** or $N_2 - N_1 = (N_2 - N_1)_{\text{th}}$. From Eq. (4.6.4)

Threshold population inversion

$$(N_2 - N_1)_{\text{th}} \approx g_{\text{th}} \frac{c \Delta v}{B_{21} n h v_o} \quad (4.6.8)$$

We can now substitute for B_{21} in terms of A_{21} (which can be determined experimentally) from Eq. (4.2.7), i.e., $A_{21}/B_{21} = 8\pi h v^3/c^3$, and also use $A_{21} = 1/\tau_{\text{sp}}$, where τ_{sp} is the spontaneous decay time

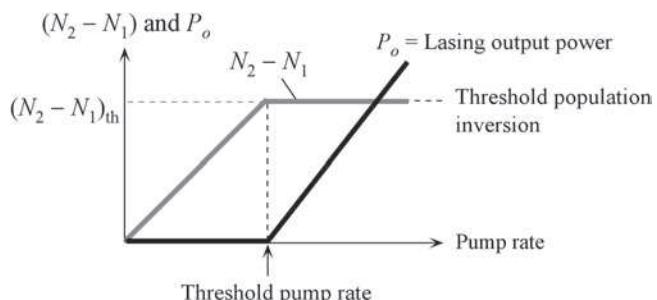
Threshold population inversion

$$(N_2 - N_1)_{\text{th}} \approx g_{\text{th}} \frac{8\pi n^2 v_o^2 \tau_{\text{sp}} \Delta v}{c^2} \quad (4.6.9)$$

Initially the medium must have a gain coefficient g greater than g_{th} . This allows the oscillations to build up in the cavity until a steady state is reached when $g = g_{\text{th}}$. By analogy, an electrical oscillator circuit has an overall gain (loop gain) of unity once a steady state is reached and oscillations are maintained. Initially, however, when the circuit is just switched on, the overall gain is greater than unity. The oscillations start from a small noise voltage and become amplified, that is built-up, until the overall round-trip (or loop) gain becomes unity and a steady state operation is reached. The reflectance of the mirrors R_1 and R_2 are important in determining the threshold population inversion as they control g_{th} in Eq. (4.6.9). It should be apparent that the laser device emitting coherent emission is actually a **laser oscillator**.

The examination of the steady state continuous wave (CW) coherent radiation output power P_o and the population difference $(N_2 - N_1)$ in a laser as a function of the pump rate would reveal the highly simplified behavior shown in Figure 4.23. Until the pump rate can bring $(N_2 - N_1)$ to the threshold value $(N_2 - N_1)_{\text{th}}$, there would be no coherent radiation output. When the pumping rate exceeds the threshold value, then $(N_2 - N_1)$ remains clamped at $(N_2 - N_1)_{\text{th}}$ because this controls the optical gain g , which must remain at g_{th} . Additional pumping increases the *rate of stimulated transitions* and hence increases the photon concentration and the optical output power P_o . Also note that we have not considered how pumping actually modifies N_1 and N_2 , except that $(N_2 - N_1)$ is proportional to the pumping rate as in Figure 4.23.²² Further, the characteristics shown in Figure 4.23 refer to plotting values under steady state or CW operation, and exclude the fact that initially $(N_2 - N_1)$ must be greater than $(N_2 - N_1)_{\text{th}}$ for the oscillations to build-up.

FIGURE 4.23 A simplified description of a laser oscillator. $(N_2 - N_1)$ and coherent output power (P_o) vs. pump rate under continuous wave steady state operation.



²²To relate N_1 and N_2 to the pumping rate we have to consider the actual energy levels involved in the laser operation (three or four levels) and develop *rate equations* to describe the transitions in the system; see, for example, J. Wilson and J. F. B. Hawkes, *Optoelectronics: An Introduction*, 3rd Edition (Prentice-Hall, Pearson Education, 1998), Ch. 5.

EXAMPLE 4.6.1 Threshold population inversion for the He-Ne laser

Consider a He-Ne gas laser operating at the wavelength 632.8 nm (equivalent to $\nu_o = 473.8 \text{ THz}$). The tube length $L = 40 \text{ cm}$ and mirror reflectances are approximately 95% and 100%. The linewidth $\Delta\nu$ is 1.5 GHz, the loss coefficient α_s is 0.05 m^{-1} , the spontaneous decay time constant τ_{sp} is roughly 100 ns, and $n \approx 1$. What are the threshold gain coefficient and threshold population inversion?

Solution

The threshold gain coefficient from Eq. (4.6.7) is

$$g_{\text{th}} = \alpha_s + \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right) = (0.05 \text{ m}^{-1}) + \frac{1}{2(0.4 \text{ m})} \ln\left[\frac{1}{(0.95)(1)}\right] = 0.114 \text{ m}^{-1}$$

The threshold population inversion from Eq. (4.6.9) is

$$\begin{aligned} \Delta N_{\text{th}} &\approx g_{\text{th}} \frac{8\pi n^2 \nu_o^2 \tau_{\text{sp}} \Delta\nu}{c^2} \\ &= (0.114 \text{ m}^{-1}) \frac{8\pi(1)^2 (473.8 \times 10^{12} \text{ s}^{-1})^2 (100 \times 10^{-9} \text{ s})(1.5 \times 10^9 \text{ s}^{-1})}{(3 \times 10^8 \text{ m s}^{-1})^2} \\ &= 1.1 \times 10^{15} \text{ m}^{-3} \end{aligned}$$

Note that this is the threshold population inversion for Ne atoms in configurations $2p^55s^1$ and $2p^53p^1$. The spontaneous decay time τ_{sp} is the natural decay time of Ne atoms from E_2 ($2p^55s^1$) to E_1 ($2p^53p^1$). This time must be much longer than the spontaneous decay from E_1 to lower levels to allow a population inversion to be built up between E_2 and E_1 , which is the case in the He-Ne laser. Equation (4.6.8) was a simplified derivation that used two energy levels: E_2 and E_1 . The He-Ne case is actually more complicated because the excited Ne atom can decay from E_2 not only to E_1 but to other lower levels as well. While the He-Ne is lasing, the optical cavity ensures that the photon density in cavity promotes the E_2 to E_1 transitions to maintain the lasing operation.

C. Output Power and Photon Lifetime in the Cavity

Consider the laser cavity shown in Figure 4.24 that has been pumped so that it is operating under steady state, and emitting CW radiation. Under steady state lasing operation, the coherent photon concentration N_{ph} would neither decay with distance nor with time inside the cavity,

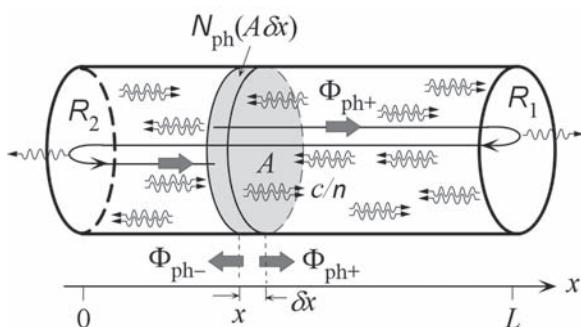


FIGURE 4.24 A pictorial visualization of photons inside a laser optical cavity bouncing back and forth between the cavity ends with some being transmitted. N_{ph} is the *photon concentration* inside the cavity. $\Phi_{\text{ph}+}$ is the *photon flux* in the $+x$ direction. (Schematic illustration only. In reality, the photons in the cavity are much longer.)

inasmuch as g_{th} just balances the total loss α_t . Further, there will be a coherent photon flux along the $+x$ direction $\Phi_{\text{ph}+}$ and another flux $\Phi_{\text{ph}-}$ along $-x$. In the small elemental volume $A\delta x$ shown in Figure 4.24, half of $N_{\text{ph}}(A\delta x)$ is moving toward the $+x$ direction with a velocity c/n . The time δt it takes for photons to travel δx is $\delta t = n\delta x/c$. The photon flux $\Phi_{\text{ph}+}$ is therefore $(1/2)N_{\text{ph}}(A\delta x)/(A\delta t)$ or

Photon
flux in $+x$
direction

$$\Phi_{\text{ph}+} = \frac{1}{2}N_{\text{ph}}(c/n) \quad (4.6.10)$$

The power flowing toward the $+x$ direction is simply $h\nu_o A\Phi_{\text{ph}+}$. A fraction $(1 - R_1)$ will be transmitted at the right-hand-side mirror so that the output power is

Output
power

$$P_o \approx \frac{1}{2}A(1 - R_1)h\nu_o N_{\text{ph}}(c/n) \quad (4.6.11)$$

Thus, the output power, as expected, depends on the photon concentration N_{ph} in the cavity, and the latter depends on the rate of stimulated transitions.

The total cavity loss coefficient $\alpha_t = \alpha_s + (1/2L)\ln(R_1R_2)^{-1}$ has two parts, in which α_s is the internal loss coefficient and $(1/2L)\ln(R_1R_2)^{-1}$ represents the **end losses**. These losses would cause the photon concentration N_{ph} to decay exponentially with time with a characteristic time constant, called the **photon cavity lifetime** τ_{ph} , that depends inversely on α_t . Consider the photon flux $\Phi_{\text{ph}+}$ shown in Figure 4.24 that starts at x and is traveling toward the $+x$ direction. It becomes reflected at R_1 and then at R_2 , and then it returns back to x with a smaller magnitude as we are considering just the effect of all losses represented by α_t . The round-trip time interval $\delta t = 2nL/c$. The attenuated photon flux after one round trip is $\Phi_{\text{ph}}R_1R_2\exp(-2\alpha_s L)$ and the change $\delta\Phi_{\text{ph}+}$ is $-\Phi_{\text{ph}}[1 - R_1R_2\exp(-2\alpha_s L)]$, where the negative sign indicates a decrease. Further, we can eliminate α_s and R_1R_2 by using the total attenuation coefficient α_t so that

$$\delta\Phi_{\text{ph}+} = -\Phi_{\text{ph}}[1 - \exp(-2\alpha_t L)]$$

The fractional change in the photon concentration per unit time is then

Photon
cavity
lifetime

$$\frac{\delta N_{\text{ph}}}{N_{\text{ph}}\delta t} = \frac{\delta\Phi_{\text{ph}+}}{\Phi_{\text{ph}+}\delta t} = -\frac{1 - \exp(-2\alpha_t L)}{2Ln/c} = -\frac{1}{\tau_{\text{ph}}} \quad (4.6.12)$$

Clearly, if we ignore the gain of the medium, the above equation shows that losses would cause the photon concentration N_{ph} to decay exponentially with time as $N_{\text{ph}}(t) = N_{\text{ph}}(0)\exp(-t/\tau_{\text{ph}})$.²³ Assuming small losses, we can expand $\exp(-2\alpha_t L) \approx 1 - 2\alpha_t L$ so that the **photon cavity lifetime** is given by

Photon
cavity
lifetime

$$\tau_{\text{ph}} \approx n/c\alpha_t \quad (4.6.13)$$

Thus, the total attenuation in the optical cavity, as represented by the coefficient α_t , is equivalent to a cavity photon concentration that decays exponentially with time with a time constant τ_{ph} given by Eq. (4.6.13).

²³If the change δN_{ph} is small over the time δt , then we can use differentials for changes. Further, most books simply use $\tau_{\text{ph}} = n/c\alpha_t$ for the photon cavity lifetime.

EXAMPLE 4.6.2 Output power and photon cavity lifetime τ_{ph}

Consider the He-Ne laser in Example 4.6.1 that has a tube length of 40 cm and $R_1 = 0.95$ and $R_2 = 1$. Suppose that the tube diameter is 0.8 mm, and the output power is 2.5 mW. What are the photon cavity lifetime and the photon concentration inside the cavity? (The emission frequency v_o is 474 THz.)

Solution

Using $L = 40$ cm, $R_1 = 0.95$, $R_2 = 1$, $\alpha_s = 0.05 \text{ m}^{-1}$ gives

$$\alpha_t = \alpha_s + (1/2L)\ln(R_1R_2)^{-1} = 0.05 \text{ m}^{-1} + [2(0.4 \text{ m})]^{-1}\ln[(0.95 \times 1)]^{-1} = 0.114 \text{ m}^{-1}$$

and hence from Eq. (4.6.12),

$$\tau_{ph} = [(2)(1)(0.4)] / [(3 \times 10^8)(1 - e^{-2 \times 0.114 \times 0.4})] = 30.6 \text{ ns}$$

If we use Eq. (4.6.13) we would find 29.2 ns. To find the photon concentration, we use Eq. (4.6.11),

$$\begin{aligned} P_o &= (0.0025 \text{ W}) \approx \frac{1}{2}A(1 - R_1)hv_oN_{ph}c/n \\ &= \frac{1}{2}[\pi(8 \times 10^{-3}/2)^2](1 - 0.95)(6.62 \times 10^{-34})(474 \times 10^{12})N_{ph}(3 \times 10^8)/(1) \end{aligned}$$

which gives $N_{ph} \approx 2.1 \times 10^{15}$ photons m^{-3} .

D. Optical Cavity, Phase Condition, Laser Modes

The laser oscillation condition stated in Eq. (4.6.5) which leads to the threshold gain g_{th} in Eq. (4.6.7) considers only the intensity of the radiation inside the cavity. The examination of Figure 4.22 reveals that the initial wave E_i with power P_i attains a power P_o after *one round trip* when the wave has arrived back exactly at the same position as E_f , as shown in Figure 4.22. Unless the total phase change after one round trip from E_i to E_f is a multiple of 2π , the wave E_f cannot be identical to the initial wave E_i . We therefore need the additional condition that the round-trip phase change $\Delta\phi_{\text{round-trip}}$ must be a multiple of 360°

$$\Delta\phi_{\text{round-trip}} = m(2\pi) \quad (4.6.14)$$

Phase condition for laser oscillations

where m is an integer, 1, 2, This condition ensures *self-replication* rather than self-destruction. There are various factors that complicate any calculation from the phase condition in Eq. (4.6.14). The refractive index n of the medium in general will depend on pumping (especially so in semiconductors), and the end-reflectors can also introduce phase changes. In the simplest case, we can assume that n is constant and neglect phase changes at the mirrors. If $k = 2\pi/\lambda$ is the free-space propagation constant, only those special k -values, denoted as k_m , that satisfy Eq. (4.6.14) can exit as radiation in the cavity, *i.e.*, for propagation along the cavity axis

$$nk_m(2L) = m(2\pi) \quad (4.6.15)$$

Laser cavity modes

which leads to the usual mode condition

$$m\left(\frac{\lambda_m}{2n}\right) = L \quad (4.6.16)$$

Wavelengths of laser cavity modes

Thus, our intuitive representation of modes as standing waves, as originally described by Eq. (4.5.4) and again by Eq. (4.6.16), is a simplified conclusion from the general phase condition in Eq. (4.6.14). Furthermore, the modes in Eq. (4.6.16) are controlled by the length L of the optical cavity along its axis and are called **longitudinal axial modes**. We can also write the modes in terms of their frequencies ν_m inasmuch as $\lambda_m = c/\nu_m$

*Frequencies
of laser
cavity modes*

$$\nu_m = \frac{mc}{2nL} \quad (4.6.17)$$

and the separation of modes can be found from Eq. (4.6.17), $\Delta\nu_m = c/2nL$.

In the discussions of threshold gain and phase conditions, we referred to Figure 4.22 and tacitly assumed plane EM waves traveling inside the cavity between two perfectly flat and aligned mirrors. A plane wave is an idealization as it has an infinite extent over the plane normal to the direction of propagation. All practical laser cavities have a finite transverse size, that is, perpendicular to the cavity axis. Furthermore, not all cavities have flat reflectors at the ends. In gas lasers, one or both mirrors at the tube ends may be spherical to allow a better mirror alignment as illustrated in Figures 4.25 (a) and (b). One can easily visualize off-axis self-replicating rays that can travel off the axis as shown in one example in Figure 4.25 (a). Such a mode would be non-axial. Its properties would be determined not only by the off-axis round-trip distance, but also by the transverse size of the cavity. The greater the transverse size, the more of these off-axis modes can exist. Further, notice that the ray in Figure 4.25 (a) traverses the tube length almost four times before it completes a true round trip, whereas the total round-trip path was $2L$ in the case of flat mirrors.

A better way of thinking about modes is to realize that a mode represents a particular electric field pattern in the cavity that can replicate itself after one round trip. Figure 4.25 (b) shows how a wavefront of a particular mode starts parallel to the surface of one of the mirrors, and after one round trip, it replicates itself. The wavefront curvature changes as the radiation propagates in the cavity and it is parallel to mirror surfaces at the end mirrors. Such a mode has similarities to the Gaussian beam discussed in Chapter 1. Indeed, in many applications, laser beams are modeled by assuming they are Gaussian beams.

More generally, whether we have flat or spherical end mirrors, we can find all possible allowed modes by considering what spatial field patterns at one mirror can self-replicate

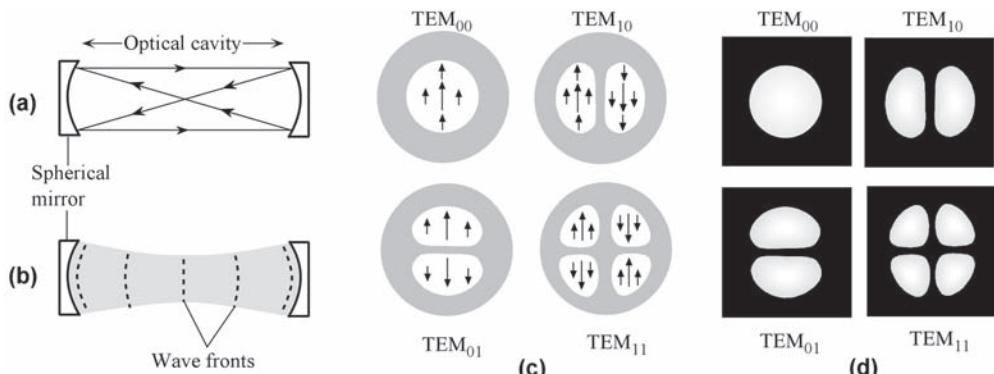


FIGURE 4.25 Laser modes. (a) An off-axis transverse mode is able to self-replicate after one round trip. (b) Wavefronts in a self-replicating wave. (c) Four low-order transverse cavity modes and their fields. (d) Intensity patterns in the modes of (c).

after one round trip²⁴ through the cavity to the other mirror and back as in the example in Figure 4.25 (b). A mode is a certain field pattern at a reflector that can propagate to the other reflector and back again and return the same field pattern. All these modes can be represented by fields \mathbf{E} and \mathbf{B} that are nearly normal to the cavity axis; they are referred to as **transverse modes** or **transverse electric and magnetic (TEM) modes**.²⁵ Each allowed mode corresponds to a distinct spatial field distribution at a reflector. These modal field patterns at a reflector can be described by three integers, p , q , and m , and designated by TEM_{pqm} . The integers p , q represent the number of nodes in the field distribution along the transverse directions y and z to the cavity axis x (put differently, across the beam cross-section). The integer m is the number of nodes along the cavity axis x and is the usual longitudinal mode number. Figures 4.25 (c) and (d) shows the field patterns for four TEM modes and the corresponding intensity patterns for four example TEM modes. Each transverse mode with a given p , q has a set of longitudinal modes (m values) but usually m is very large ($\sim 10^6$ in gas lasers) and is not written, though understood. Thus, transverse modes are written as TEM_{pq} and each has a set of longitudinal modes ($m = 1, 2, \dots$). Moreover, two different transverse modes may not necessarily have the same longitudinal frequencies implied by Eq. (4.6.7). (For example, n may not be spatially uniform and different TEM modes have different spatial field distributions.)

Transverse modes depend on the optical cavity dimensions, reflector sizes, and other size-limiting apertures that may be present in the cavity. The modes either have **Cartesian (rectangular)** or **polar (circular)** symmetry about the cavity axis. Cartesian symmetry arises whenever a feature of the optical cavity imposes a more favorable field direction; otherwise, the patterns exhibit circular symmetry. The examples in Figures 4.25 (c) and (d) possess rectangular symmetry and would arise, for example, if polarizing Brewster windows are present at the ends of the cavity.

The lowest order mode TEM_{00} has an intensity distribution that is radially symmetric about the cavity axis, and has a Gaussian intensity distribution across the beam cross-section everywhere inside and outside cavity. It also has the lowest divergence angle. These properties render the TEM_{00} mode highly desirable, and many laser designs optimize on TEM_{00} while suppressing other modes. Such lasers usually require restrictions in the transverse size of the cavity.

4.7 BROADENING OF THE OPTICAL GAIN CURVE AND LINewidth

If the upper and lower energy levels E_2 and E_1 in laser transitions were truly discrete levels, we might think that we would get a stimulated emission at one frequency only at $h\nu_o = E_2 - E_1$, and the optical gain curve would be a delta function, a narrow line at ν_o . At least in terms of modern physics, we know that this would be impossible based on the Uncertainty Principle alone. Suppose that we wish to measure the energy E_2 by some perfectly designed instrument. Suppose, further, that lifetime of the atom in this E_2 -state is τ_2 after which it ends up at E_1 . Then, according to the Uncertainty Principle, if ΔE_2 is the uncertainty in E_2 , we must obey $\Delta E_2 \tau_2 > \hbar$, which clearly shows that we cannot think of E_2 as a well-defined discrete single-valued energy level. Consequently the emitted radiation will have this finite minimum spectral width $h\Delta\nu = \Delta E_2$.

²⁴We actually have to solve Maxwell's equations with the boundary conditions of the cavity to determine what EM wave patterns are allowed. Further we have to incorporate optical gain into these equations (not a trivial task).

²⁵Or, *transverse electromagnetic modes*.

Exactly the same Uncertainty Principle argument applies to E_1 , and if τ_1 is the lifetime of the atom in a state at E_1 , from which it decays to another lower energy state, then we must obey $\Delta E_1 \tau_1 > \hbar$. Thus, the emission spectrum involves transitions from a small region ΔE_2 around E_2 to a small region ΔE_1 around E_1 , and consequently has a finite spectral width determined by τ_2 and τ_1 . This type of broadening of the emission spectrum, and hence the optical gain curve (its lineshape), is called **lifetime broadening**.

In lifetime broadening, all atoms in the medium will exhibit the same broadening. They will all have the same central frequency v_o and spectral width Δv , as shown in Figure 4.26 (a). This type of spectral broadening in which *all* the atoms in a medium generate the same emission curve with the same v_o and subjected to the same broadening mechanism is called **homogeneous broadening**. The shape of the emission curve is called a **Lorentzian lineshape**, and its mathematical form is well-known in photonics. (See Question 4.18.) The spectral width due to lifetime broadening is known as the **natural linewidth**, which is the minimum linewidth exhibited by all atoms.

Another homogeneous broadening mechanism that is of particular practical importance is the emission from ions embedded in crystals, for example, in a laser based on Nd^{3+} ions in a YAG crystal, or Ti^{3+} ions in an Al_2O_3 (sapphire) crystal. The Nd^{3+} ions in the YAG crystal ($\text{Y}_3\text{Al}_5\text{O}_{12}$) will be situated at well-defined sites, and their emission characteristics would be well-defined in the absence of any interruptions. However, lattice vibrations, *i.e.*, phonons, will interact, that is *collide*, with the Nd^{3+} ions (phonons will “jolt” the Nd^{3+} ions) and interrupt the emission process. Phonon collisions suddenly change the phase of the EM wave during emission. This type of broadening mechanism is called **collision broadening**, which is also homogeneous, because all Nd^{3+} ions experience the same broadening. Phonons interact with all the ions in the same way. Collision broadening is much stronger than lifetime broadening. The spectral width of the Nd^{3+} :YAG laser is 120 GHz and is due to phonon collision broadening.

Homogeneous collision broadening can also take place in gas lasers in which the atoms collide with each other and interrupt the radiation emission process, either terminate it or suddenly change the phase of the emitted radiation. In the case of gas lasers, collision broadening is usually called **pressure broadening** because the broadening increases with the gas pressure as more and more atoms collide with each other. While for the He-Ne laser, pressure broadening is negligible compared with Doppler broadening, it is nonetheless a contributing factor in certain gas lasers such as the CO_2 laser.

We saw that in the He-Ne laser, individual Ne atoms emitted at different frequencies due to the Doppler effect as illustrated in Figure 4.26 (b). If they were all stationary they would all have the same emission characteristic, *i.e.*, a homogeneously broadened lineshape centered at v_o ; but, they move around randomly in the gas. The Doppler effect shifts the emission curve of each Ne atom an amount that depends on the velocity of the Ne atom and its direction. Since the atoms are

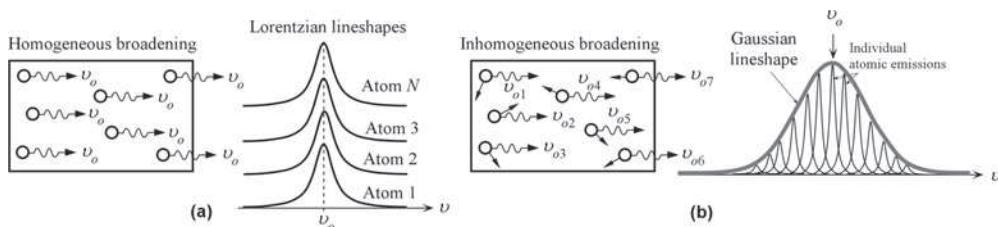


FIGURE 4.26 Lineshapes for (a) homogeneous and (b) inhomogeneous broadening. The lineshape is Lorentzian for homogeneous and Gaussian for inhomogeneous broadening.

moving randomly, the shift can be considered to be random. With a large number of Ne atoms, these different randomly shifted emission curves add to produce an overall emission lineshape that is a **Gaussian**. (The Gaussian lineshape is not unexpected since we have a very large collection of “random processes.”) This type of broadening in which each atom emits at a slightly different central frequency is called **inhomogeneous broadening**. The overall spectral width is obviously much wider than the homogeneous broadening inherent in individual atomic emissions as indicated in Figure 4.26 (b). In the case of 632.8 nm emission, homogeneous lifetime broadening is only 14 MHz, whereas inhomogeneous Doppler broadening is 1.5 GHz, two orders of magnitude wider.

Inhomogeneous broadening also occurs in solid state glass lasers, for example, in Nd³⁺:glass lasers. Each isolated Nd³⁺ ion outside the glass structure would emit the same spectrum. However, within the glass structure, the Nd³⁺ ions find themselves in different local environments, with different neighboring ions. These neighboring ions are not always the same and always exactly at the same place for each Nd³⁺ ion. These variations in the local environment from site to site in the glass result in the energy levels of the Nd³⁺ ion becoming spread. The overall emission characteristic is a Gaussian lineshape because the variations from one Nd³⁺ site to another are random. This type of inhomogeneous broadening is often termed **amorphous structure broadening**. The spectral broadening in Nd³⁺:glass lasers are typically 5–7 THz, which is much wider than that for Nd³⁺:YAG above.

Table 4.2 summarizes the basic characteristics of homogeneous and inhomogeneous broadening and provides simple examples.²⁶ The optical gain lineshape and its width are key parameters in laser design, especially so in mode-locked lasers described in the next section.

The lasing radiation output from a perfect homogeneously broadened laser medium is normally a single mode. On the other hand, the lasing output from an inhomogeneously broadened medium can be multimode or single mode, depending on the overlap of the gain curve with the cavity modes.

Suppose that we manage to pump a homogeneously broadened gain medium above threshold gain g_{th} for an instant as shown in Figure 4.27 (a). Suppose that there are three modes of the cavity that are allowed under the gain curve above g_{th} , marked at ν_o , ν_1 , and ν'_1 in Figure 4.27 (a).

TABLE 4.2 Comparison of homogeneous and inhomogeneous broadening mechanisms on the emission spectrum, or the optical gain curve

	Main characteristics	Lineshape	Examples of lasers
Homogeneous broadening	All atoms emit the same spectrum with the same center frequency ν_o Single-mode lasing output	Lorentzian function	Nd ³⁺ :YAG; phonon collision broadening Pressure broadening in CO ₂ lasers
Inhomogeneous broadening	Different atoms emit at slightly differing central frequencies, due to random processes shifting the peak emission frequency Multimode or single-mode lasing output	Gaussian function	He-Ne lasers; Doppler broadening Nd ³⁺ : glass lasers; amorphous structure broadening

²⁶There are also many practical examples in which the lineshape is neither exactly Lorentzian nor Gaussian, but a combination of the two. When the two functions are suitably combined, the result is a Voigt lineshape.

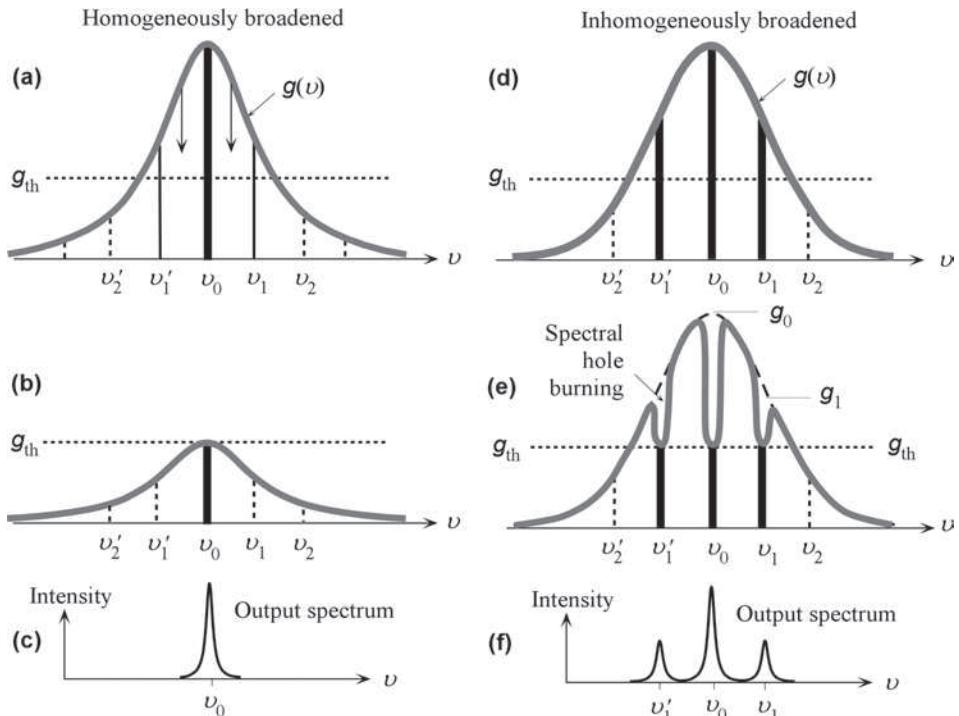


FIGURE 4.27 (a) An ideal homogeneously broadened gain medium that has been pumped above threshold for an instant. The mode at v_0 is the most intense and de-excites atoms by stimulated emission and reduces the population inversion for all the modes, which reduces the gain. (b) The gain $g(v)$ is reduced until the threshold is reached, and only the v_0 mode can oscillate. (c) The output spectrum is a single mode. (d) An ideal inhomogeneously broadened gain medium that has been pumped above threshold. The atoms behave independently and each has its own gain curve. The de-excitation of atoms by stimulated radiation at v_0 does not affect the atoms at v_1 and v_1' . (e) The gain $g(v)$ must be reduced at v_0 , v_1 , and v_1' until threshold is reached to sustain steady state oscillations. The gain spectrum thus has dips at v_0 , v_1 , and v_1' . (f) The output spectrum has three modes at v_0 , v_1 , and v_1' .

The v_0 -mode has the highest intensity, and the v_0 -radiation de-excites (by stimulated emission) a much greater number of atoms than v_1 - and v_1' -radiation. Remember that all the atoms have the same gain curve, which means that the v_0 -radiation removes a large number of excited atoms from interacting with v_1 - or v_1' -radiation. The emissions stimulated by the v_0 -radiation cause even more atoms to become de-excited, leaving very little number of atoms for the v_1 - and v_1' -radiation. Put differently, the stimulated emissions induced by the highest intensity v_0 -radiation reduce the population inversion $N_2 - N_1$ and hence the gain for all, as illustrated by the vertical downward arrows in Figure 4.27 (a). The modes at v_1 and v_1' , however, are affected very badly since their gains were already less than that of v_0 . As the gain curve is reduced, in the end, as shown in Figure 4.27 (b), only the v_0 -radiation can sustain itself. The output spectrum is a single mode as shown in Figure 4.27 (c).

Consider now a laser gain medium that has inhomogeneous broadening and suppose that the gain curve has three modes with gains above g_{th} as shown in Figure 4.27 (d). In this case, each atom behaves independently with its own gain curve that is centered away from the others as in Figure 4.26 (b). Consequently, the v_0 -mode radiation only interacts with those atoms that have their individual gain curves overlapping the v_0 -radiation. Similarly, v_1 -radiation interacts

only with those atoms that have their individual gain curves overlapping the ν_1 -radiation. When the ν_0 -radiation de-excites atoms, it does not de-excite those that are at frequencies ν_1 and ν'_1 . We know that stimulated emissions de-excite atoms and reduce the population inversion and hence the gain. Consequently, under steady state operation, the gain for the ν_0 -mode is reduced from its peak g_0 down to g_{th} , which is the required gain for maintaining steady state oscillations. Similarly, the gain for ν_1 -mode is reduced from g_1 down to g_{th} , as shown in Figure 4.27 (e). To maintain steady state laser oscillations, $g(\nu)$ must have dipped regions at ν_0 , ν_1 , and ν'_1 where $g(\nu)$ becomes g_{th} . The dip in $g(\nu)$ at a mode is called **spectral hole burning**. It might seem as though all three modes should have the same intensity as they all have $g = g_{\text{th}}$, but this is not the case. Consider the ν_0 - and ν_1 -modes. The reduction of the gain from its peak at g_0 to g_{th} , and its maintenance at g_{th} involves stimulated emissions for de-excitation. More stimulated emissions are needed to reduce g_0 to g_{th} than for reducing g_1 to g_{th} . Thus ν_0 -mode has greater stimulated emissions and more optical power than the ν_1 -mode. The output spectrum is illustrated in Figure 4.27 (f). (As one might have surmised, for simplicity only, the mode ν_0 was made coincident with the peak of the gain curve.)

4.8 PULSED LASERS: Q-SWITCHING AND MODE LOCKING

A. Q-Switching

The Q -factor of an optical resonant cavity, like its finesse, is a measure of the cavity's frequency selectiveness. The higher the Q -factor, the more selective the resonator is, or the narrower the spectral width. It is also a measure of the energy stored in the resonator per unit energy dissipated, *e.g.*, losses at the reflecting surfaces or scattering in the cavity, per cycle of oscillation.²⁷ **Q -switching** refers to a laser whose optical resonant cavity is switched from a low Q to a high Q to generate an intense laser pulse, as described below.

A laser operation needs a good optical resonant cavity with low losses, that is, a high Q -factor, to allow the gain to reach the threshold gain. The large radiation intensity within the cavity stimulates further emissions. However, by temporarily removing one of the reflectors as in Figure 4.28 (a), we can switch the Q -factor of the cavity to some low value. While the optical

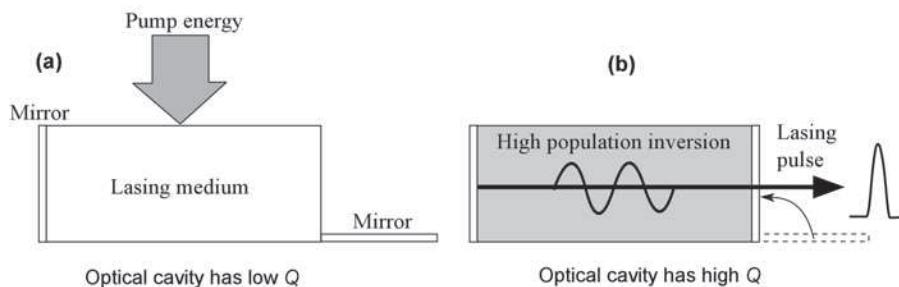


FIGURE 4.28 (a) The optical cavity has a low Q so that pumping takes the atoms to a very high degree of population inversion; lasing is prevented by not having a right-hand mirror. (b) The right mirror is “flung” to make an optical resonator, Q is switched to a high value which immediately encourages lasing emissions. There is an intense pulse of lasing emission which brings down the excess population inversion.

²⁷ See Section 1.11 in Chapter 1. All optical resonators have a Q -factor, which is another way of expressing their finesse.

cavity has low Q , the lasing is suppressed. Without the reflections from the mirrors, the photon energy density in the cavity is small, and the active medium can be pumped to achieve a large population inversion, greater than what would have normally been achieved had the mirrors been in place. When the Q of the optical cavity is switched to a high value, as in Figure 4.28 (b), an intense lasing emission is generated. While the Q is low, the pumped lasing medium is effectively a very high gain photon amplifier. There is too much loss in the optical resonator (*i.e.*, no optical feedback) to achieve a lasing oscillation. As soon as the Q is switched to a high value, the low loss in the optical resonator allows lasing oscillations to occur, which deplete the population inversion and decrease the gain until the population inversion falls below the threshold value and lasing oscillations cease.

The sequence of events involved in Q -switching a laser is illustrated in Figure 4.29. The pump, a flash tube in many cases, that provides optical excitation, is triggered at time t_1 as shown in Figure 4.29 (a). It must have sufficient intensity to quickly build up $(N_2 - N_1)$ to well above the normal threshold value $(N_2 - N_1)_{\text{th}}$ as indicated in Figure 4.29 (b). The pumping power needed must be provided before excited ions (or atoms) in the medium can spontaneously decay to the lower laser energy level. The optical cavity's Q is switched at time t_2 , before the excited ions decay spontaneously to the lower laser level, as shown in Figure 4.29 (c). The result is a nearly sudden formation of an optical cavity, which enables lasing radiation to be quickly built up inside the cavity from t_2 to t_3 in Figures 4.29 (c)–(d). The radiation buildup from t_2 to t_3 is of the order of nanoseconds and starts with stray photons as in normal CW laser operation. Eventually, the rapid build-up of radiation in the cavity stimulates lasing transitions, and acts as positive feedback, yielding a pulse of coherent radiation as output as in Figure 4.29 (d). The duration of the lasing emission (the width of the output pulse) depends on both the Q -switching time, and the photon cavity lifetime, *i.e.*, how long photons are in the cavity before they leave. The Q -switching speed must be faster than the rate of oscillation buildup; otherwise the lasing oscillation builds up slowly. Table 4.3 summarizes the properties of two commercial Q -switched Nd³⁺:YAG lasers. Notice that the Q -switched pulses are a few

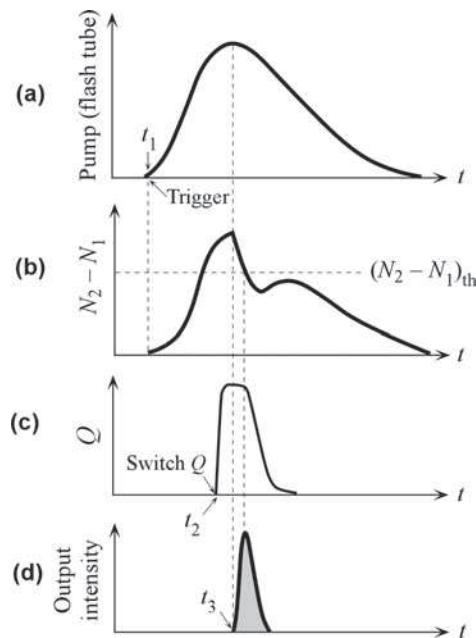


FIGURE 4.29 A simplified schematic of Q -switching in the generation of short laser pulses. (a) The pump profile, *i.e.*, the flash tube output. (b) The evolution of the population difference $N_2 - N_1$ with time. (c) The switching of the optical cavity Q during the pumping procedure while the population inversion is very large and greater than the normal threshold population. (d) The output light pulse.

TABLE 4.3 Selected typical characteristics of two commercial *Q*-switched Nd^{3+} :YAG lasers

Property	Nd^{3+} :YAG	Nd^{3+} :YAG
	NL220 by EKSPLA	P-1064-150-HE by Alphalas
Emission wavelength (nm)	1064	1064
Output pulse energy (mJ)	10	1.5
Pulse width (ns)	7	1.1
Repetition rate (Hz)	1000	100
Peak power (kW)	~1400	~1400
<i>Q</i> -switch	EO	Passive
Output beam	TEM_{00}	TEM_{00}
Output beam M^2	<1.5	—
Beam divergence (mrad)	<1.5	6
Beam diameter (mm)	2.5	0.3

Note: Usually higher-energy output pulses require operation at a lower repetition frequency.

nanoseconds long, and the actual power in the laser pulse is 1.4–1.8 MW; extremely high compared with CW power levels.

There are a number of ways in which *Q*-switching can be conveniently implemented as illustrated in three examples in Figures 4.30 (a)–(c). The first example uses a rotating reflector, which in this case is a prism. The cavity *Q* is switched during the rotation of the prism every time a prism face aligns with the cavity axis, which is shown in Figure 4.30 (a). Another technique is to use a *saturable absorber*, also called a *bleachable dye*, in the optical cavity as in Figure 4.30 (b). A saturable absorber's absorption decreases with increasing light intensity so that it becomes transparent only at high intensities. When the absorber becomes transparent under a large radiation intensity, the dye is said to be bleached. Initially the absorber is opaque and keeps the cavity *Q* low by absorbing the radiation, but as the intensity builds up the absorber becomes transparent, which corresponds to switching the cavity *Q* to a high value. Another technique that is widely employed is the use of an *electro-optic* (EO) switch in the optical cavity as illustrated in Figure 4.30 (c). As will be explained in Chapter 6, such an EO switch is essentially an electro-optic (or a piezoelectric) crystal with electrodes, and works in combination with a polarizer. EO switches can be very fast (*e.g.*, less than nanoseconds), which is one of their distinct advantages. For the present purposes it is clear that when the EO switch is turned transparent by the application of a voltage, the cavity *Q* is switched high, which results in the generation of a lasing pulse.

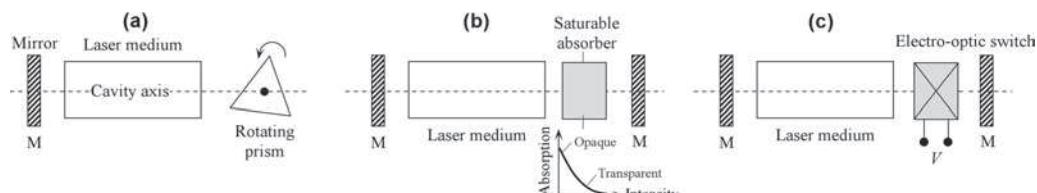


FIGURE 4.30 (a) *Q*-switching by using a rotating prism. (b) *Q*-switching by using a saturable absorber. (c) *Q*-switching by using an electro-optic (EO) switch. Normally a polarizer is also needed before or after the switch but this is part of the EO switch in this diagram.

B. Mode Locking

Mode locking is used to generate short and intense laser light pulses at a certain repetition rate that depends on the laser construction. A **mode-locked laser** is a laser that has been constructed to have one transverse mode and many (N) longitudinal modes, as in Figure 4.31 (a), that have the *same phase*. The N longitudinal modes then *reinforce* each other to generate an intense pulse of lasing emission at fixed time intervals as illustrated in Figure 4.31 (b). The repetition rate depends on the laser cavity length L and the speed of light in the lasing medium, c , assuming that the refractive index n is nearly 1 (e.g., we have a gas laser). Normally, the longitudinal modes of a laser cavity would be “independent” with random relative phases. In such a case, the output intensity from the laser would simply be the sum of the individual mode intensities, *i.e.*, $I_1 + I_2 + \dots + I_N$. If, on the other hand, the modes have been forced to have the same phase, then these mode oscillations would *reinforce* each other, *the optical fields would add*, in such a way that they generate an intense optical pulse at a certain repetition rate. In this mode-locked case, the intensity in a pulse is proportional to $(E_1 + E_2 + \dots + E_N)^2$, and can be enormously larger than the case when the modes are not locked, as indicated in Figure 4.31 (b) where the horizontal dashed line, CW, represents the intensity in the unlocked laser. The repetition rate in a mode-locked laser is the reciprocal of the round-trip time of a light pulse in the optical cavity.

Suppose that $T = 2L/c$ is the round-trip time of a light pulse in the optical resonator shown in Figure 4.31 (c), where L is the cavity length and c is the velocity of light. The pulse repeats itself every $T = 2L/c$ seconds. The situation is analogous to taking a fixed number of sine waves with multiple frequencies, adjusting their phases so that one can obtain the maximum amplitude from their summation. The achievement of such an intense light pulse at every T seconds requires that the “modes are locked,” that is, the relative phases of the modes have been correctly adjusted and fixed to yield the required maximum output intensity. When modes have been locked, there must be an optical pulse in the resonator that is traveling between the mirrors with exactly the required round-trip time of $T = 2L/c$. (If the refractive index n is not unity, we need to use c/n instead of c .)

The width Δt of an individual light pulse depends on the frequency width $\Delta\nu$ of the laser optical gain curve; normally $\Delta\nu$ is taken as the full width half maximum width. We know that $\Delta\nu$ also determines how many modes N are there in the output. The pulse width $\Delta t \approx 1/\Delta\nu \approx T/N$. In practice, it is not very difficult to obtain mode locking since the required output must correspond

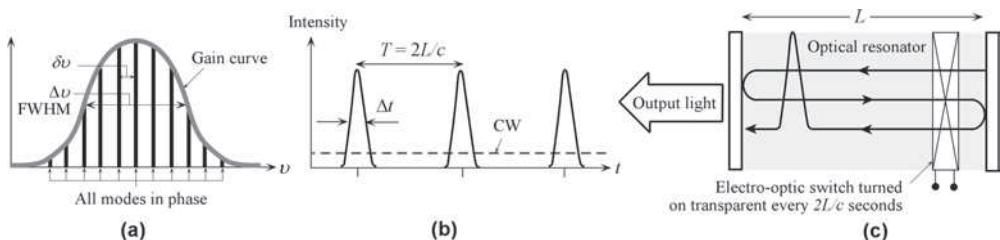


FIGURE 4.31 (a) A mode-locked laser has its N modes all in phase so that the modes add correctly to generate a short laser pulse every T seconds. $\Delta\nu$ is the full width at half maximum. (b) The output light intensity from a mode-locked laser is a periodic series of short intense optical pulses that are separated in time by $T = 2L/c$, the round-trip time for the pulse in the resonator. (c) A laser can be mode-locked by using an EO switch in the optical cavity that becomes transparent exactly at the right time, every $T = 2L/c$ seconds. Each time the pulse in the resonator impinges on the left mirror, every $T = 2L/c$ seconds, a portion of it is transmitted, which constitutes the output from a mode-locked laser.

to an intense pulse train of repetition rate $1/T$. This output itself corresponds to a single optical pulse in the resonator bouncing back and forward between the mirrors with a round-trip time T . Each time this pulse impinges on a partially reflecting mirror, a portion of it is transmitted as an output pulse as in Figure 4.31 (c). Suppose that we insert an *electro-optic switch* that is switched to be made transparent at every T seconds (as we did for Q -switching). The switch is on to be transparent only when the pulse is there and only for the duration of the pulse. Then, the only possible situation is where a single optical pulse can bounce back and forward in the resonator and this corresponds to locking the modes. This is an example of **active mode locking**. The difference from the electro-optically Q -switched laser is that in mode locking, the EO switch must be turned on every T seconds for a very short time, so the timing in sequence is critical.

Another possibility is **passive mode locking** where a saturable absorber is used in the optical cavity (again, similar to Q -switching). The optical cavity is lossy for low light intensities. Such an absorber would only allow a high intensity light pulse to exist in the cavity since it is only transparent at high intensities. The latter would correspond to various modes having the right phases to yield an intense pulse, that is, mode locking. The difference from the Q -switching case is that the absorber in a mode-locked laser must be able to respond faster than the time T .

The output pulses from a mode-locked laser are spatially separated by $2L$ because the pulses come out at every T seconds from the output mirror and T_c is $2L$. A mode-locked He-Ne laser typically has a pulse width of roughly 600 ps whereas it is 150 fs for a mode-locked Nd³⁺:glass laser. It might be thought that we have generated a more powerful output laser beam, a stream of high intensity pulses, than the CW (unlocked) case, but this is not true. The peak power in an individual mode-locked pulse, roughly the energy in the pulse divided by the pulse width Δt , can be extremely high, often in megawatts, but the time-averaged power (over many pulses) is the same as that in the CW unlocked laser.

4.9 PRINCIPLE OF THE LASER DIODE²⁸

Consider a degenerately doped direct bandgap semiconductor *pn* junction whose band diagram is shown in Figure 4.32 (a). By degenerate doping we mean that the Fermi level E_{Fp} in the *p*-side is in the valence band (VB) and E_{Fn} in the *n*-side is in the conduction band (CB). All energy levels up to the Fermi level can be taken to be occupied by electrons as in Figure 4.32 (a). In the absence of an applied voltage, the Fermi level is continuous across the diode, $E_{Fp} = E_{Fn}$. The depletion region or the space charge layer (SCL) in such a *pn* junction is very narrow. There is a built-in voltage V_o that gives rise to a potential energy barrier eV_o that prevents electrons in the CB of *n*⁺-side diffusing into the CB of the *p*⁺-side.²⁹ There is a similar barrier stopping the hole diffusion from *p*⁺-side to *n*⁺-side.

Recall that when a voltage is applied to a *pn* junction device, the change in the Fermi level from end-to-end is the electrical work done by the applied voltage, that is, $\Delta E_F = eV$. Suppose

²⁸The first semiconductor lasers used GaAs *pn* junctions, as reported by the American researchers Robert N. Hall *et al.* (General Electric Research, Schenectady), Marshall I. Nathan *et al.* (IBM, Thomas J. Watson Research Center), and Robert H. Rediker *et al.* (MIT), in 1962. Nick Holonyak (General Electric at Syracuse), also in 1962, reported a laser diode based on the compound GaAsP, which emitted in the visible (red).

²⁹The potential energy barrier eV_o in a nondegenerate *pn* junction is normally taken as the energy required for an electron to move from E_c on the *n*-side to E_c on the *p*-side since electrons are near the bottom of the CB at E_c . In the present case, electrons occupy states from E_c up to E_{Fn} . The exact analysis of the problem is therefore more complicated; however $E_{Fn} - E_c$ is small compared to eV_o .

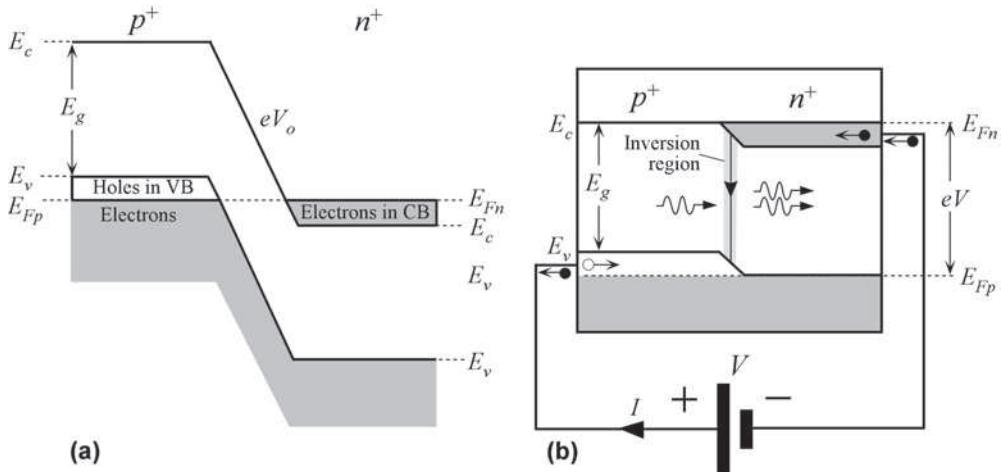


FIGURE 4.32 (a) The energy band diagram of a degenerately doped pn with no bias. eV_o is the potential energy barrier against electron diffusion from the n^+ -side to the p^+ -side. Note that $(E_{Fn} - E_c)$ and $(E_v - E_{Fp})$ are small compared with eV_o . (The diagram is exaggerated.) (b) Band diagram with a sufficiently large forward bias to cause population inversion and hence stimulated emission. (As before, filled black circles are electrons and empty circles are holes. Only electrons flow in the external circuit.)

that this degenerately doped pn junction is forward biased by a voltage V greater than the band-gap voltage; $eV > E_g$ as shown in Figure 4.32 (b). The separation between E_{Fn} and E_{Fp} is now the applied potential energy or eV . The applied voltage diminishes the built-in potential barrier to almost zero, which means that electrons flow into the SCL and flow over to the p^+ -side to constitute the diode current. There is a similar reduction in the potential barrier for holes from p^+ - to n^+ -side. The final result is that electrons from n^+ -side and holes from p^+ -side flow into the SCL, and this SCL region is no longer depleted, as apparent in Figure 4.32 (b). If we draw the energy band diagram with $E_{Fn} - E_{Fp} = eV > E_g$ this conclusion is apparent. In this region, there are *more* electrons in the conduction band at energies near E_c than electrons in the valence band near E_v as illustrated by density of states diagram for the junction region in Figure 4.33 (a). In other words, there is a **population inversion** between energies near E_c and those near E_v around the junction.

This population inversion region is a layer along the junction and is called the **inversion layer** or the **active region**. An incoming photon with an energy of $(E_c - E_v)$ cannot excite an

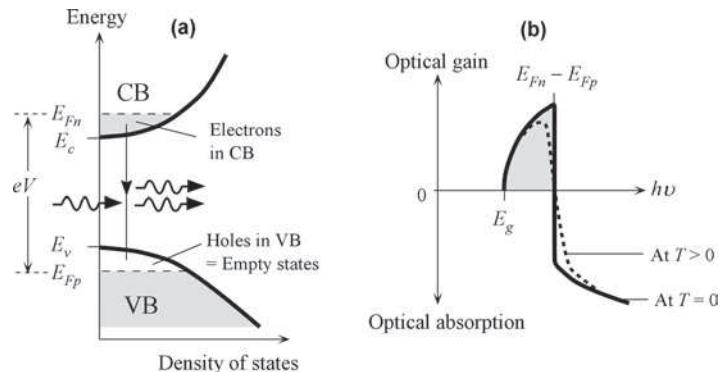


FIGURE 4.33 (a) The density of states and energy distribution of electrons and holes in the conduction and valence bands, respectively, in the SCL under forward bias such that $E_{Fn} - E_{Fp} > E_g$. Holes in the VB are empty states. (b) Gain vs. photon energy ($h\nu$). The effect of temperature is also shown.

electron from E_v to E_c as there are almost none near E_v . It can, however, *stimulate* an electron to fall down from E_c to E_v as shown in Figure 4.32 (b). Put differently, the incoming photon stimulates direct recombination. The region where there is population inversion and hence more stimulated emission than absorption, or the active region, has an **optical gain** because an incoming photon is more likely to cause stimulated emission than being absorbed. The optical gain depends on the photon energy (and hence on the wavelength) as apparent by the energy distributions of electrons and holes in the conduction and valence bands in the active layer in Figure 4.33 (a). At low temperatures ($T \approx 0$ K), the states between E_c and E_{Fn} are filled with electrons and those between E_{Fp} and E_v are empty. Photons with energy greater than E_g but less than $E_{Fn} - E_{Fp}$ cause stimulated emissions, whereas those photons with energies greater than $E_{Fn} - E_{Fp}$ become absorbed. Figure 4.33 (b) shows the expected dependence of optical gain and absorption on the photon energy at low temperatures ($T \approx 0$ K). As the temperature increases, the Fermi–Dirac function spreads the energy distribution of electrons in the CB to above E_{Fn} and holes below E_{Fp} in the VB. Put differently, electrons from below E_{Fn} are spread to energies above E_{Fn} . The result is a reduction and also some smearing of the optical gain curve as indicated in Figure 4.33 (b). The optical gain depends on $E_{Fn} - E_{Fp}$, which depends on the applied voltage, and hence on the diode current.

It is apparent that population inversion between energies near E_c and those near E_v is achieved by the injection of carriers across the junction under a sufficiently large forward bias. The pumping mechanism is therefore the forward diode current and the pumping energy is supplied by the external battery. This type of pumping is called **injection pumping**.

In addition to population inversion we also need to have an *optical cavity* to implement a laser oscillator, that is, to build up the intensity of stimulated emissions by means of an optical resonator. This would provide a continuous coherent radiation as output from the device. Figure 4.34 shows schematically the structure of a **homojunction laser diode**. The *pn* junction uses the same direct bandgap semiconductor material throughout, for example GaAs, and hence has the name homojunction. The ends of the crystal are cleaved to be flat and optically polished to provide reflection and hence form an optical cavity. Photons that are reflected from the cleaved surfaces stimulate more photons of the same frequency and so on. This process builds up the intensity of the radiation in the cavity. The wavelength of the radiation that can build up in the cavity is determined by the length L of the cavity because

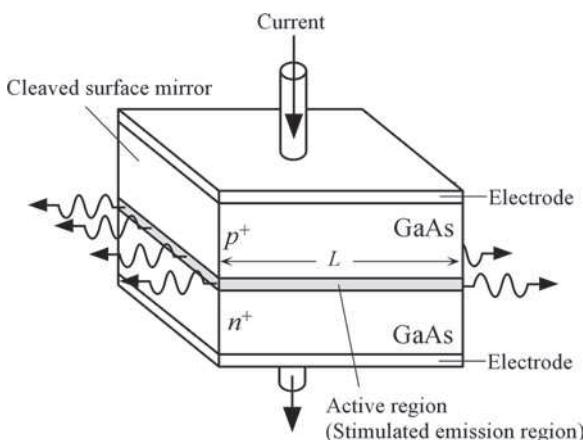
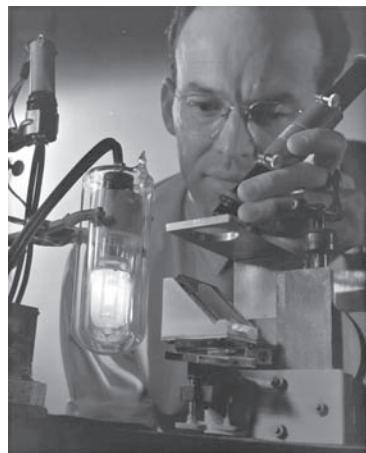


FIGURE 4.34 A schematic illustration of a GaAs homojunction laser diode. The cleaved surfaces act as reflecting mirrors.

Robert Hall and his colleagues, while working at General Electric's Research and Development Center in Schenectady, New York, were among the first groups of researchers to report a working semiconductor laser diode in 1962. He obtained a U.S. patent in 1967, entitled "Semiconductor junction laser diode" for his invention. When Robert Hall retired from GE in 1987, he had been awarded more than 40 patents. (R. N. Hall et al., *Phys. Rev. Letts.*, 9, 366, 1962; Courtesy of GE.)



only multiples of the half-wavelengths can exist in such an optical cavity as explained in Section 4.6D, *i.e.*,

*Modes in
an optical
cavity*

$$m \frac{\lambda}{2n} = L \quad (4.9.1)$$

where, as before, m is an integer, called a **mode number**, n is the refractive index of the semiconductor, and λ is the free-space wavelength. Each radiation pattern satisfying the above relationship is essentially a **resonant frequency** of the cavity, that is, a **mode** of the cavity. The separation between possible modes of the cavity (or separation between allowed wavelengths) $\Delta\lambda_m$ can be readily found from Eq. (4.9.1) as in the case of the He-Ne gas laser previously.

The dependence of the optical gain of the medium on the wavelength of radiation can be deduced from the energy distribution of the electrons in the CB and holes in the VB around the junction as in Figures 4.33 (a) and (b). The exact output spectrum from the laser diode depends both on the nature of the optical cavity and the optical gain vs. wavelength characteristics. Lasing radiation is only obtained when the optical gain in the medium can overcome the photon losses from the cavity, which requires the diode current I to exceed a certain threshold value I_{th} . (This is similar to the condition we had for gas lasers where pumping must exceed some threshold pumping required for the threshold gain.) Below I_{th} , the light from the device is due to spontaneous emission and not stimulated emission. The light output is then composed of incoherent photons that are emitted randomly and the device behaves like an LED. For diode currents above I_{th} , the device emits coherent lasing emission.

We can identify two critical diode currents. First is the diode current that provides just sufficient injection to lead to stimulated emissions just balancing absorption. This is called the **transparency current** I_T inasmuch as there is then no net photon absorption; the medium is perfectly transparent. Above I_T there is optical gain in the medium though the optical output is not yet a continuous wave coherent radiation. Lasing oscillations occur only when the optical gain in the medium can overcome the photon losses from the cavity, that is, when the optical gain g reaches the threshold gain g_{th} . This occurs at the **threshold current** I_{th} . Those cavity resonant frequencies that experience the threshold optical gain can resonate within the cavity. Some of this cavity radiation is transmitted out from the cleaved ends as these are not perfectly reflecting (typically about

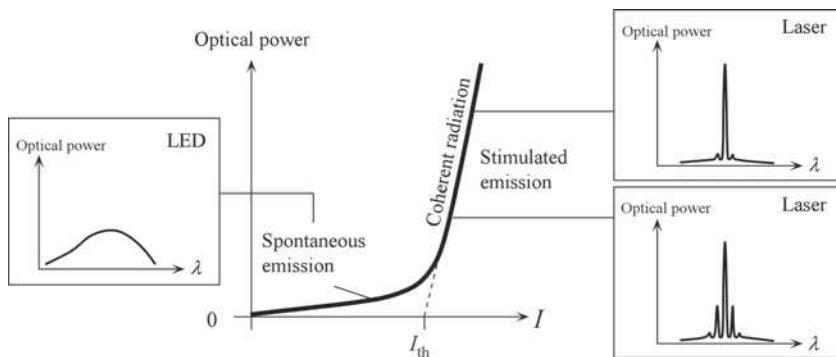


FIGURE 4.35 Typical output optical power vs. diode current (I) characteristics and the corresponding output spectrum of a laser diode. I_{th} is the threshold current and corresponds to the extension of the coherent radiation output characteristic onto the I -axis.

32% reflecting without any antireflection coating). Figure 4.35 shows the output light intensity as a function of diode current. Below the threshold current I_{th} , the light output is incoherent radiation due to the spontaneous recombination of injected electrons and holes. Above I_{th} , the light output becomes coherent radiation consisting of cavity wavelengths (or modes); and the output intensity increases steeply with the current. The number of modes in the output spectrum and their relative strengths depend on the diode current as illustrated in Figure 4.35. Notice that as the current increases, one of the intense modes becomes prominent at the expense of less intense modes. The prominent mode has so much intensity that it ends up using most of the injected electrons; there are insufficient electrons left for other modes to satisfy the threshold gain condition at those frequencies.

The main problem with the homojunction laser diode is that the threshold current density J_{th} is too high for practical uses. For example, the threshold current density is of the order of $\sim 500 \text{ A mm}^{-2}$ for GaAs pn junctions at room temperature, which means that the GaAs homojunction laser can be operated continuously only at very low temperatures. However, J_{th} can be reduced by orders of magnitude by using **heterostructure** semiconductor laser diodes.

4.10 HETEROSTRUCTURE LASER DIODES

The reduction of the threshold current I_{th} to a practical value requires improving the rate of stimulated emission and also improving the efficiency of the optical cavity. First, we can confine the injected electrons and holes to a narrow region around the junction. This narrowing of the active region means that less current is needed to establish the necessary concentration of carriers for population inversion. Second, we can build a dielectric waveguide around the optical gain region to increase the photon concentration and hence the probability of stimulated emission. This way we can reduce the loss of photons traveling off the cavity axis. We therefore need both **carrier confinement** and **photon or optical confinement**. Both of these requirements are readily achieved in modern laser diodes by the use of heterostructured devices as in the case of high-intensity double heterostructure LEDs. However, in the case of laser diodes, there is an additional requirement for maintaining a good optical cavity that will increase stimulated emission over spontaneous emission.

Figure 4.36 (a) shows a **double heterostructure** (DH) device based on two junctions between different semiconductor materials with different bandgaps. In this case, the semiconductors are $\text{Al}_{1-x}\text{Ga}_x\text{As}$ (or simply AlGaAs) with $E_g \approx 2 \text{ eV}$ and GaAs with $E_g \approx 1.4 \text{ eV}$. The p -GaAs region is a thin layer, typically about $0.1 \mu\text{m}$, and constitutes the **active layer** in which

Izuo Hayashi (left) and Morton Panish (1971) at Bell Labs were able to design the first semiconductor laser that operated continuously at room temperature. The need for semiconductor heterostructures for efficient laser diode operation was put forward by Herbert Kroemer in the United States and Zhores Alferov in Russia in 1963. (Notice the similarity of the energy band diagram on the chalkboard with that in Figure 4.36.) (Reprinted with permission of Alcatel-Lucent USA Inc.)



lasing recombination takes place. Both p -GaAs and p -AlGaAs are heavily p -type doped and are degenerate with E_F in the valence band. When a sufficiently large forward bias is applied, E_c of n -AlGaAs moves above E_c of p -GaAs, which leads to a large injection of electrons from the CB of n -AlGaAs into the CB of p -GaAs as shown in Figure 4.36 (b). These electrons, however, are *confined* to the CB of p -GaAs since there is a potential barrier ΔE_c between p -GaAs and p -AlGaAs due to the change in the bandgap (there is also a small change in E_v but we ignore this). Inasmuch as p -GaAs is a thin layer, the concentration of injected electrons in the p -GaAs layer can be increased quickly even with moderate increases in forward current. This effectively reduces the threshold current for population inversion or optical gain. Thus, even moderate forward currents can inject sufficient number of electrons into the CB of p -GaAs to establish the necessary electron concentration for population inversion in this layer.

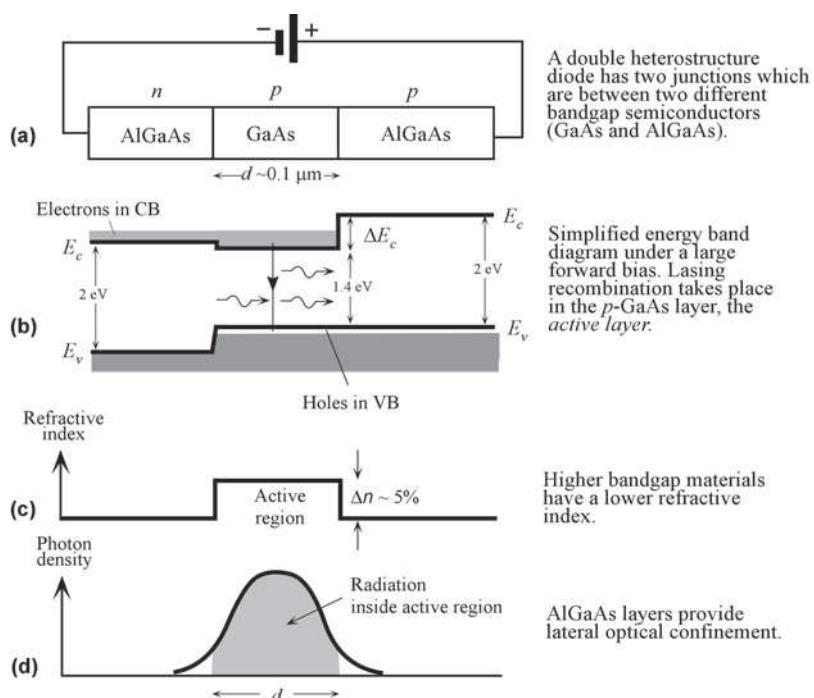


FIGURE 4.36 The basic principle of operation of a double heterostructure laser.

A wider bandgap semiconductor generally has a lower refractive index. AlGaAs has a lower refractive index than that of GaAs. The change in the refractive index defines an optical dielectric waveguide, as illustrated in Figure 4.36 (c), that confines the photons to the active region of the optical cavity and thereby increases the photon concentration. The photon concentration across the device is shown in Figure 4.36 (d). This increase in the photon concentration increases the rate of stimulated emissions. Thus, both carrier and optical confinement lead to a reduction in the threshold current density. Without double heterostructure devices we would not have practical solid state lasers that can be operated continuously at room temperature.

A typical structure of a **double heterostructure laser diode** is similar to a double heterostructure LED and is shown schematically in Figure 4.37. The doped layers are grown epitaxially on a crystalline substrate which in this case is n -GaAs. The double heterostructure described above consists of the first layer on the substrate, n -AlGaAs, the active p -GaAs layer, and the p -AlGaAs layer. There is an additional p -GaAs layer, called **contacting layer**, next to p -AlGaAs. It can be seen that the electrodes are attached to the GaAs semiconductor rather than AlGaAs. This choice allows for better contacting, that is, smaller contact resistance. The p - and n -AlGaAs layers provide carrier and optical confinement in the vertical direction by forming heterojunctions with p -GaAs. The active layer is p -GaAs, which means that the lasing emission will be in the range 870–900 nm depending on the doping level. This layer can also be made to be $Al_yGa_{1-y}As$ but of different composition than the confining $Al_xGa_{1-x}As$ layers, and still preserve heterojunction properties. By modifying the composition of the active layer, we can control the wavelength of the lasing emission over 650–900 nm. The advantage of AlGaAs/GaAs heterojunction is that there is only a small lattice mismatch between the two crystal structures and hence negligible strain induced interfacial defects (*e.g.*, dislocations) in the device. Such defects invariably act as nonradiative recombination centers and hence reduce the rate of radiative transitions.

An important feature of this laser diode is the **stripe geometry**, or stripe contact on p -GaAs. The current density J from the stripe contact is not uniform laterally. J is greatest

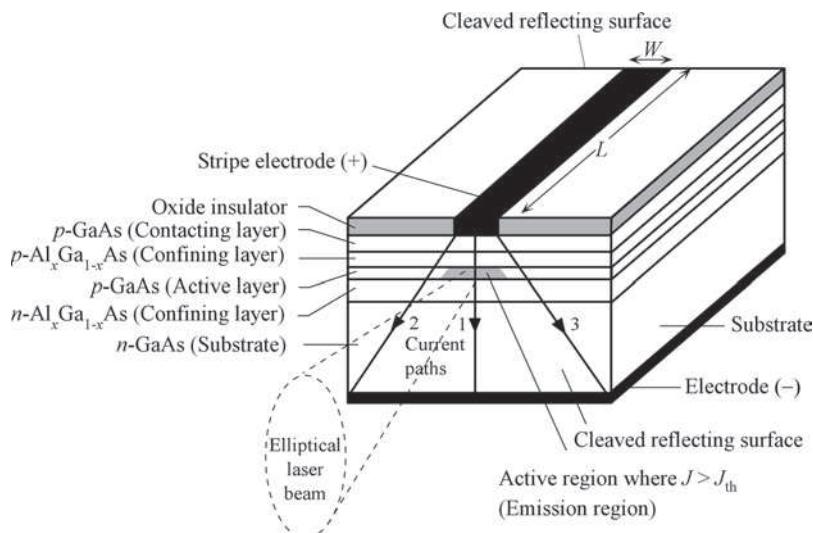


FIGURE 4.37 Schematic illustration of the structure of a double heterojunction stripe contact laser diode.

along the central path, 1, and decreases away from path 1, toward 2 or 3. The current is confined to flow within paths 2 and 3. The current density paths through the active layer where J is greater than the threshold value J_{th} , as shown in Figure 4.37, define the **active region** where population inversion and hence optical gain takes place. The lasing emission emerges from this active region. The width of the active region, or the optical gain region, is therefore defined by the current density from the stripe contact. Optical gain is highest where the current density is greatest. Such lasers are called **gain guided**. There are two advantages to using a stripe geometry. First, the reduced contact area also reduces the threshold current I_{th} . Second, the reduced emission area makes light coupling to optical fibers easier. Typical stripe widths (W) may be as small as a few microns leading to typical threshold currents that may be tens of milliamperes.

The simple double heterostructure laser diode structure in Figure 4.37 has an optical cavity defined by the end surfaces, *i.e.*, the facets, of the crystal. The lasing emission emerges from one of the crystal facets, that is, from an area on a crystal face perpendicular to the active layer. The radiation is emitted from the *edge* of the crystal. These DH devices are usually marketed as **Fabry–Perot cavity, edge emitting** laser diodes, or simply FP laser diodes, and edge emission is implied. Since the refractive index of GaAs is about 3.6, the reflectance is 0.32 (or 32%). The laser efficiency can therefore be further improved and the finesse of the FP cavity enhanced by reducing the reflection losses from the crystal facets by, for example, suitably coating the end surfaces. Further, by fabricating a dielectric mirror (as explained in Chapter 1) at the rear facet, that is, a mirror consisting of a number of quarter-wavelength semiconductor layers of different refractive index, it is possible to bring the reflectance close to unity and thereby improve the optical gain of the cavity. This corresponds to a reduction in the threshold current.

The width, or the lateral extent, of the optical gain region in the stripe geometry DH laser in Figure 4.37 is defined by the current density, and changes with the current. More importantly, the lateral optical confinement of photons to the active region is poor because there is no marked change in the refractive index laterally. It would be advantageous to laterally confine the photons to the active region to increase the rate of stimulated emissions. This can be achieved by shaping the refractive index profile in the same way the vertical confinement was defined by the heterostructure. Figure 4.38 illustrates schematically the structure of such a DH laser diode where the active layer, p -GaAs, is bound both vertically and laterally by a wider bandgap semiconductor, AlGaAs, which has lower refractive index. The active layer (GaAs) is effectively buried within a wider bandgap material (AlGaAs) and the structure is hence called a **buried double heterostructure** laser diode. Since the active layer is surrounded by a lower index material (AlGaAs), it behaves as a dielectric waveguide and ensures

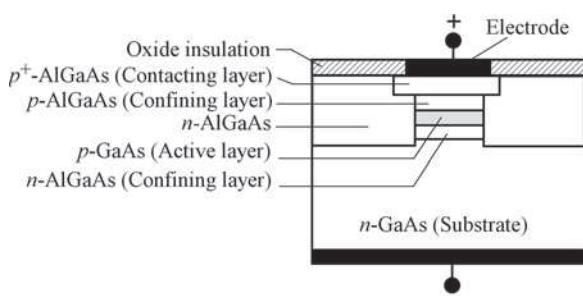


FIGURE 4.38 A simplified schematic diagram of a double heterostructure semiconductor laser device that has its active region (p -GaAs) buried within the device in such a way that it is surrounded by a low refractive index material (AlGaAs) rendering the active region as a waveguide.

that the photons are confined to the active or optical gain region. The photon confinement increases the rate of stimulated emission and hence the efficiency of the diode. Inasmuch as the optical power is confined to the waveguide defined by the refractive index variation, these diodes are called **index guided**. Further, if the buried heterostructure has the right lateral dimensions compared with the wavelength of the radiation, then only the **fundamental lateral mode** can exist in this waveguide structure as in the case of dielectric waveguides. There may, however, be several longitudinal modes.

A laser diode in which the lasing emission occurs only in one mode, both longitudinally and laterally, is called a **single-mode laser diode**. Lateral confinement, and hence a single lateral mode (*i.e.*, TE₀₀) can be easily obtained in buried DH laser diodes by ensuring that the buried active layer has a sufficiently small lateral size, *e.g.*, a few microns. In the stripe geometry LDs, a single lateral mode can be obtained by keeping the width of the stripe (*W*) narrow in Figure 4.37. On the other hand, operation at a single longitudinal mode, as shown in Example 4.10.1, typically needs the cavity length to be sufficiently short to allow only one mode to exit within the optical gain bandwidth. There are several techniques for generating single-mode lasers. One commonly used technique is to restrict the allowed wavelengths in the optical cavity by using wavelength-selective reflectors, as discussed later in Section 4.14.

The laser diode heterostructures based on Al_{1-x}Ga_xAs alloys are suitable for emissions from about 680 nm (red) to about 900 nm (IR). For operation in the optical communication wavelengths of 1.3 μm and 1.55 μm, typical heterostructures are based on InGaAsP (In_{1-x}Ga_xAs_{1-y}P_y) alloys grown on InP substrates. InGaAsP quaternary alloys have a narrower bandgap than InP, and a greater refractive index. The composition of the InGaAsP alloy is adjusted to obtain the required bandgap for the active and confining layers. Figure 4.39 shows a highly simplified schematic structure of a buried (index guided) DH LD laser diode for use in optical communications. The active layer (InGaAsP) is surrounded by wider bandgap, lower refractive index InP. Notice that the active layer with the right InGaAsP composition is surrounded by lower refractive index, higher bandgap InP both horizontally and vertically. Layers are grown on an InP substrate. The contact covers the whole surface so that the current must be limited to the central active layer region. The latter is achieved by using InP *np* junctions around the active layer that are reverse biased and therefore prevent the current flow outside the central active region. There is a *p*-InGaAs layer between the electrode and the *p*-InP that serves as a contacting layer; that is, this layer facilitates current injection into the *p*-InP by reducing the contact resistance.

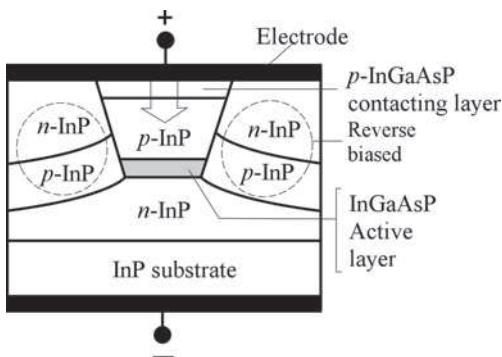
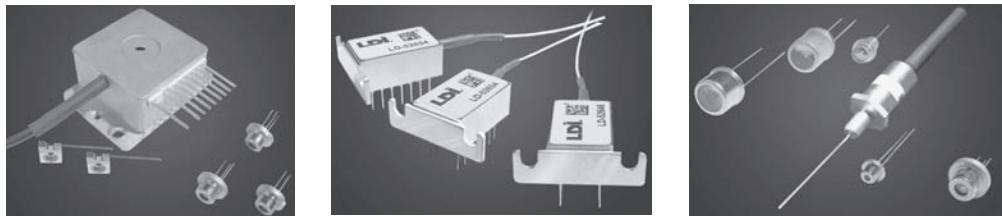


FIGURE 4.39 A highly simplified schematic sketch of a buried heterostructure laser diode for telecom applications. The active layer (InGaAsP) is surrounded by the wider bandgap, lower refractive index InP material. Layers are grown on an InP substrate. The InP *np* junction is reverse biased and prevents the current flow outside the central active region.



Left: High power (0.5–7 W) CW laser diodes with emission at 805 nm and a spectral width of 2.5 nm. Applications include medical systems, diode pumped lasers, analytical equipment, illuminators, reprographics, laser initiated ordnance, etc. Center: Typical pigtailed laser diodes for telecom. These are Fabry–Perot laser diodes operating at peak wavelengths of 1310 nm and 1550 nm with spectral widths of 2 nm and 1.3 nm, respectively. The threshold currents are 6 mA and 10 mA, and they can deliver 2 mW of optical power into a single-mode fiber. Right: High power 850 nm and 905 nm pulsed laser diodes for use in range finders, ceilometers, weapon simulation, optical fuses, surveying equipment, etc. (Courtesy of OSI Laser Diode Inc.)

EXAMPLE 4.10.1 Modes in a semiconductor laser and the optical cavity length

Consider an AlGaAs-based heterostructure laser diode that has an optical cavity of length 200 μm . The peak radiation is at 870 nm and the refractive index of GaAs is about 3.6. What is the mode integer m of the peak radiation and the separation between the modes of the cavity? If the optical gain vs. wavelength characteristics has a FWHM wavelength width of about 6 nm, how many modes are there within this bandwidth? How many modes are there if the cavity length is 20 μm ?

Solution

Figure 4.19 schematically illustrates the cavity modes, the optical gain characteristics, and a typical output spectrum from a laser. The wavelength λ of a cavity mode and length L are related by Eq. (4.9.1),

$$m \frac{\lambda}{2n} = L$$

where n is the refractive index of the semiconductor medium, so that

$$m = \frac{2nL}{\lambda} = \frac{2(3.6)(200 \times 10^{-6})}{(870 \times 10^{-9})} = 1655.1 \quad \text{or} \quad 1655 \text{ (integer)}$$

The wavelength separation $\Delta\lambda_m$ between the adjacent cavity modes m and $(m + 1)$ in Figure 4.19 is

$$\Delta\lambda_m = \frac{2nL}{m} - \frac{2nL}{m+1} \approx \frac{2nL}{m^2} = \frac{\lambda^2}{2nL}$$

where we assumed that the refractive index n does not change significantly with wavelength from one mode to another. Thus, the separation between the modes for a given peak wavelength increases with decreasing L . When $L = 200 \mu\text{m}$

$$\Delta\lambda_m = \frac{(870 \times 10^{-9})^2}{2(3.6)(200 \times 10^{-6})} = 5.26 \times 10^{-10} \text{ m} \quad \text{or} \quad 0.526 \text{ nm}$$

If the optical gain has a bandwidth of $\Delta\lambda_{1/2}$, then there will be $\Delta\lambda_{1/2}/\Delta\lambda_m$ number of modes, or $(6 \text{ nm})/(0.526 \text{ nm})$, that is, 11 modes.

When $L = 20 \mu\text{m}$, the separation between the modes becomes

$$\Delta\lambda_m = \frac{(870 \times 10^{-9})^2}{2(3.6)(20 \times 10^{-6})} = 5.26 \text{ nm}$$

Then $(\Delta\lambda_{1/2})/\Delta\lambda_m = 1.14$ and there will be one mode that corresponds to about 870 nm. In fact, m must be an integer so that choosing the nearest integer, $m = 166$, gives $\lambda = 867.5\text{ nm}$ (choosing $m = 165$ gives 872.7 nm). It is apparent that reducing the cavity length suppresses higher modes. Note that the optical bandwidth depends on the diode current.

4.11 QUANTUM WELL DEVICES

As in the case of light-emitting diodes (LEDs), quantum wells (QWs) are also widely used in modern laser diodes. A typical **quantum well-based** laser is essentially a heterostructure device, as in Figure 4.36, in which the thin GaAs layer becomes ultrathin, typically less than 20 nm. We now have an ultrathin narrow bandgap GaAs sandwiched between two wider bandgap semiconductors, which results in a QW as illustrated in Figure 4.40 (a). The confinement length d , the size of the QW, is so small that we can treat the electron as in a one-dimensional potential energy (PE) well in the x -direction and as if it were free in the yz plane. The energy of the electron in the QW is quantized along the x -direction of confinement, which is shown in Figure 4.40 (b) as E_1, E_2, \dots , corresponding to the quantum number n being 1, 2, ..., respectively. Similarly, the hole energy is also quantized as E'_1, E'_2, \dots , with a corresponding quantum number n' being 1, 2, etc., respectively. The electron is free in the yz plane (along y and z), which means that we must add this kinetic energy to the quantized energy levels. The electrons in the conduction band (CB) form a 2D (two-dimensional) free electron gas. As explained in Chapter 3,

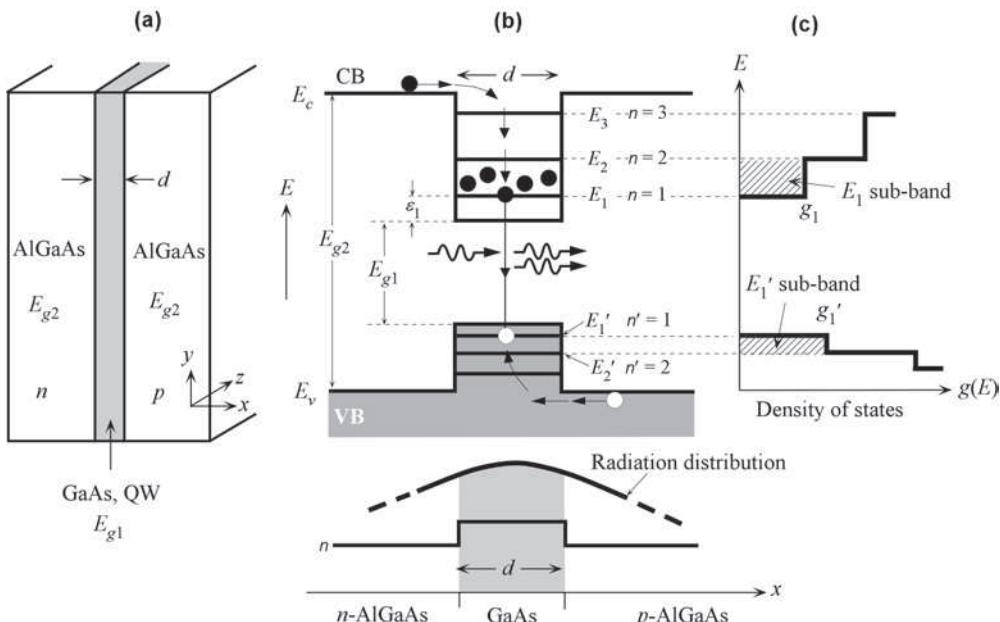


FIGURE 4.40 (a) A single quantum well of bandgap E_{g1} sandwiched between two semiconductors of wider bandgap E_{g2} . (b) The electron energy levels, and stimulated emission. The electrons and holes are injected from n -AlGaAs and p -AlGaAs, respectively. The refractive index variation tries to confine the radiation to GaAs but d is too thin, and most of the radiation is in the AlGaAs layers rather than within d . (c) The density of states $g(E)$ is a step-like function, and is finite at E_1 and E'_1 . The E_1 sub-band for electrons and E'_1 sub-band for holes are also shown. The electrons in the E_1 sub-band have kinetic energies in the yz plane.

the density of states $g(E)$ ³⁰ for a 2D electron gas is a step-like function with energy as shown in Figure 4.40 (c). Notice that the region between E_1 and E_2 has been hatched in Figure 4.40 (c) to indicate that the energies between E_1 and E_2 are due to the kinetic energy of the electron in the yz plane (and almost continuous). We can therefore identify a sub-band of energies starting at E_1 and another sub-band starting at E_2 , and so on. The density of states at E_1 , and up to E_2 , is g_1 , which is large and finite. Similarly for holes, the density of states at E'_1 is g'_1 , large and finite.

A **single quantum well** (SQW) device has only one quantum well surrounded by wider bandgap semiconductors as shown in Figures 4.40 (a) and (b) for a thin GaAs layer (thickness d) sandwiched between two thick AlGaAs layers. Usually the QW layer is undoped, whereas the AlGaAs confining layers are heavily doped. Under a sufficient forward bias, the electrons will be injected from n -AlGaAs, and holes from p -AlGaAs into the QW's CB and VB, respectively. These injected electrons very quickly thermalize and start filling states at and near E_1 . Since at E_1 there is a finite and substantial density of states (g_1), the electrons in the conduction band do not have to spread far in energy to find states. In the bulk semiconductor on the other hand, the density of states at the bottom of the band (at E_c) is zero and increases slowly with energy (as $E^{1/2}$), which means that the electrons are spread more deeply into the conduction band in search for states. Thus, in a QW, a large concentration of electrons can easily occur at E_1 , whereas this is not the case in the bulk semiconductor at E_c . Similarly, the majority of holes in the valence band will be around E'_1 since there are sufficient states at this energy. Thus, under a forward bias, the injected electrons readily populate the ample number of states at E_1 , which means that the electron concentration at E_1 increases rapidly with the current and hence, population inversion occurs quickly without the need for a large current to bring in a great number of electrons. Stimulated transitions of electrons from E_1 to E'_1 leads to a lasing emission as illustrated in Figure 4.40 (b). In practice, we should also include the kinetic energy of the electron before and after the transition, which means that the radiative transitions occur from the E_1 sub-band to the E'_1 sub-band. When the injected electron concentration is high (under a large current), E_2 and E'_2 sub-bands will also be involved.

There are two distinct advantages to a QW laser diode. First is that the threshold current for population inversion and hence lasing emission should be markedly reduced with respect to that for bulk semiconductor devices. Second, since the majority of the electrons are at or near E_1 and holes are at or near E'_1 , the range of emitted photon energies should be very close to $E_1 - E'_1$. Consequently, the optical gain curve, and hence the bandwidth, in a SQW laser diode is narrower than its bulk counterpart. We note that the allowed radiative transitions in a QW follow a **selection rule** that requires $\Delta n = n - n'$ to be zero. Thus, the transition E_1 to E'_1 is allowed, as is E_2 to E'_2 , but E_1 to E'_2 has a very low probability.

It is clear that the SQW provides a higher probability of radiative transitions due to the large densities of states at E_1 and E'_1 , and the confinement of the electrons and holes: a distinct advantage. However, the SQW can still be further improved. The radiation is not well confined to the QW region because the QW is too thin as shown in Figure 4.40 (b); the radiation spreads out into the neighboring layers in a similar fashion to the spread of radiation into the cladding in an optical fiber that has a very thin core (a very small V -number). The optical confinement can be improved by having **optical confinement layers**, or **cladding layers**, surrounding the QW. The second problem is that at high currents, the QW can be flooded with electrons and lose its function as a QW.

The above two problems with the SQW are easily overcome by using **multiple quantum wells** (MQWs) as shown Figure 4.41. In MQW lasers, the structure has alternating

³⁰Since g (san serif) is used for optical gain, g (times roman) is used for the density of states function in this chapter.

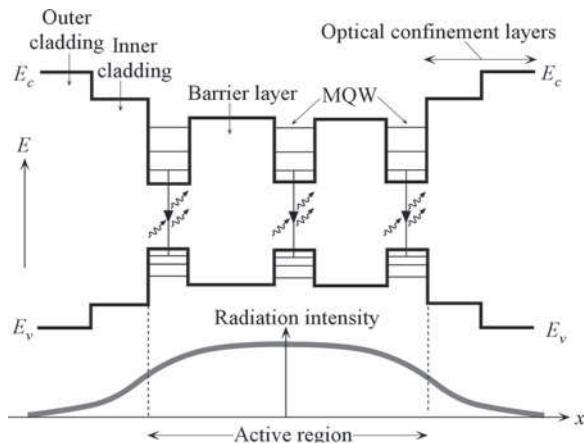


FIGURE 4.41 A simplified schematic diagram of multiple quantum well heterostructure laser diode. Electrons are injected by the forward current into quantum wells. The light intensity distribution is also shown. Most of the light is in the active region.

ultrathin layers of wide and narrow bandgap semiconductors as schematically sketched in Figure 4.41. The smaller bandgap layers are the individual QWs where electron confinement and lasing transitions take place, whereas the wider bandgap layers are the **barrier layers**. The active region now has multiple quantum wells. The active layer is cladded by two optical confinement layers, labeled as inner and outer cladding layers, with wider bandgaps, hence lower refractive indices. Most of the radiation is now within the active region containing the QWs as shown in Figure 4.41, which is the desired goal for increasing the stimulated emission rate. QW flooding is eliminated because the injected electrons are shared by the QWs. The only drawback is that the design must ensure that all the QWs are injected with sufficient electrons to exhibit optical gain.

The green, blue, and violet laser diodes in the market are based on MQWs using $\text{In}_x\text{Ga}_{1-x}\text{N}$ (E_{g1}) and GaN (E_{g2}) thin semiconductor layers. The overall LD structure is similar to the blue LEDs described in Chapter 3 [see Figure 3.39 (c)], but there is an optical cavity in the present case to ensure lasing operation.

As in the case of LEDs, there are major advantages to incorporating QWs into the laser diode structure. First, the threshold current can be lowered. Second is the overall higher efficiency of the device. Third, the optical gain curve is narrower, which should allow single-mode operation to be achieved more readily. Although the optical gain curve is narrower than the corresponding bulk device, the output spectrum from a quantum well device is not necessarily a single mode. The number of modes depends on the individual widths of the quantum wells. It is, of course, possible to combine an MQW design with an optical cavity using dielectric mirrors (*i.e.*, wavelength-selective reflectors) to generate only one mode. Many commercially available LDs are currently MQW devices.

EXAMPLE 4.11.1 A GaAs quantum well

Consider a very thin GaAs quantum well sandwiched between two wider bandgap semiconductor layers of AlGaAs ($\text{Al}_{0.33}\text{Ga}_{0.67}\text{As}$ in the present case). The QW depths from E_c and E_v are approximately 0.28 eV and 0.16 eV, respectively. Effective mass m_e^* of a conduction electron in GaAs is approximately $0.07m_e$, where m_e is the electron mass in vacuum. Calculate the first two electron energy levels for a quantum well of thickness 10 nm. What is the hole energy in the QW above E_v of GaAs if the hole

effective mass $m_h^* \approx 0.50m_e$? What is the change in the emission wavelength with respect to bulk GaAs, for which $E_g = 1.42$ eV? Assume infinite QW depths for the calculations.

Solution

As we saw in Chapter 3 (Section 3.12), the electron energy levels in the QW are with respect to the CB edge E_c in GaAs. Suppose that ε_n is the electron energy with respect to E_c in GaAs, or $\varepsilon_n = E_n - E_c$ in Figure 4.40 (b). Then, the energy of an electron in a one-dimensional infinite potential energy well is

$$\varepsilon_n = \frac{h^2 n^2}{8m_e^* d^2} = \frac{(6.626 \times 10^{-34})^2 (1)^2}{8(0.07 \times 9.1 \times 10^{-31})(10 \times 10^{-9})^2} = 8.62 \times 10^{-21} \text{ J} \quad \text{or} \quad 0.0538 \text{ eV}$$

where n is a quantum number, 1, 2, ..., and we have used $d = 10 \times 10^{-9}$ m, $m_e^* = 0.07m_e$, and $n = 1$ to find $\varepsilon_1 = 0.054$ eV. The next level from the same calculation with $n = 2$ is $\varepsilon_2 = 0.215$ eV.

The hole energy levels below E_v in Figure 4.40 (b) are given by

$$\varepsilon_{n'} = \frac{h^2 n'^2}{8m_h^* d^2}$$

where n' is the quantum number for the hole energy levels above E_v . Using $d = 10 \times 10^{-9}$ m, $m_h^* \approx 0.5m_e$, and $n' = 1$, we find $\varepsilon'_1 = 0.0075$ eV.

The wavelength of emission from bulk GaAs with $E_g = 1.42$ eV is

$$\lambda_g = \frac{hc}{E_g} = \frac{(6.626 \times 10^{-34})(3 \times 10^8)}{(1.42)(1.602 \times 10^{-19})} = 874 \times 10^{-9} \text{ m (874 nm)}$$

In the case of QWs, we must obey the selection rule that the radiative transition must have $\Delta n = n' - n = 0$. Thus, the radiative transition is from ε_1 to ε'_1 so that the emitted wavelength is

$$\begin{aligned} \lambda_{\text{QW}} &= \frac{hc}{E_g + \varepsilon_1 + \varepsilon'_1} = \frac{(6.626 \times 10^{-34})(3 \times 10^8)}{(1.42 + 0.0538 + 0.0075)(1.602 \times 10^{-19})} \\ &= 838 \times 10^{-9} \text{ m (838 nm)} \end{aligned}$$

The difference is $\lambda_g - \lambda_{\text{QW}} = 36$ nm. We note that we assumed an infinite PE well. If we actually solve the problem properly by using a finite well depth,³¹ then we would find $\varepsilon_1 \approx 0.031$ eV, $\varepsilon_2 \approx 0.121$ eV, $\varepsilon'_1 \approx 0.007$ eV. The emitted photon wavelength is 848 nm and $\lambda_g - \lambda_{\text{QW}} = 26$ nm.

4.12 ELEMENTARY LASER DIODE CHARACTERISTICS

The output spectrum from a laser diode depends on two factors: the nature of the optical resonator used to build the laser oscillations and the optical gain curve (lineshape) of the active medium. The optical resonator is essentially a Fabry–Perot cavity as illustrated in Figure 4.42 (a) which can be assigned a length L , width W , and height H . The height is the same as the layer thickness d of the active layer in the heterostructure LD as shown in Figure 4.36 (a). The length L determines the **longitudinal mode** separation, whereas the width W and height H determine the **transverse modes, or lateral modes** in LD nomenclature. If the transverse dimensions (W and H) are sufficiently small, only the lowest transverse mode, TEM_{00} mode, will exit. This TEM_{00} mode, however, will have longitudinal modes whose separation depends on L . Figure 4.42 (a)

³¹ Clear explanations and the proper calculations for QWs can be found in Mark Fox, *Optical Properties of Solids*, 2nd Edition (Oxford University Press, 2010), Ch. 6.

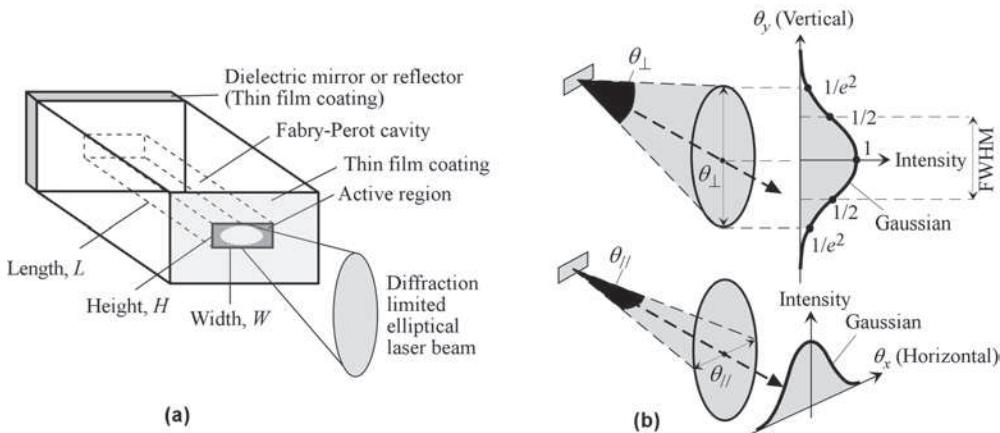


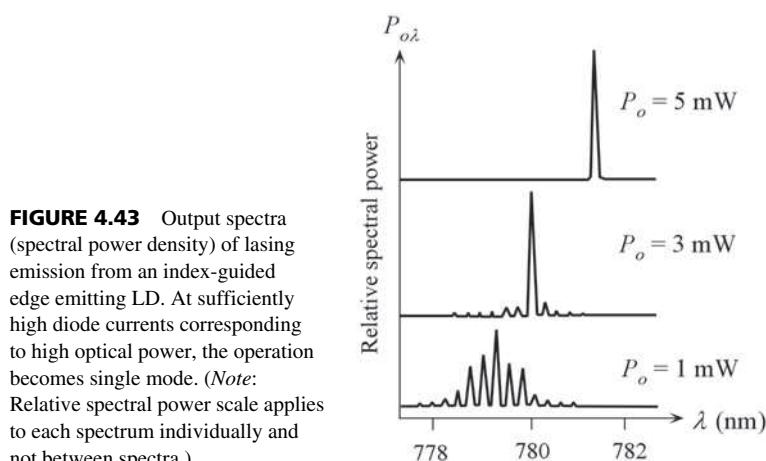
FIGURE 4.42 (a) The laser cavity definitions and the output laser beam characteristics. (b) Laser diode output beam astigmatism. The beam is elliptical, and is characterized by two angles, θ_{\perp} and θ_{\parallel} .

also shows that the emerging laser beam exhibits divergence. This is due to the diffraction of the waves at the cavity ends. The smallest aperture (W in the figure) causes the greatest diffraction.

The emerging beam from an edge emitting LD has an elliptical cross-section and exhibits astigmatism as shown in Figure 4.42 (b). The angular spread in the vertical (y) and horizontal (x) directions (perpendicular or parallel to the plane of the diode junction) are not the same which leads to a nearly elliptical beam cross-section. Typically, the active layer cavity height H is smaller than the cavity width W . Since the angular spread of the emerging laser beam is diffraction controlled by the emission aperture, smaller the diffracting height H , the wider is the diffracted output beam in the vertical direction. The vertical angular spread θ_{\perp} is more than the horizontal spread θ_{\parallel} as illustrated in Figure 4.42 (b).

The far field intensity distributions (far from the crystal facet) in the y and x directions are nearly Gaussian in the angular displacements θ_y and θ_x . The spreads θ_{\perp} and θ_{\parallel} refer to the angular separation of the e^{-2} points on the Gaussian curve as indicated in Figure 4.42 (b). In some specifications, θ_{\perp} and θ_{\parallel} refer to the angular separation of half-intensity points (full width at half maximum, FWHM); the two are related by $\theta(e^{-2}) = 1.7\theta$ (FWHM). The output divergence is normally quoted as $\theta_{\perp} \times \theta_{\parallel}$. The output beam from the LD in most applications is either put through a lens-prism system to collimate the beam, or it is coupled, using a suitable (an anamorphic) lens, into a fiber. The manipulation of the LD output beam through optics into a well-defined collimated beam with a circular cross-section ($\theta_{\perp} = \theta_{\parallel}$) is usually referred to as the **collimation** and **circularization** of the output beam. (Circularization should not be confused with circularly polarized light.) The collimation and circularization can be quite complicated depending on the exact requirements and space available for the optics.

The actual modes that exist in the output spectrum of a LD will depend on the optical gain these modes will experience. The term output spectrum strictly refers to the spectral power vs. λ characteristic of the emitted radiation. The **spectral power** $P_{o\lambda}$ is defined as $dP_o/d\lambda$, that is, optical power emitted per unit wavelength. A typical example is shown in Figure 4.43 for an index guided LD. The area under the $P_{o\lambda}$ vs. λ spectrum curve represents the total power P_o emitted. The spectrum, is either multimode or single mode depending on the optical resonator structure and the pumping current level. As apparent from Figure 4.43, there is a shift in the peak emission wavelength with the current. In most LDs, the shift in the peak wavelength is due to the Joule heating of the semiconductor



at high currents, which changes the properties of the semiconductor and the laser cavity. High stability LDs have a thermoelectric cooler to maintain the temperature at the desired level. The multimode spectrum at low output power typically becomes single mode at high output powers. The shift in the peak wavelength is partly due to the Joule heating of the semiconductor at high currents. In contrast, the output spectrum of most gain-guided LDs tends to remain multimode even at high diode currents. For many applications such as optical communications, the LD has to be designed so that the emission is single mode and the shift in the emission wavelength with increasing output power is small.

The LD output characteristics are temperature sensitive. Figure 4.44 shows the changes in the optical output power vs. diode current characteristics with the case temperature. As the temperature increases, the threshold current increases steeply, typically as the exponential of the absolute temperature, that is,

$$I_{\text{th}} = A \exp(T/T_o) \quad (4.12.1)$$

Threshold current and temperature

where A and T_o are constants that have the units of current (A) and temperature (K), respectively; T_o is sometimes called the characteristic temperature for the I_{th} vs. T dependence.

The output spectrum also changes with the temperature. In the case of a single-mode LD, the peak emission wavelength λ_o exhibits jumps at certain temperatures as apparent in Figures 4.45 (a) and (b). A jump corresponds to a **mode hop** in the output. That is, at the new

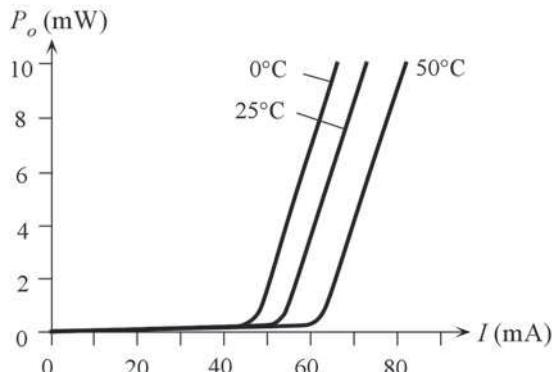


FIGURE 4.44 Output optical power vs. diode current at three different temperatures. The threshold current shifts to higher temperatures.

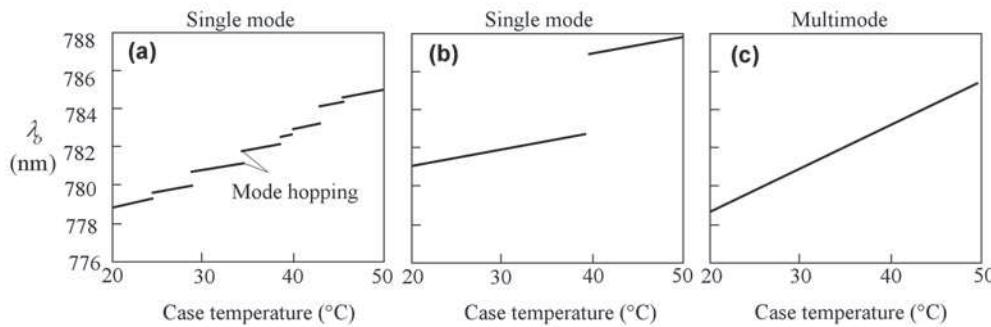


FIGURE 4.45 Peak wavelength λ_o vs. case temperature characteristics. (a) Mode hops in the output spectrum of a single-mode LD. (b) Restricted mode hops and none over the temperature range of interest ($20\text{--}40^\circ\text{C}$). (c) Output spectrum from a multimode LD.

operating temperature, another mode fulfills the laser oscillation condition, which means a discrete change in the laser oscillation wavelength. Between mode hops, λ_o increases slowly with the temperature due to the slight increase in the refractive index n (and also the cavity length) with temperature. If mode hops are undesirable, then the device structure must be such to keep the modes sufficiently separated. In contrast, the output spectrum of a gain-guided laser has many modes, so that λ_o vs. T behavior tends to follow the changes in the bandgap (the optical gain curve) rather than the cavity properties as illustrated in Figure 4.45 (c). Highly stabilized LDs are usually marketed with thermoelectric coolers integrated into the diode package to control the device temperature. Many high power laser diodes have a photodetector within the packaging to monitor the output intensity. The output from this photodetector is called the **monitoring current**, whose value (a fraction of a milliamp) is specified by the manufacturer for a given output optical power from the laser diode.

One commonly stated important and useful laser diode parameter is the slope efficiency, which determines the optical power P_o in the lasing emission in terms of the diode current I above the threshold current I_{th} . The **slope efficiency** (SE) η_{slope} is the slope of P_o vs. I above threshold

$$\eta_{\text{slope}} = \left(\frac{\Delta P_o}{\Delta I} \right)_{\text{above threshold}} \approx \frac{P_o}{I - I_{\text{th}}} \quad (4.12.2)$$

Slope efficiency

and is measured in W A^{-1} or mW mA^{-1} . The slope efficiency depends on the LD structure as well as the semiconductor packaging, and typical values for commonly available LDs are close to 1 W/A . There are several laser diode efficiency definitions as follows:

The **external quantum efficiency** η_{EQE} of a laser diode is defined as

$$\eta_{\text{EQE}} = \frac{\text{Number of output photons from the diode per second}}{\text{Number of injected electrons into the diode per second}} = \frac{P_o/h\nu}{I/e} \approx \frac{eP_o}{E_g I} \quad (4.12.3)$$

External quantum efficiency

The term *differential* is normally used to identify efficiencies related to changes above the threshold. The **external differential quantum efficiency** η_{EDQE} is defined for operation above threshold as

$$\eta_{\text{EDQE}} = \frac{\text{Increase in number of output photons from diode per second}}{\text{Increase in number of injected electrons into diode per second}} \quad (4.12.4a)$$

External differential quantum efficiency

*External
differential
quantum
efficiency*

$$\eta_{\text{EDQE}} = \frac{\Delta P_o/h\nu}{\Delta I/e} = \eta_{\text{slope}} \frac{e}{h\nu} \approx \left(\frac{e}{E_g} \right) \frac{P_o}{I - I_{\text{th}}} \quad (4.12.4b)$$

The **internal quantum efficiency** (IQE) η_{IQE} , as in the case of LEDs, measures what fraction of injected electrons recombines radiatively. If τ_r is the radiative recombination time and τ_{nr} is the nonradiative recombination time, then

*Internal
quantum
efficiency*

$$\eta_{\text{IQE}} = \frac{\text{Rate of radiative recombination}}{\text{Rate of all recombination processes}} = \frac{1/\tau_r}{1/\tau_r + 1/\tau_{nr}} \quad (4.12.5)$$

Equivalently, we could have defined IQE as the number of generated photons divided by the number of injected electrons.

The **internal differential quantum efficiency** (IDQE) represents the increase in the number of photons generated inside the gain medium per unit increase in the injected electrons, *i.e.*, the current, above threshold. Thus, *above threshold*,

*Internal
differential
quantum
efficiency*

$$\eta_{\text{IDQE}} = \frac{\text{Increase in number of photons generated internally per second}}{\text{Increase in number of injected electrons into diode per second}} \quad (4.12.6)$$

Put differently, if the current increases by ΔI above threshold, increase in the injected electrons is $\Delta I/e$. The increase in the number of photons generated *internally* is then $\eta_{\text{IDQE}} \times \Delta I/e$. However, these photons have to escape the cavity into the output for which we would need the **extraction efficiency** (EE) η_{EE} .

Suppose that we are interested in the output from one of the mirrors, say R_1 . The *loss* at R_1 is the transmitted output. If there were no losses at the other mirror ($R_2 = 1$) and no internal losses ($\alpha_s = 0$), then the radiation in the cavity would eventually be coupled out from R_1 and the EE efficiency η_{EE} would be unity. Thus, what is important is the relative ratio of the loss due to R_1 to all losses in the cavity. The **extraction efficiency** is defined as the ratio of loss coefficient due to R_1 to that due to total losses, α_t ,

$$\text{Extraction efficiency} = \frac{\text{Loss from the exit cavity end}}{\text{Total loss}} \quad (4.12.7)$$

i.e.,

*Extraction
efficiency*

$$\eta_{\text{EE}} = (1/2L) \ln(1/R_1)/\alpha_t$$

where $\alpha_t = \alpha_s + (1/2L) \ln(1/R_1) + (1/2L) \ln(1/R_2)$ is the total loss.

The **power conversion efficiency**, η_{PCE} , or the **external power efficiency**, gauges the overall efficiency of the conversion from the input of electrical power to the output of optical power, *i.e.*,

*External
power
efficiency*

$$\eta_{\text{PCE}} = \frac{\text{Optical output power}}{\text{Electrical input power}} = \frac{P_o}{IV} \approx \eta_{\text{EQE}} \left(\frac{E_g}{eV} \right) \quad (4.12.8)$$

Although this is not generally quoted in data sheets, it can be easily determined from the output power at the operating diode current and voltage. In some modern LDs this may be as high as 30%.

The expressions on the right-hand side of Eqs. (4.12.2) to (4.12.4) and (4.12.8) give the corresponding formulas in terms of P_o , I , V , and E_g for each definition and are approximate due to the assumptions used in deriving these equations. It is left as an exercise to show that these equations follow directly from the definitions. (Assume $h\nu \approx E_g$.)

Table 4.4 shows the main characteristics of a few commercial laser diodes emitting in the red (639–670 nm) and violet (405 nm) from low to high output powers. All are MQW laser diodes. The violet emission uses InGaN/GaN MQWs, whereas the red LDs are based on AlGaInP/GaInP alloys. Higher output power diodes have larger geometries and require larger threshold currents as can be seen from the table. Power conversion efficiencies of the LDs in the table are in the range 6–30%, which are typical values. Figure 4.46 shows typical values for the threshold current I_{th} , slope efficiency, and power conversion efficiency for a selection of commercial red LDs (36) with different optical output powers from 3 mW to 500 mW. The figure also highlights that there can be significant variations in I_{th} even at a given power level between different LDs, as can be seen for the 10-mW output. The trend in higher threshold current for higher output power, however, is quite clear. The η_{slope} values in Table 4.4 range from 0.9 to 1.7, which are typical for commercial LDs.

TABLE 4.4 Typical characteristics for a few selected red and violet commercial laser diodes. All LDs are MQW structures and have FP cavities

LD	P_o (mW)	λ (nm)	I_{th} (mA)	I (mA)	V (V)	θ_\perp	θ_\parallel	η_{slope} (W A $^{-1}$)	η_{PCE} (%)
Red	500	670	400	700	2.4	21°	10°	1.0	30
Red	100	660	75	180	2.5	18°	9°	1.0	22
Red	50	660	60	115	2.3	17°	10°	0.90	19
Red	10	639	30	40	2.3	21°	8°	1.0	11
Violet	400	405	160	390	5.0	45°	15°	1.7	21
Violet	120	405	45	120	5.0	17°	8°	1.6	20
Violet	10	405	26	35	4.8	19°	8.5°	1.1	6

Note: Violet lasers are based on InGaN/GaN MQW, and red LEDs use mainly AlGaInP/GaInP MQW.

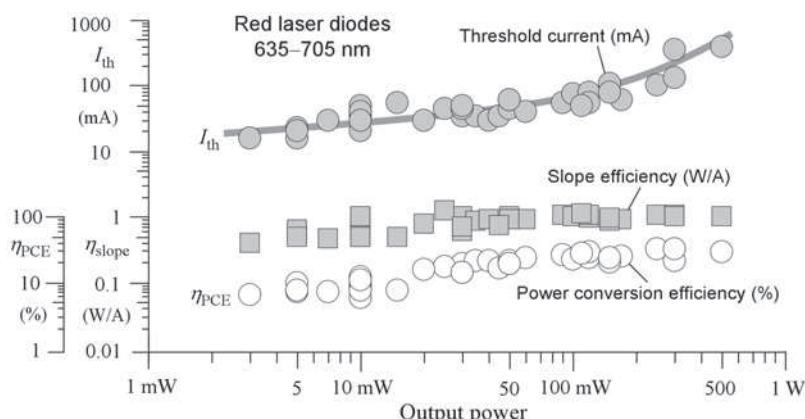
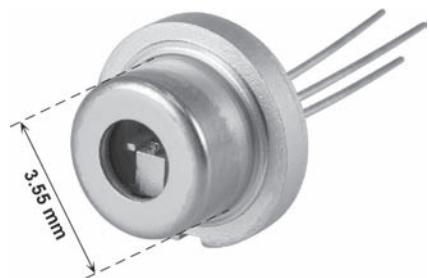


FIGURE 4.46 Typical values for the threshold current I_{th} , slope efficiency (η_{slope}), and power conversion efficiency (η_{PCE}) for 36 commercial red LDs with different optical output powers from 3 mW to 500 mW.

A red emitting (642 nm) high power laser diode that can provide CW laser beam at 150 mW. The threshold current is 110 mA, operating current and voltage are 280 mA and 2.6 V respectively, for 150 mW output power. The device structure is based on AlGaInP MQWs. (*Opnext-Hitachi laser diode. Courtesy of Thorlabs.*)



EXAMPLE 4.12.1 Laser output wavelength variation with temperature

The refractive index n of GaAs is approximately 3.6 and it has a temperature dependence $d\eta/dT \approx 2.0 \times 10^{-4} \text{ K}^{-1}$. Estimate the change in the emitted wavelength at around 870 nm per degree change in the temperature for a given mode.

Solution

Consider a particular given mode with wavelength λ_m ,

$$m\left(\frac{\lambda_m}{2n}\right) = L$$

If we differentiate λ_m with respect to temperature,

$$\frac{d\lambda_m}{dT} = \frac{d}{dT} \left[\frac{2}{m} nL \right] \approx \frac{2L}{m} \frac{dn}{dT}$$

where we neglected the change in the cavity length with temperature. Substituting for L/m in terms of λ_m ,

$$\frac{d\lambda_m}{dT} = \frac{\lambda_m}{n} \frac{dn}{dT} = \frac{870 \text{ nm}}{3.6} (2 \times 10^{-4} \text{ K}^{-1}) = 0.048 \text{ nm K}^{-1}$$

Note that we have used n for a passive cavity, whereas n above should be the effective refractive index of the *active* cavity which will also depend on the optical gain of the medium, and hence its temperature dependence is likely to be somewhat higher than the dn/dT value we used. It is left as an exercise to show that the changes in λ_m due to the expansion of the cavity length with temperature is much less than that arising from dn/dT . The linear expansion coefficient of GaAs is $6 \times 10^{-6} \text{ K}^{-1}$.

EXAMPLE 4.12.2 Laser diode efficiencies for a sky-blue LD

Consider a 60-mW blue LD (Nichia SkyBlue NDS4113), emitting at a peak wavelength of 488 nm. The threshold current is 30 mA. At a forward current of 100 mA and a voltage of 5.6 V, the output power is 60 mW. Find the slope efficiency, PCE, EQE, and EDQE.

Solution

From the definition in Eq. (4.12.2),

$$\eta_{\text{slope}} = P_o/(I - I_{\text{th}}) = (60 \text{ mW})/(100 - 30 \text{ mA}) = 0.86 \text{ mW mA}^{-1}$$

From Eq. (4.12.8), PCE is

$$\eta_{\text{PCE}} = P_o/IV = (60 \text{ mW}) / [(100 \text{ mA})(5.6 \text{ V})] = 0.11 \text{ or } 11\%$$

We can find the EQE from Eq. (4.12.3) but we need $h\nu$, which is hc/λ . In eV,

$$h\nu(\text{eV}) = 1.24/\lambda(\mu\text{m}) = 1.24/0.488 = 2.54 \text{ eV}$$

EQE is given by Eq. (4.12.3)

$$\begin{aligned}\eta_{\text{EQE}} &= (P_o/h\nu)/(I/e) = [(60 \times 10^{-3})/(2.54 \times 1.6 \times 10^{-19})]/[(100 \times 10^{-3})/(1.6 \times 10^{-19})] \\ &= 0.24 \text{ or } 24\%\end{aligned}$$

Similarly, η_{EDQE} is given by Eq. (4.12.4b) above threshold,

$$\begin{aligned}\eta_{\text{EDQE}} &= (\Delta P_o/h\nu)/(\Delta I/e) \approx (P_o/h\nu)/[(I - I_{\text{th}})/e] \\ &= [(60 \times 10^{-3})/(2.54 \times 1.6 \times 10^{-19})]/[(100 \times 10^{-3} - 30 \times 10^{-3})/(1.6 \times 10^{-19})] \\ &= 0.34 \text{ or } 34\%\end{aligned}$$

The EDQE is higher than the EQE because most injected electrons above I_{th} are used in stimulated recombinations. EQE gauges the total conversion efficiency from all the injected electrons brought by the current to coherent output photons. But, a portion of the current is used in pumping the gain medium.

EXAMPLE 4.12.3 Laser diode efficiencies

Consider an InGaAs FP semiconductor laser diode that emits CW radiation at 1310 nm. The cavity length (L) is 200 μm . The internal loss coefficient $\alpha_s = 20 \text{ cm}^{-1}$, $R_1 = R_3 \approx 0.33$ (cleaved ends). Assume that the internal differential quantum efficiency, IDQE, is close to 1. The threshold current is 5 mA. What is the output power P_o at $I = 20 \text{ mA}$? The forward voltage is about 1.3 V. What are the EDQE and conversion efficiency?

Solution

From the definition of IDQE in Eq. (4.12.6), the number of internal coherent photons generated per second above threshold is $\eta_{\text{IDQE}}(I - I_{\text{th}})/e$. Thus,

$$\text{Internal optical power generated} = h\nu \times \eta_{\text{IDQE}}(I - I_{\text{th}})/e$$

The extraction efficiency η_{EE} then couples a portion of this optical power into the output radiation. The output power P_o is then $\eta_{\text{EE}} \times h\nu \times \eta_{\text{IDQE}}(I - I_{\text{th}})/e$. Thus,

$$P_o = \eta_{\text{EE}} \eta_{\text{IDQE}} h\nu (I - I_{\text{th}})/e \quad (4.12.9) \quad \begin{matrix} \text{Output} \\ \text{power vs.} \\ \text{current} \end{matrix}$$

The slope efficiency from Eq. (4.12.2) is

$$\eta_{\text{slope}} = \Delta P_o/\Delta I = \eta_{\text{EE}} \eta_{\text{IDQE}} (h\nu/e) \quad (4.12.10) \quad \begin{matrix} \text{Slope} \\ \text{efficiency} \end{matrix}$$

Further, from the definition of EDQE and Eq. (4.12.9) is

$$\eta_{\text{EDQE}} = (\Delta P_o/h\nu)/(\Delta I/e) = (P_o/h\nu)/[(I - I_{\text{th}})/e] = \eta_{\text{EE}} \eta_{\text{IDQE}} \quad (4.12.11) \quad \begin{matrix} \text{External} \\ \text{differential} \\ \text{quantum} \\ \text{efficiency} \end{matrix}$$

We can now calculate the quantities needed. The total loss coefficient is

$$\alpha_t = \alpha_s + (1/2L) \ln(1/R_1 R_2) = 2000 + (2 \times 200 \times 10^{-6})^{-1} \ln(0.33 \times 0.33)^{-1} = 7543 \text{ m}^{-1}$$

The extraction efficiency is

$$\eta_{EE} = (1/2L) \ln(1/R_1)/\alpha_t = (2 \times 200 \times 10^{-6})^{-1} \ln(1/0.33)/(7543) = 0.37 \text{ or } 37\%$$

Thus, using $I = 20 \text{ mA}$ and $h\nu = hc/\lambda$ in Eq. (4.12.9),

$$P_o = (0.37)(1)[(6.62 \times 10^{-34})(3 \times 10^8)/(1310 \times 10^{-9})][(0.02 - 0.005)/(1.6 \times 10^{-19})] \\ = 5.2 \text{ mW}$$

The slope efficiency from Eq. (4.12.10) is

$$\eta_{slope} = \Delta P_o / \Delta I = (5.2 \text{ mW} - 0)/(20 \text{ mA} - 5 \text{ mA}) = 0.35 \text{ mW mA}^{-1}$$

The EDQE from Eq. (4.12.11) is

$$\eta_{EDQE} = \eta_{EE} \eta_{IDQE} = 0.37 \text{ or } 37\%$$

The power conversion efficiency $\eta_{PCE} = P_o/IV = 5.2 \text{ mW}/(20 \text{ mA} \times 1.3 \text{ V}) = 0.20 \text{ or } 20\%.$

4.13 STEADY STATE SEMICONDUCTOR RATE EQUATIONS: THE LASER DIODE EQUATION

A. Laser Diode Equation

Consider a double heterostructure ILD under forward bias as in Figures 4.47 (a) and (b). The current I injects electrons into the active region d and the radiative recombination of these electrons with holes in this layer generates the coherent radiation in the cavity. Suppose that n is the injected electron concentration and N_{ph} is the coherent photon concentration in the active layer. We have to be careful since not all the coherent radiation will be within the active region as illustrated in Figure 4.47 (c) and in Figure 4.36 (d), even if there is a waveguide formed by the refractive index profile between the active layer and the neighboring confining layers as shown in Figure 4.36 (c). A **radiation confinement factor** Γ is used to represent the fraction of the

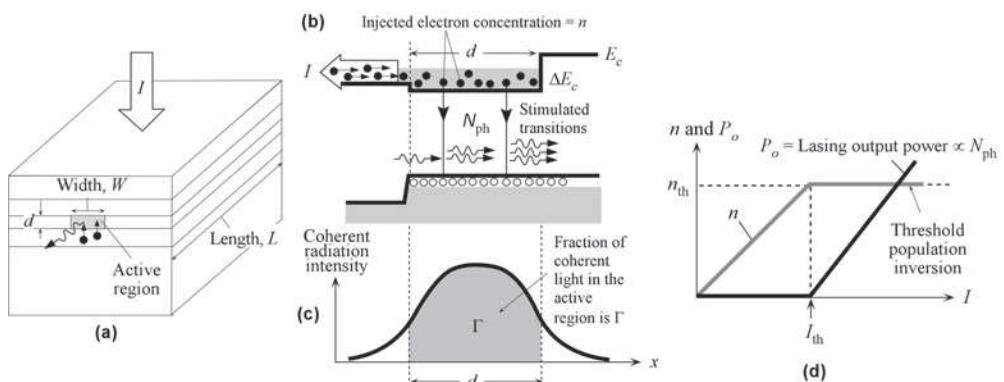


FIGURE 4.47 A highly simplified and idealized description of a semiconductor laser diode for deriving the LD equation. (a) The heterostructure laser diode structure. (b) The current I injects electrons in the conduction band, and these electrons recombine radiatively with the holes in the active region. (c) The coherent radiation intensity across the device; only a fraction Γ is within the active region where there is optical gain. (d) Injected electron concentration n and coherent radiation output power P_o vs. diode current I . The current represents the pump rate.

coherent radiation within the active layer, which is unity for perfect optical confinement. In the following we simply assume $\Gamma = 1$.

The current carries the electrons into the active layer where they recombine with holes radiatively as indicated in Figure 4.47 (b). Under steady state operation, the rate of electron injection into the active layer by the current I is equal to their rate of recombination by spontaneous emission *and* stimulated emission (neglecting nonradiative recombinations). Thus,

$$\begin{aligned}\text{Rate of electron injection by current } I &= \text{Rate of spontaneous emissions} \\ &\quad + \text{Rate of stimulated emissions}\end{aligned}$$

Thus, per unit volume, we have

$$\frac{I}{eLWd} = \frac{n}{\tau_r} + CnN_{ph} \quad (4.13.1)$$

where τ_r is the average time for spontaneous radiative recombination, simply called the **radiative lifetime**, and C is a proportionality constant for stimulated emissions (a bit like B_{21}) that, among other factors, includes the probability per unit time that a photon stimulates an electron to undergo a radiative transition. The second term on the right represents the stimulated emission rate, which depends on the concentration of electrons in the conduction band, n , and the coherent photon concentration N_{ph} in the active layer. N_{ph} includes only those coherent photons encouraged by the optical cavity, that is, a mode of the cavity. As the current increases and provides more pumping, N_{ph} increases (helped by the optical cavity), and eventually the stimulated term dominates the spontaneous term (as illustrated in Figure 4.35). The output light power P_o is proportional to N_{ph} .

Consider the coherent photon concentration N_{ph} in the cavity. Under *steady state conditions*,

$$\text{Rate of coherent photon loss in the cavity} = \text{Rate of stimulated emissions}$$

that is,

$$\frac{N_{ph}}{\tau_{ph}} = CnN_{ph} \quad (4.13.2)$$

where τ_{ph} is the *average time* for a photon to be lost from the cavity due to transmission through the end-faces, and also by absorption and scattering in the semiconductor cavity. It is the **photon cavity lifetime** as calculated in Example 4.6.2. If α_t is the total attenuation coefficient representing all these loss mechanisms, inside the cavity and also at the ends of the cavity, then the power in a light wave, in the absence of amplification, decreases as $\exp(-\alpha_t x)$, which is equivalent to a decay in time as $\exp(-t/\tau_{ph})$, where $\tau_{ph} = n/(c\alpha_t)$ and n is the refractive index.

In semiconductor laser science, the threshold electron concentration n_{th} and threshold current I_{th} refer to that condition when the stimulated emission just overcomes the spontaneous emission and the total loss mechanisms inherent in τ_{ph} . This occurs when the injected electron concentration n reaches n_{th} , the **threshold electron concentration**. From Eq. (4.13.2), this is when

$$n_{th} = \frac{1}{C\tau_{ph}} \quad (4.13.3) \quad \text{Threshold concentration}$$

This is the point when coherent radiation gain in the active layer by stimulated emission just balances all the cavity losses (represented by τ_{ph}) plus losses by spontaneous emission which

is random. When the current exceeds I_{th} , the output optical power increases sharply with the current as apparent in Figure 4.35, so we can just as well take $N_{\text{ph}} \approx 0$ when $I = I_{\text{th}}$ in Eq. (4.13.1), which gives

Threshold
current

$$I_{\text{th}} = \frac{n_{\text{th}}eLWd}{\tau_r} \quad (4.13.4)$$

Clearly the threshold current decreases with d , L , and W which explains the reasons for the heterostructure and stripe geometry lasers and the avoidance of the homojunction laser. Further, τ_{ph} can be made longer, and hence n_{th} and I_{th} smaller, by reducing the cavity losses.

When the current exceeds the threshold current, the excess carriers above n_{th} brought in by the current recombine by stimulated emission. The reason is that above threshold, the active layer has optical gain and therefore builds up coherent radiation quickly, and stimulated emission rate depends on N_{ph} . The steady state electron concentration remains constant at n_{th} , though the rates of carrier injection and stimulated recombination have increased. Above threshold, we therefore have $n = n_{\text{th}}$ in Eq. (4.13.1), and we can substitute for n_{th}/τ_r from Eq. (4.13.4)

$$\frac{I}{eLWd} = \frac{n_{\text{th}}}{\tau_r} + Cn_{\text{th}}N_{\text{ph}} \quad (4.13.5)$$

which shows that above threshold, increasing I , as expected, increases N_{ph} . We can tidy up Eq. (4.13.5) by first substituting for C from Eq. (4.13.3) and then substituting for n_{th} from Eq. (4.13.4) to find

Coherent
photon
concentration

$$N_{\text{ph}} = \frac{\tau_{\text{ph}}}{eLWd}(I - I_{\text{th}}) \quad (4.13.6)$$

To find the optical output power P_o , consider the following. It takes $\Delta t = nL/c$ seconds for photons to cross the laser cavity length L . Only half of the photons, $1/2 N_{\text{ph}}$, in the cavity would be moving toward the output face of the crystal at any instant. Only a fraction $(1 - R)$ of the radiation power will escape. Thus, the output optical power P_o is

$$P_o = \frac{\left(\frac{1}{2} N_{\text{ph}}\right)(\text{Cavity Volume})(\text{Photon energy})}{\Delta t} (1 - R)$$

and using N_{ph} from Eq. (4.13.6), we finally obtain the **laser diode equation**

Laser
diode
equation

$$P_o = \left[\frac{hc^2\tau_{\text{ph}}(1 - R)}{2en\lambda L} \right] (I - I_{\text{th}}) \quad (4.13.7)$$

We can also express Eq. (4.13.7) in terms of the output coherent radiation intensity $I_o (= P_o/Wd)$ and the current densities $J (= I/WL)$ and $J_{\text{th}} (= I_{\text{th}}/WL)$,

Laser
diode
equation

$$I_o = \left[\frac{hc^2\tau_{\text{ph}}(1 - R)}{2en\lambda d} \right] (J - J_{\text{th}}) \quad (4.13.8)$$

We neglected the small spontaneous radiation that will also be present in the output in Eq. (4.13.7) in our derivation. This spontaneous radiation in a laser operating above the threshold would be small.

Equation (4.13.7) represents a linear coherent light output power vs. diode current behavior as illustrated in Figure 4.47 (d). Note the similarity between the case for semiconductors and that for gas lasers in Figure 4.23. The plot in Figure 4.47 (d) represents steady state operation. When the current is above I_{th} , the injected electrons rapidly recombine and emit coherent photons, and n_{th} does not change. n becomes clamped at $n = n_{\text{th}}$ for $I > I_{\text{th}}$. The coherent radiation appears as output only when $I > I_{\text{th}}$. At a given current level above the threshold, upon applying the bias, initially the gain in the active regions is actually more than the threshold gain, and n exceeds the threshold concentration n_{th} until oscillations are built and steady state is reached; at that point $n = n_{\text{th}}$ and $g = g_{\text{th}}$. The semiquantitative steady state approach above is a special case of the more general semiconductor rate equations described in more advanced textbooks where the time response of the laser diode is also analyzed.

B. Optical Gain Curve, Threshold, and Transparency Conditions

The simple rate equation analysis above does not allow us to determine n_{th} and hence J_{th} without knowing the threshold gain. We know that the threshold population inversion is determined by the threshold gain as in Eq. (4.6.8). The latter equation was derived for a two-level system, and we need its corresponding equation applicable for semiconductors; this can be found in more advanced textbooks. There is one further modification. The threshold gain g_{th} must be modified to account for the less than 100% confinement of radiation into the active region. This is easily done by rewriting the g_{th} -condition as

$$\Gamma g_{\text{th}} = \alpha_t = \alpha_s + \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right) \quad (4.13.9)$$

Threshold
optical
gain

where Γ represents the fraction of the coherent optical radiation within the active region. The gain g works on the radiation within the cavity, which means that we must multiply g with Γ to account for less than perfect optical confinement. In Eq. (4.13.9), α_s represents various losses in the semiconductor active region as well as the cladding.³² The right-hand term in Eq. (4.13.9) represents the total losses inside and at the ends of the cavity. The α_s losses in semiconductor lasers are free carrier absorption in the active region, scattering in both the active and cladding regions, and absorption in the cladding (which has no net optical gain). As Γ decreases, resulting in less coherent radiation confinement, g_{th} has to be larger so that Eq. (4.13.9) can be satisfied and the gain Γg_{th} can make up for the losses on the right-hand side of Eq. (4.13.9). The active region formed by the heterostructure is normally sufficiently small ($d \sim 0.1 \mu\text{m}$) that there is only one mode (*i.e.*, a single lateral mode) and this mode has a mode field diameter that can substantially penetrate the confining (cladding) layers. The optical confinement factor Γ can easily be as low as 0.2.

The optical gain spectrum $g(v)$ under sufficiently large forward bias has a frequency dependence that is sketched in Figure 4.33 (b), when the LD has been forward biased to place the quasi-Fermi levels E_{Fn} in the CB and E_{Fp} in the VB as in Figure 4.33 (a). To find the optical gain curve $g(v)$ we need to consider the density of states in the CB and the VB, their occupation statistics through the Fermi–Dirac function, and positions of E_{Fn} and E_{Fp} ; the latter depend on the injected carrier concentrations n . Figure 4.48 (a) shows the optical gain curve $g(v)$ as a function

³²In the case of semiconductors, some books use α_i to represent internal losses, instead of α_s , but the meaning is the same. α_t in Eq. (4.13.9) is an *effective distributed attenuation coefficient* through the cavity length that represents all losses, that is, both internal and end losses.

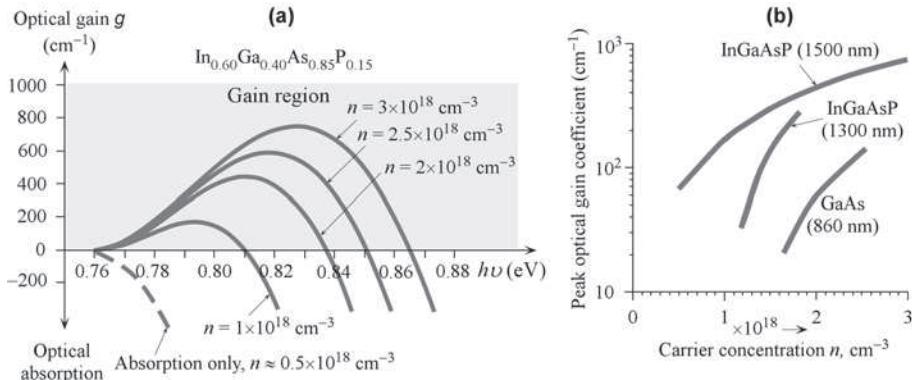


FIGURE 4.48 (a) Optical gain g vs. photon energy for an InGaAsP active layer (in a 1500-nm LD) as a function of injected carrier concentration n from $1 \times 10^{18} \text{ cm}^{-3}$ to $3 \times 10^{18} \text{ cm}^{-3}$. (The model described in Leuthold *et al.*, *J. Appl. Phys.*, 87, 618, 2000, was used to find the gain spectra at different carrier concentrations.) (b) The dependence of the peak gain coefficient (maximum g) on the injected carrier concentration n for GaAs (860 nm), $\text{In}_{0.72}\text{Ga}_{0.28}\text{As}_{0.4}\text{P}_{0.4}$ (1300 nm), and $\text{In}_{0.60}\text{Ga}_{0.40}\text{As}_{0.85}\text{P}_{0.15}$ (1500 nm) active layers. (Source: Data combined from J. Singh, *Electronic and Optoelectronic Properties of Semiconductor Structures*, Cambridge University Press, 2003, p. 390; N. K. Dutta, *J. Appl. Phys.*, 51, 6095, 1980; J. Leuthold *et al.*, *J. Appl. Phys.*, 87, 618, 2000.)

of photon energy for an $\text{In}_{0.60}\text{Ga}_{0.40}\text{As}_{0.85}\text{P}_{0.15}$ (1500 nm) active layer with different amounts of injected carrier concentration n . Notice also the shift in the peak frequency (where the gain is maximum) and the increase in the bandwidth with increasing injection. What is important is the injected carrier concentration that makes the gain g equal to the threshold gain g_{th} , which would allow laser oscillations to be maintained in the cavity. As the LD current and hence the injected carrier concentration n increase, the peak gain becomes larger as shown in Figure 4.48 (b) for $\text{In}_{0.60}\text{Ga}_{0.40}\text{As}_{0.85}\text{P}_{0.15}$ (1500 nm). The optical gain in InGaAsP alloys is greater than that in GaAs.

There is one further definition of importance in laser diode engineering. The threshold current and carrier concentration I_{th} and n_{th} correspond to the gain coefficient g_{th} just balancing all the cavity losses, including end losses, *i.e.*, $g_{\text{th}} = \alpha_s$, so that a steady state laser oscillation can be maintained. **Transparency**, on the other hand, refers to the condition that makes the gain just balance the absorption within the active medium so that the net optical gain is unity within the active medium. Thus, the active medium provides no net gain, and neither does it attenuate so that it becomes **transparent** to the radiation through it. The diode current and the injected carrier concentration for attaining transparency are denoted by I_T and n_T , which are lower than the threshold I_{th} and n_{th} .

EXAMPLE 4.13.1 Threshold current and optical output power from a Fabry–Perot heterostructure laser diode

Consider GaAs DH laser diode that lases at 860 nm. It has an active layer (cavity) length L of 250 μm . The active layer thickness d is 0.15 μm and the width W is 5 μm . The refractive index is 3.6, and the attenuation coefficient α_s inside the cavity is 10^3 m^{-1} . The required threshold gain g_{th} corresponds to a threshold carrier concentration $n_{\text{th}} \approx 2 \times 10^{18} \text{ cm}^{-3}$. The radiative lifetime τ_r in the active region can

be found (at least approximately) by using $\tau_r = 1/Bn_{\text{th}}$, where B is the direct recombination coefficient, and assuming strong injection as will be the case for laser diodes [see Eq. (3.8.7) in Chapter 3]. For GaAs, $B \approx 2 \times 10^{-16} \text{ m}^3 \text{ s}^{-1}$. What are the threshold current density and threshold current? Find the output optical power at $I = 1.5I_{\text{th}}$, and the external slope efficiency η_{slope} . How would $\Gamma = 0.5$ affect the calculations?

Solution

The reflectances at the each end are the same (we assume no other thin film coating on the ends of the cavity) so that $R = (n - 1)^2/(n + 1)^2 = 0.32$. The total attenuation coefficient α_t and hence the threshold gain g_{th} , assuming $\Gamma = 1$ in Eq. (4.13.9), is

$$g_{\text{th}} = \alpha_t = (10 \text{ cm}^{-1}) + \frac{1}{(2 \times 250 \times 10^{-4} \text{ cm})} \ln \left[\frac{1}{(0.32)(0.32)} \right] = 55.6 \text{ cm}^{-1}$$

From Figure 4.48 (b), at this gain of 56 cm^{-1} , $n_{\text{th}} \approx 2 \times 10^{18} \text{ cm}^{-3}$. This is the threshold carrier concentration that gives the right gain under ideal optical confinement, with $\Gamma = 1$.

$$\text{The radiative lifetime } \tau_r = 1/Bn_{\text{th}} = 1/[(2 \times 10^{-16} \text{ m}^3 \text{ s}^{-1})(2 \times 10^{24} \text{ m}^{-3})] = 2.5 \text{ ns}$$

Since $J = I/WL$, the threshold current density from Eq. (4.13.4) is

$$\begin{aligned} J_{\text{th}} &= \frac{n_{\text{th}}ed}{\tau_r} = \frac{(2 \times 10^{24} \text{ m}^{-3})(1.6 \times 10^{-19} \text{ C})(0.15 \times 10^{-6} \text{ m})}{(2.5 \times 10^{-9} \text{ s})} \\ &= 1.9 \times 10^7 \text{ A m}^{-2} \quad \text{or} \quad 1.9 \text{ kA cm}^{-2} \quad \text{or} \quad 19 \text{ A mm}^{-2} \end{aligned}$$

The threshold current itself is

$$I_{\text{th}} = (WL)J_{\text{th}} = (5 \times 10^{-6} \text{ m})(250 \times 10^{-6} \text{ m})(1.9 \times 10^7 \text{ A m}^{-2}) = 0.024 \text{ A} \quad \text{or} \quad 24 \text{ mA}$$

The photon cavity lifetime depends on α_t , and is given by

$$\tau_{\text{ph}} = n/(c\alpha_t) = 3.6/[(3 \times 10^8 \text{ m s}^{-1})(5.56 \times 10^3 \text{ m}^{-1})] = 2.16 \text{ ps}$$

The laser diode output power is

$$P_o = \left[\frac{hc^2\tau_{\text{ph}}(1 - R)}{2en\lambda L} \right] (I - I_{\text{th}}) = \frac{(6.626 \times 10^{-34})(3 \times 10^8)^2(2.16 \times 10^{-12})(1 - 0.32)}{2(1.6 \times 10^{-19})(3.6)(860 \times 10^{-9})(250 \times 10^{-6})} (I - I_{\text{th}})$$

that is,

$$P_o = (0.35 \text{ W A}^{-1})(I - I_{\text{th}}) = (0.35 \text{ mW mA}^{-1})(I - 24 \text{ mA})$$

When $I = 1.5I_{\text{th}} = 36 \text{ mA}$, $P_o = (0.35 \text{ mW mA}^{-1})(36 \text{ mA} - 24 \text{ mA}) = 4.2 \text{ mW}$.

The slope efficiency is the slope of the P_o vs. I characteristic above I_{th} , thus,

$$\eta_{\text{slope}} = \frac{\Delta P_o}{\Delta I} = \left[\frac{hc^2\tau_{\text{ph}}(1 - R)}{2en\lambda L} \right] = 0.35 \text{ mW mA}^{-1}$$

We can now repeat the problem, say for $\Gamma = 0.5$, which would give $\Gamma g_{\text{th}} = \alpha_t$, so that $g_{\text{th}} = 55.6 \text{ cm}^{-1}/0.5 = 111 \text{ cm}^{-1}$. From Figure 4.48 (b), at this gain of 111 cm^{-1} , $n_{\text{th}} \approx 2.5 \times 10^{18} \text{ cm}^{-3}$. The new radiative lifetime $\tau_r = 1/Bn_{\text{th}} = 1/[(2.0 \times 10^{-16} \text{ m}^3 \text{ s}^{-1})(2.5 \times 10^{24} \text{ m}^{-3})] = 2.0 \text{ ns}$.

The corresponding threshold current density is

$$J_{\text{th}} = n_{\text{th}}ed/\tau_r = (2.5 \times 10^{24} \text{ m}^{-3})(1.6 \times 10^{-19} \text{ C})(0.15 \times 10^{-6} \text{ m})/(2.0 \times 10^{-9} \text{ s}) = 30 \text{ A mm}^{-2}$$

and the corresponding threshold current I_{th} is 37.5 mA.

There are several important notes to this problem. First, the threshold concentration $n_{\text{th}} \approx 2 \times 10^{18} \text{ cm}^{-3}$ was obtained graphically from Figure 4.48 (b) by using the g_{th} value we need. Second is that, at best, the calculations represent rough values since we also need to know how the mode spreads into the cladding where there is no gain but absorption and, in addition, what fraction of the current is lost to nonradiative recombination processes. We can increase α_s to account for absorption in the cladding, which would result in a higher g_{th} , larger n_{th} , and greater I_{th} . If τ_{nr} is the nonradiative lifetime, we can replace τ_r by an effective recombination time τ such that $\tau^{-1} = \tau_r^{-1} + \tau_{nr}^{-1}$, which means that the threshold current will again be larger. We would also need to reduce the optical output power since some of the injected electrons are now used in nonradiative transitions. Third is the low slope efficiency compared with commercial LDs. η_{slope} depends on τ_{ph} , the photon cavity lifetime, which can be greatly improved by using better reflectors at the cavity ends, e.g., by using thin film coating on the crystal facets to increase R .

4.14 SINGLE FREQUENCY SEMICONDUCTOR LASERS

A. Distributed Bragg Reflector LDs

Ideally, the output spectrum from a laser device should be as narrow as possible, which generally means that we have to allow only a single mode to exist in the cavity. There are a number of device structures that operate with an output spectrum that has high modal purity and hence a very narrow spectral width. Such LDs are often called **single frequency lasers**, even though it would be, in principle, impossible to get a perfect single frequency. (Obviously, a single-mode operation is implied.)

One method of ensuring a single mode of radiation in the laser cavity is to use frequency selective dielectric mirrors at the cleaved surfaces of the semiconductor. The **distributed Bragg reflector (DBR)**, as shown in Figure 4.49 (a), is a mirror that has been designed like a reflection-type diffraction grating; it has a periodic corrugated structure at the end of the active region. Intuitively, partial reflections of waves from the corrugations interfere constructively (*i.e.*, they reinforce each other) to give a reflected wave only when the wavelength inside the medium corresponds to twice the corrugation period as illustrated in Figure 4.49 (b). For simplicity, the corrugations have been approximated by step changes in the refractive index from n_1 to n_2 to n_1 and so on. For example, two partially reflected waves at two n_1/n_2 interfaces in Figure 4.49 (b) such as A and B have an optical path difference of 2Λ , where Λ is the corrugation period. They can only interfere constructively if 2Λ is a multiple of the wavelength within the medium.

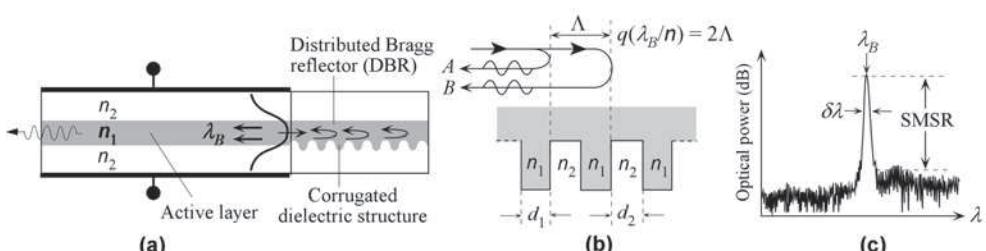


FIGURE 4.49 (a) The basic principle of the distributed Bragg reflection (DBR) laser. (b) Partially reflected waves at the corrugations can constitute a reflected wave only when the wavelength satisfies the Bragg condition. Reflected waves A and B interfere constructively when $q(\lambda_B/n) = 2\Lambda$. (c) Typical output spectrum. SMSR is the side mode suppression ratio.

Each of these wavelengths is called a **Bragg wavelength** λ_B and given by the familiar condition for in-phase interference³³

$$q \frac{\lambda_B}{n} = 2\Lambda \quad (4.14.1) \quad \text{Bragg wavelength}$$

where n is the *effective* refractive index of the corrugated medium (explained below) and $q = 1, 2, \dots$ is an integer called the **diffraction order**. The DBR therefore has a high reflectance around λ_B but low reflectance away from λ_B . The result is that only the particular optical cavity mode that is closest to the Bragg wavelength in Eq. (4.14.1), within the optical gain curve, can lase and exist in the output. The refractive index n in Eq. (4.14.1) must be viewed as an *effective* or *average refractive index* of the corrugated medium. The waves A and B in Figure 4.49 (b) involve reflections from n_1/n_2 interfaces, but there are also reflections from n_2/n_1 interfaces. If we do a proper interference accounting of all the reflections from the DBR under normal incidence, then the result would be the familiar diffraction condition in Eq. (4.14.1), but with n representing an effective or average index given by $n_{av} = (n_1 d_1 + n_2 d_2)/\Lambda$, where d_1 and d_2 are the thicknesses of the n_1 and n_2 regions in the grating. The effective index n_{av} would be different if the refractive index variation profile is different than that shown. (Exact calculations of the mode frequency is beyond the scope of this book.) Moreover, while the simplest DBR laser shown in Figure 4.49 (a) has a grating only at one end of the cavity, it is possible to have a grating at both ends to enhance the spectral selectivity and radiation intensity in the cavity; the latter would result in a lower threshold current. The output spectrum from a typical DBR laser is a very narrow mode centered at (or very close to) λ_B as indicated in Figure 4.49 (c). The spectral width, depending on the design, can be less than 1 MHz. Table 4.5 summarizes some of the important properties of a commercial DBR LD. There are various modifications and improvements to the basic DBR design illustrated in Figure 4.49 (a) that can be found in more advanced texts.

There is one important quantity, called the **side mode suppression ratio** (SMSR), that quantifies the intensity of the single-mode peak emission with respect to that of suppressed modes, which are all away from λ_B as illustrated in Figure 4.49 (d). SMSR is normally measured in dB, and typically reported values are usually 45 dB, that is, a suppression ratio of 3×10^4 .

B. Distributed Feedback LDs

In a simple Fabry–Perot cavity-type laser, the crystal facets provide the necessary optical feedback into the cavity to build up the photon concentration. In the **distributed feedback (DFB)** laser, as shown in Figure 4.50 (a), there is a corrugated layer, called the **guiding layer**, next to the active layer; radiation spreads from the active layer to the guiding layer. The mode field diameter of the radiation in the active layer is such that it covers the guiding layer. (Remember that the thickness d of the active layer is small.) These corrugations in the refractive index act as optical feedback over the length of the cavity by producing partial reflections. Thus optical feedback is *distributed* over the cavity length. Intuitively, we might infer that only those Bragg wavelengths λ_B related to the corrugation periodicity Λ as in Eq. (4.14.1) can interfere constructively and thereby exist in the “cavity” in a similar fashion to that illustrated in Figure 4.49 (b).

³³We saw in Chapter 1 that the reflectance from alternating layers of high and low refractive index structures (as in dielectric mirror) is maximum when the free-space wavelength is λ_B , where λ_B is given by Eq. (4.14.1) in which n is an effective (average) index given by $n_{av} = (n_1 d_1 + n_2 d_2)/\Lambda$ if the modulation consists of periodic step changes in the refractive index from n_1 to n_2 to n_1 , etc.

TABLE 4.5 Selected properties of DBR, DFB, and external cavity (EC) laser diodes

LD	λ_o (nm)	$\delta\nu, \delta\lambda$	SMSR (dB)	P_o (mW)	I (mA)	η_{slope} (mA)	Comment
DBR ^a	1063	2 MHz, 8 fm	45	80	200	0.8	GaAs DBR LD for spectroscopy and metrology, includes monitor current, TEC, and thermistor.
DFB ^b	1063	2 MHz, 8 fm	45	80	190	0.2	GaAs DFB LD for spectroscopy and metrology, includes monitor current, TEC, and thermistor.
DFB ^c	1550	10 MHz, 0.08 pm	45	40	300	0.3	Pigtailed to a fiber, includes monitor current, TEC, and thermistor. CW output for external modulation. For use in long-haul DWDM.
DFB ^d	1653	0.1 nm	35	5	30	0.23	Pigtailed to a single-mode fiber, includes monitor current, TEC, and thermistor. Mainly for fiber optic sensing.
EC ^e	1550	50 kHz, 0.4 fm	45	40	300	0.2	Pigtailed. Tunable over $\Delta\nu = 3$ GHz. Mainly for communications.

Notes: λ_o is the peak emission wavelength (λ_o for the DFB LD); I is typical operating current; fm is 10^{-15} s; $\delta\nu$ and $\delta\lambda$ are spectral widths; SMSR is the side mode suppression ratio; TEC is thermoelectric cooler. ^aEagleyard, EYP-DBR-1080-00080-2000-TOC03-0000;

^bEagleyard, EYP-DFB-1083-00080-1500-TOC03-0000; ^cFurukawa-Fitel, FOL15DCWD; ^dInPhenix, IPDFD1602; ^eCovega SFL1550S, marketed by Thorlabs.

However, the operating principle of the DFB laser is totally different. Radiation is fed from the active into the guiding layer along the whole cavity length, so that the corrugated medium can be thought of as possessing an optical gain. Partially reflected waves experience gain, and we cannot simply add these without considering the optical gain and also possible phase changes [Eq. (4.14.1) assumes normal incidence and ignores any phase change on reflection]. A left-traveling wave in the guiding layer experiences partial reflections, and these reflected waves are optically amplified by the medium to constitute a right-going wave.

In a Fabry–Perot cavity, a wave that is traveling toward the right becomes reflected to travel toward the left. At any point in the cavity, as a result of end-reflections, we therefore have these right- and left-traveling waves interfering, or being “coupled.” These oppositely traveling waves, assuming equal amplitudes, can only set up a standing wave, a mode, if they are *coherently coupled*, which requires that the round-trip phase change is 2π . In the DFB structure,

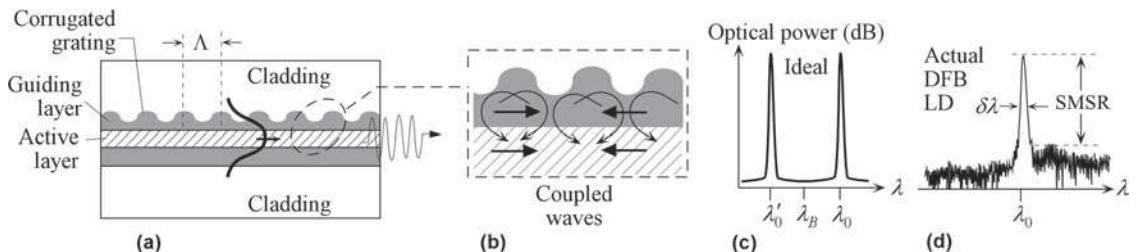


FIGURE 4.50 (a) Distributed feedback laser structure. The mode field diameter is normally larger than the active layer thickness and the radiation spreads into the guiding layer. (b) There are left and right propagating waves, partial reflections from the corrugation, and optical amplification within the cavity, which has both the active layer and the guiding layer. (c) Ideal lasing emission output has two primary peaks above and below λ_B . (d) Typical output spectrum from a DFB laser has a single narrow peak with a $\delta\lambda$ typically very narrow and much less than 0.1 nm.

traveling waves are reflected partially and *periodically* as they propagate in the cavity as illustrated in Figure 4.50 (b). The left- and right-traveling waves can only coherently couple to set up a mode if their frequency is related to the spatial corrugation period Λ , taking into account that the medium alters the wave-amplitudes via optical gain.³⁴ The allowed DFB modes are not exactly at Bragg wavelengths but are symmetrically placed about λ_B . If λ_m is an allowed DFB lasing mode then

$$\lambda_m = \lambda_B \pm \frac{\lambda_B^2}{2nL} (m + 1) \quad (4.14.2)$$

DFB laser wavelengths

where m is a mode integer, 0, 1, 2, ..., and L is the effective length of the diffraction grating, corrugation length (same as the cavity length in this very simple example). The index n in Eq. (4.14.2) is again the *effective* refractive index. The relative threshold gain for higher modes is so large that only the $m = 0$ mode effectively lases. A perfectly symmetric device has two equally spaced modes, at λ_0 and λ_0' , placed around λ_B as in Figure 4.50 (c), corresponding to $m = 0$. In reality, either inevitable asymmetry introduced by the fabrication process, or asymmetry introduced on purpose, leads to only one of the modes to appear as shown in Figure 4.50 (d). Further, typically the corrugation length L is so much larger than the period Λ that the second term in Eq. (4.14.2) is very small and the lasing wavelength is very close to λ_B . Commercial DFB lasers also have thin optical coatings at the ends of the crystal facets to modify the reflectance, *i.e.*, the transmittance. For example, the facet for the radiation output would have an antireflection coating to encourage the transmission of the lasing cavity mode through this facet, and the other facet would have an optical coating to increase the reflectance.

There are numerous commercially available single-mode DFB lasers in the market for various applications such as telecom, instrumentation, metrology, interferometry, and spectroscopy. Typically the spectral widths of telecom DFB LDs are less than 0.1 nm at the communications channel of 1.55 μm . There are high-end DFB LDs that have exceptionally narrow spectral widths, a few MHz (corresponding to a few femtometers of linewidth $\delta\lambda$ in wavelength), and can be used in various spectroscopic applications where spectral purity is important. Table 4.5 highlights the properties of a few commercial DFB lasers.

One technical drawback of using DFB LDs in instrumentation is the shift of the peak wavelength λ_0 with small changes in the temperature, that is, $d\lambda_0/dT$; both Λ and n change with temperature. The increase in λ_0 is $\sim 0.1 \text{ nm } ^\circ\text{C}^{-1}$. Most DFB LDs come with thermoelectric coolers (TEC) that control the temperature and limit $d\lambda_0/dT$ below certain acceptable limits. Some DFB diode manufacturers, in fact, use the $d\lambda_0/dT$ shift as a means to *tune* the laser output over a range of wavelengths, usually a few nanometers; obviously the temperature has to be very closely controlled.

When the DFB LD is modulated by its current, as in optical communications, the frequency of the optical field in the output optical pulse should remain the same at the operating frequency ω_0 throughout the optical pulse duration. However, in practice, the light pulse exhibits what is called a **frequency chirp**, that is, the frequency of the optical field changes over the time duration of the optical pulse. The reason is that the refractive index n depends on the gain of the active medium, which depends on the electron concentration, n . But, n depends on the current

³⁴These partially reflected waves travel in a medium that has the refractive index modulated periodically (with periodicity Λ). We have to consider how left- and right-propagating waves in this corrugated structure are coupled. The rigorous theory is beyond the scope of this book and is called “Coupled Mode Theory.”

(increases n) and coherent radiation intensity (decreases n through stimulated emissions).³⁵ Any change in n causes a phase change, say $\Delta\phi$, in the emitted radiation. The rate of change of this phase change during modulation, *i.e.*, $d\Delta\phi/dt$, is equivalent to an instantaneous frequency shift $\Delta\omega$, which is the **frequency chirp** defined above. Further, the increase in the current during modulation would also result in some additional Joule heating of the semiconductor. However, the TEC is not sufficiently fast to maintain the temperature constant, and hence prevent the shift in ω_0 . The narrow spectral widths quoted in MHz (or less) normally refer to *instantaneous* spectral widths inasmuch as fluctuations during operation will cause small shifts in λ_0 , and these shifts would appear as though $\delta\lambda$ is broader. Some manufacturers use an effective $\delta\lambda$ to account for the latter effect. The output spectrum would appear as though broadened, which would not be the same spectral width under stabilized CW operation.

C. External Cavity LDs

The main problem with the DFB laser is that the output radiation suffers from frequency chirping. It would be highly desirable to have LDs that are immune to frequency chirping and also possess stability against any drift. By having the optical cavity external to the laser diode, in so-called **external cavity diode lasers** (ECDL), these problems can be overcome, though at the expense of size. In addition, one can increase the photon cavity lifetime (τ_{ph}) and narrow the spectral width further. The basic idea involved in ECDLs is quite simple. The light from one of the facets of the LD is coupled into an external optical resonator, and a wavelength-selective component, such as a diffraction grating or an interference filter, is used within the external cavity to select the wavelength of interest (λ_o), within the gain bandwidth of the LD.

Figure 4.51 is a schematic diagram of one of the many types of ECDL that are available commercially. The radiation from the left facet B of the LD is collected, expanded, and collimated by the lens L_2 . The collimated beam passes through an angled interference filter (IF).

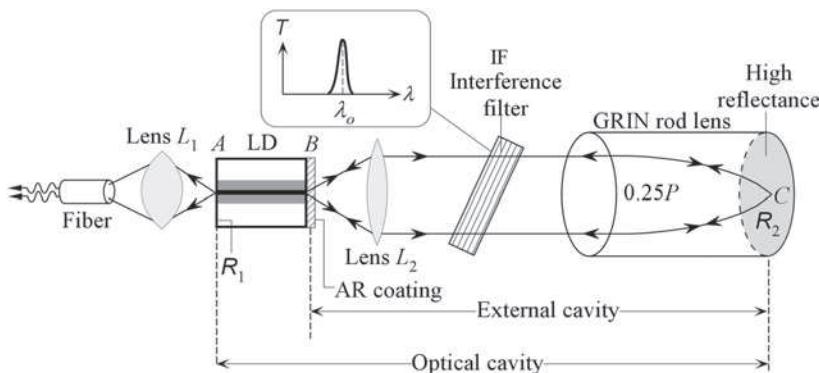


FIGURE 4.51 A simplified diagram of an external cavity diode laser (ECDL), which uses an angled interference filter (IF) to select the wavelength λ_o (depends on the tilt angle of the IF), and the optical cavity has a GRIN rod lens with one end coated for full reflection back to the LD. The output is taken from the left facet of the LD.

³⁵It might be thought that the electron concentration should be clamped at n_{th} , but this is only true under CW operation. We need to solve the full rate equations to obtain the transient response, which is normally done in advanced textbooks, *e.g.*, see A. Yariv and P. Yeh, *Photonics*, 6th Edition (Oxford University Press, 2007), Ch. 15.



A commercial external cavity diode laser using the basic principle shown in Figure 4.51 (U.S. Patent 6,556,599, Bookham Technology). The output is a single mode at 785 nm ($\pm 1.5\text{ pm}$) with a linewidth less than 200 kHz, and coupled into a fiber. The output power is 35 mW, and the SMSR is 50 dB. (ECDL, SWL-7513-P. Courtesy of Newport, USA.)

The latter has excellent transmittance T at one particular wavelength λ_o , within the gain bandwidth of the LD, with a very narrow spectral width $\delta\lambda$. It is somewhat similar to a dielectric mirror, but the alternating dielectric layers are chosen in such a way to transmit at one wavelength rather than reflect. The transmitted wavelength λ_o depends on the angle of the IF to the incident radiation. Further, IFs need the incident light to be collimated for the full utilization of their properties (such as wavelength selectivity and a narrow $\delta\lambda$). The radiation that passes the IF is at λ_o and enters a GRIN rod lens (a graded index glass rod that acts like a lens) whose end C has been coated to obtain high reflectance, R_2 . The light reflected from the right face C of the GRIN rod travels back, passes through the IF again, and is fed back into the LD by the lens L_2 . The outside surface of the LD's facet B is coated with an antireflection (AR) coating to prevent any of this light being reflected; we need to couple the light from the external cavity back into the LD. The light then propagates along the active region of the LD, experiences optical gain, and becomes reflected at the left facet A of the LD (reflectance R_1) to complete a round trip. Some of the light, $(1 - R_1)$ fraction, escaping from the left end A is collected by the lens L_1 , and fed into a fiber, or collimated as an output beam. The true optical cavity is actually between A and C . The external cavity is between B and C , and much longer than the LD's cavity AB . The AR coating on the LD's facet B suppresses the role of LD's cavity AB . Since the optical path AC is much longer than the path AB , perturbations to the optical path due, for example, to changes in the refractive index in the active region along AB , do not cause shifts in the wavelength. Properly designed ECDLs do not suffer from frequency chirping, which is their advantage. Further, by using a highly wavelength-selective element in the external cavity, such as an interference filter or a suitable diffracting grating, the output radiation can be made to have very narrow spectral width $\delta\nu$, e.g., less than 200 kHz for the ECDL in Figure 4.51.

Many ECDLs are **tunable lasers** over a range of wavelengths; that is, the center emission wavelength λ_o can be varied over the gain bandwidth of the LD. Tuning is done most easily by changing the transmission wavelength (λ_o) through the wavelength-selective element in the external cavity. In Figure 4.51, one can change the tilt angle of the IF to tune the external cavity. There are several other types of ECDLs, most of which use a diffraction grating. If we were to replace the IF in Figure 4.51 by a tilted reflection-type diffraction grating, and eliminate the GRIN rod, we would obtain what is called a **Littrow configuration** ECDL. The EC would be between the grating and the LD's facet B . The orientation of the diffraction grating would select the output wavelength. The Littrow grating-based ECDLs are widely used since they are quite simple to construct. As shown in Table 4.5, ECDLs can have very narrow linewidths and high SMSRs.

EXAMPLE 4.14.1 DFB LD wavelength

Consider a DFB laser that has a corrugation period Λ of $0.22\text{ }\mu\text{m}$ and a grating length of $400\text{ }\mu\text{m}$. Suppose that the effective refractive index of the medium is 3.5. Assuming a first-order grating, calculate the Bragg wavelength, the mode wavelengths, and their separation.

Solution

The Bragg wavelength from Eq. (4.14.1) is

$$\lambda_B = \frac{2\Lambda n}{q} = \frac{2(0.22\text{ }\mu\text{m})(3.5)}{1} = 1.5400\text{ }\mu\text{m}$$

and the symmetric mode wavelengths about λ_B are

$$\lambda_m = \lambda_B \pm \frac{\lambda_B^2}{2nL} (m + 1) = 1.5400 \pm \frac{(1.5400\text{ }\mu\text{m})^2}{2(3.5)(400\text{ }\mu\text{m})} (0 + 1)$$

so that the $m = 0$ mode wavelengths are

$$\lambda_0 = 1.53915 \quad \text{or} \quad 1.54085\text{ }\mu\text{m}$$

The two are separated by $0.0017\text{ }\mu\text{m}$ or 1.7 nm. Due to a design asymmetry, only one mode will appear in the output and for most practical purposes the mode wavelength can be taken as λ_B . *Note:* The wavelength calculation was kept to five decimal places because λ_m is very close to λ_B .

4.15 VERTICAL CAVITY SURFACE EMITTING LASERS³⁶

Figure 4.52 shows the basic concept of the vertical cavity surface emitting laser (VCSEL). A VCSEL has the optical cavity axis along the direction of current flow rather than perpendicular to the current flow, as in conventional laser diodes. The active region length is very short compared with the lateral dimensions so that the radiation emerges from the “surface” of the cavity rather than from its edge. The reflectors at the ends of the cavity are **dielectric mirrors**, that is, **distributed Bragg reflectors** (DBRs) made from alternating high and low refractive index quarter-wave thick multilayers. Such dielectric mirrors provide a high degree of wavelength-selective reflectance at the required free-space wavelength λ . If the thicknesses of alternating layers are d_1 and d_2 with refractive indices n_1 and n_2 , then the condition

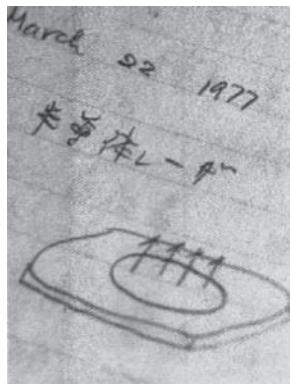
$$n_1 d_1 + n_2 d_2 = \frac{1}{2} \lambda \tag{4.15.1}$$

leads to the constructive interference of all partially reflected waves at the interfaces, and the alternating layers function as a dielectric mirror (as explained in Chapter 1). Since the wave is reflected because of a periodic variation in the refractive index, as in a grating, the dielectric mirror is essentially a **distributed Bragg reflector**. Equation (4.15.1) is equivalent to the Eq. (4.14.1) for the distributed Bragg reflector in which the effective index n is $(n_1 d_1 + n_2 d_2)/\Lambda$.

³⁶The VCSEL was conceived by Kenichi Iga (Tokyo Institute of Technology) in 1977, and proposed at a conference in 1978 with the first device implementation in 1979. (See K. Iga, *Jpn. J. Appl. Phys.*, 47, 1, 2008.)



This VCSEL diode provides a single transverse mode emission 795 nm. The spectral width is less than 100 MHz, and the output power is 0.15 mW at 2 mA. (*Courtesy of Vixar Inc.*)



Sketch of the VCSEL in Kenichi Iga's laboratory book (1997). Professor Iga was at the Tokyo Institute of Technology at the time. (*See K. Iga, Jpn J. Appl. Phys., 47, 1, 2008; Courtesy of Kenichi Iga.*)



Kenichi Iga, currently (2012) the President of the Tokyo Institute of Technology, was first to conceive the VCSEL, and played a pioneering role in the development of VCSELs. (*Courtesy of Kenichi Iga.*)

The wavelength in Eq. (4.15.1), that is, the Bragg wavelength λ_B , is chosen to lie within the optical gain of the active layer. High reflectance end mirrors are needed because the short cavity length L reduces the optical gain of the active layer inasmuch as the optical gain is proportional to $\exp(gL)$, where g is the optical gain coefficient. There may be 30–60 or so pairs of layers in the dielectric mirrors to obtain the required reflectances; roughly $\sim 99\%$ for the bottom and $\sim 95\%$ for the top

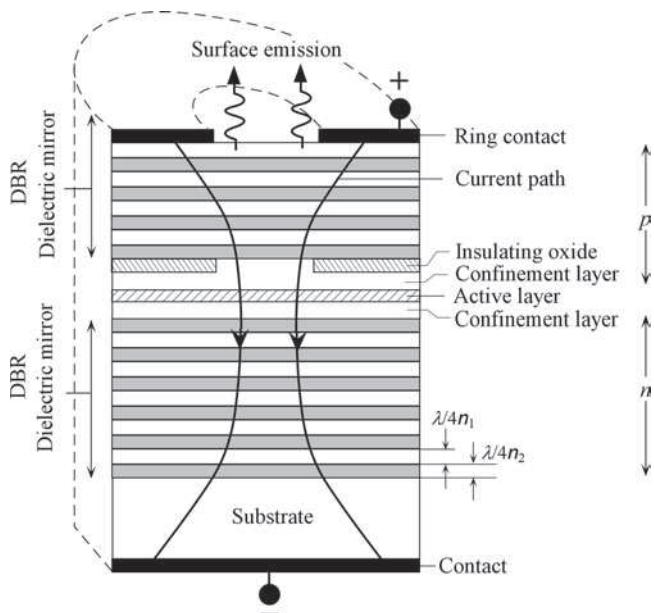
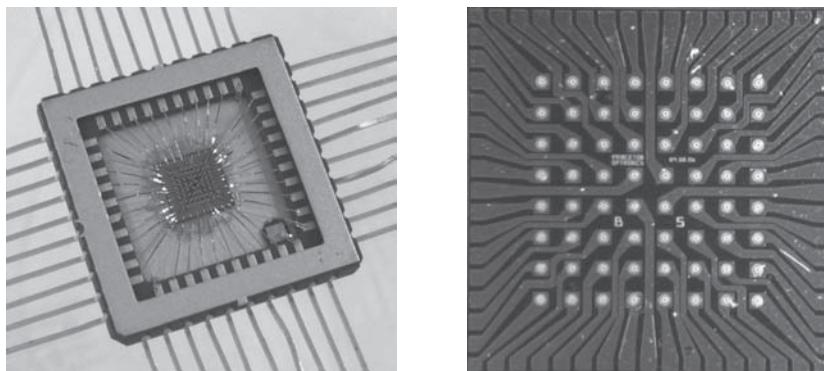


FIGURE 4.52 A simplified schematic illustration of a vertical cavity surface emitting laser. The cross-section is circular.

emitting side. The whole optical cavity in Figure 4.52 looks “vertical” if we keep the current flow the same as in a conventional laser diode cavity.

The active layer is generally very thin ($<0.1\text{ }\mu\text{m}$) and is likely to be a multiple quantum well for reduced threshold current. The active layer is sandwiched between two confinement layers. There is a thin insulating oxide layer that covers the region outside the central part, and hence confines the current flow to the central region to increase the current density flowing into the active region, and lower the threshold current for lasing. The required semiconductor layers are grown by epitaxial growth on a suitable substrate. The emission can be through the top dielectric mirror, as in Figure 4.52, or, as in some VCSELs, through the GaAs substrate. The latter are called **substrate** or **bottom emitting** VCSELs. For example, a 980-nm emitting VCSEL device has InGaAs as the active layer to provide the 980-nm emission, and a GaAs crystal is used as a substrate which is transparent at 980 nm. The dielectric mirrors are then alternating layers of AlGaAs with different compositions and hence different bandgaps and refractive indices. In practice, the current flowing through the dielectric mirrors give rise to an undesirable voltage drop and methods are used to feed the current into the active region more directly, for example, by depositing “peripheral” contacts close to the active region. There are various sophisticated VCSEL structures, and Figure 4.52 is only one very simplified example.

The vertical cavity is generally circular in its cross-section so that the emitted beam has a **circular cross-section**, which is a distinct advantage. The height of the vertical cavity may be as small as several microns. Therefore, the longitudinal mode separation is sufficiently large to allow only one longitudinal mode to operate. However, there may be one or more lateral or transverse modes depending on the lateral size of the cavity. In practice there is only one single transverse mode (STM) (and hence one mode) in the output spectrum for cavity diameters less than $\sim 10\text{ }\mu\text{m}$. The spectral line width from a VCSEL that has single transverse mode is typically much less than 0.1 nm. Specially designed VCSELs can have spectral widths ($\Delta\nu$) as low as 100 MHz, that is, a linewidth ($\Delta\lambda$) of 0.2 pm. Many VCSELs in the market have several lateral modes, and the spectral width is about $\sim 0.5\text{ nm}$, significantly less than a conventional



Left: A packaged addressable VCSEL array with 8×8 individually addressable laser devices. The chip is $3\text{ mm} \times 3\text{ mm}$. Right: A closer view of the chip. (Courtesy of Princeton Optronics, USA.)

longitudinal multimode Fabry–Perot (FP) laser diode. There is, however, a shift in the emission wavelength with temperature as in the case of DFB and DBR LDs. For VCSELs, the wavelength increases by about 0.05 nm per 1°C increase, which is slightly better than typical values quoted for DBR LDs (0.06–0.1 nm °C⁻¹). Table 4.6 summarizes some of the properties of selected single and multi-transverse mode VCSELs. Compared to conventional FP and DBR LDs, two features should be immediately apparent. First is the lower threshold current; and hence a lower overall operating current. Second is the lower optical power emitted. However, we should compare emitted intensities, power emitted per unit area, which is ~1 kW cm⁻², and higher than edge emitting FP LDs.

With cavity dimensions in the microns range, such a VCSEL laser is referred to as a **microlaser**. One of the most significant advantages of microlasers is that they can be arrayed to construct a **matrix emitter** that is a broad area surface emitting laser source. Such laser arrays have important potential applications in optical interconnect and optical computing technologies. Further, such laser arrays can provide a higher optical power than that available from a single conventional laser diode. Powers reaching a few watts are easily obtained from such matrix lasers.

VCSELs are used in many diverse applications such as short-haul optical communications (local area networks, data communications up to 10 Gb s⁻¹), sensors, encoders, bar code readers, printing, optical mice, spectroscopy, and smart cables. Graded index fibers have now widely replaced multimode step index fibers in short-haul communications. These graded index fibers need narrow linewidth emitters to avoid the material dispersion (which is significant at 850 nm) overwhelming the intrinsic (and optimized) intermode dispersion of the fiber³⁷; VCSELs are highly suitable for such use in gigabit Ethernet (*e.g.*, 2.5–10 Gb s⁻¹). When many VCSELs are integrated onto a single chip as arrays, the emission power from the chip can be quite large. There are commercial array chips, and modules with several chips, that can easily emit 100 W or more. In addition, there are chips with VCSEL arrays that are addressable so that each VCSEL diode on the array can be externally controlled.

VCSELs offer a number of distinct advantages in terms of much easier manufacturability (with high yield and volume), testing, packaging, coupling into optical fibers, and

TABLE 4.6 Selected characteristics of a few commercial VCSELs

λ (nm)	$d\lambda/dT$		I_{th} (mA)	I (mA)	V (V)	P_o (mW)	η_{slope} (mW mA ⁻¹)	VCSEL
	$\Delta\lambda$ (nm)	(nm K ⁻¹)						
680	<0.1	0.046	21°	0.4	3	2.8	0.95	STM ^a
680	0.5	0.04	15°	2.5	8	2.3	4	MTM ^a
850	0.0002	0.06	17°	0.5	2	2.2	0.75	STM ^b
850	<0.65	0.06	26°	1	6	1.9	2	MTM ^c

Notes: STM and MTM refer to single transverse mode and multi-transverse mode devices. ^aVixar, 680-000-x001, single- and multimode devices; ^bLaser Components (ULM Photonics) ULM775-03-TN; ^cOclaro, 850 nm 8.5 Gb/s Multimode VCSEL Chip. *I* and *V* are typical operating current and voltage values.

³⁷The material dispersion is quite significant at 850 nm, so that, unless a very narrow linewidth light source is used (*e.g.*, a VCSEL), the material dispersion would be greater than the intermodal dispersion characteristic of the graded index fiber.

the implementation of 2D arrays, which can be addressable as well. Advances on VCSELs are likely to make this laser diode technology one of the leaders in photonics over the next decade.³⁸

4.16 SEMICONDUCTOR OPTICAL AMPLIFIERS

A semiconductor laser structure can also be used as an optical amplifier that amplifies light waves passing through its active region as illustrated in Figure 4.53 (a). The wavelength of radiation to be amplified must fall within the optical gain bandwidth of the laser. Such a device would not be a laser oscillator, emitting lasing emission without an input, but an optical amplifier with input and output ports for light entry and exit. In the **traveling wave (TW) semiconductor optical amplifier (SOA)**, the ends of the optical cavity have antireflection (AR) coatings so that the optical cavity does not act as an efficient optical resonator, a condition for laser oscillations. Laser oscillations are therefore prevented. Light, for example, from an optical fiber, is coupled into the active region of the laser structure. As the radiation propagates through the active layer, optically guided by this layer, it becomes amplified by the induced stimulated emissions, and leaves the optical cavity with a higher intensity. Obviously the device must be pumped to achieve optical gain (population inversion) in the active layer. Random spontaneous emissions in the active layer feed “noise” into the signal and broaden the spectral width of the passing radiation. This can be overcome by using an optical filter at the output to allow the original light wavelength to pass through. Typically, such laser amplifiers are buried heterostructure devices and have optical gains of ~ 20 dB depending on the efficiency of the AR coating.

The **Fabry–Perot semiconductor optical amplifier**, as shown in Figure 4.53 (b), is similar to the conventional laser oscillator, but is operated below the threshold current for lasing oscillations. The active region has an optical gain but not sufficient to sustain a self-lasing output, *i.e.*, the gain coefficient g is below the threshold gain g_{th} . Light passing through such an active region will be amplified by stimulated emissions but, because of the presence of an optical resonator, there will be internal multiple reflections as indicated in Figure 4.53 (b). These multiple reflections lead to the gain being highest at the resonant frequencies ν_m of the cavity within the optical gain bandwidth of the active region material. Optical frequencies around the cavity resonant

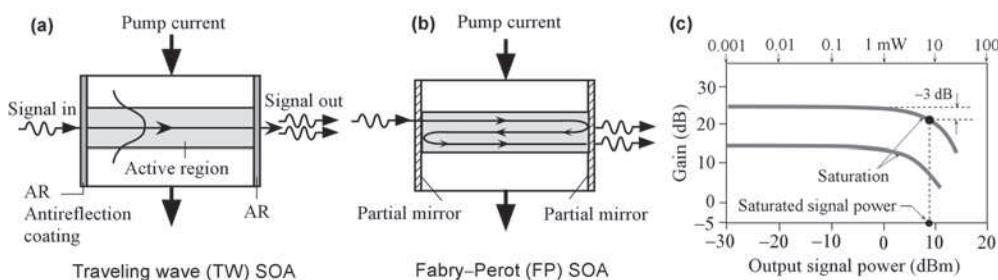


FIGURE 4.53 (a) Traveling wave (TW). (b) Fabry–Perot semiconductor optical amplifier (SOA). (c) Gain vs. output power for a 1500-nm TW SOA (GaInAsP). SOAs exhibit saturation at high output powers; saturation lowers the gain. Higher gain (25 dB) is achieved by passing a higher current through the SOA. (*Source:* Selected data used from T. Saitoh and T. Mukai, *IEEE J. Quantum Electron.*, QE23, 1010, 1987.)

³⁸See, for example, the excellent review by J. S. Harris *et al.*, *Semicond. Sci. Technol.*, 26, 1, 2011.

frequencies will experience higher gain than those away from resonant frequencies. The FP SOA is basically a **regenerative amplifier** in which positive feedback (reflections back into the cavity from the end mirrors) has been used to boost the overall gain. Although the FP amplifier can have a higher gain than the TW amplifier, due to the positive feedback, it is less stable; if the gain is not properly controlled, it can oscillate—lase.

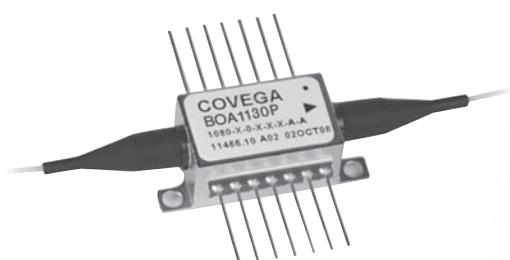
At high output signal power, the gain exhibits saturation, just as in EDFA, which is shown in Figure 4.53 (c). The reason is that at high output powers, there is a high concentration of photons in the active region, which increases the stimulated recombination rate. The increase in rate of stimulated transitions decreases the population inversion, and hence reduces gain. SOA specifications normally also quote values for the saturation output signal power P_{osat} at which the gain would have dropped by 3 dB as illustrated in Figure 4.53 (c). For a given SOA, P_{osat} depends on the gain, being lower for smaller gain, because a lower gain implies a smaller population inversion, which can be depleted more easily.

The gain of SOAs also exhibit some polarization dependence, *i.e.*, dependence on the direction of the optical field. We can view the active layer of the SOA as a planar dielectric waveguide, which implies that there will be two types of traveling waves, TE and TM polarized. These will be amplified by different amounts along the active layer. **Polarization-dependent gain**, ΔG_p , is defined as the difference between maximum and minimum gain for different orientations of the optical field (TE and TM waves), and is quoted in dB. Table 4.7 summarizes some of the properties of a few selected TW SOAs. Most SOAs come coupled (pigtailed) to a fiber.

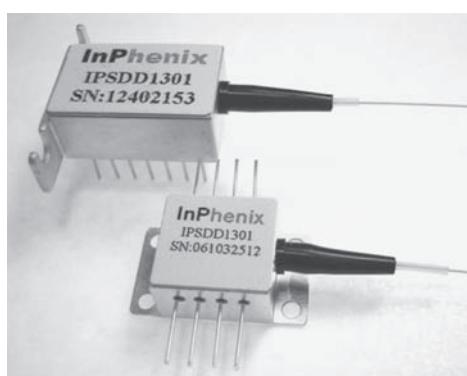
TABLE 4.7 Various properties of three selected TW SOAs

λ (nm)	Gain (dB)	BW (nm)	I_F (mA)	V_F (V)	P_{osat} (dBm)	NF (dB)	ΔG_p (dB)	Comment
850	25	40	150	2.2	8	7		Pigtailed (fiber coupled) ^a
1310	22	45	250	~2	10	7	0.5	Pigtailed, in-line amplifier ^b
1550	20	35	200	~2	11	6	0.5	Pigtailed preamplifier ^c

Notes: BW is the bandwidth of the amplifier, wavelengths over which it amplifies the optical signal; I_F and V_F are the forward current and voltage under normal operation with the given gain; ΔG_p is the polarization-dependent gain. ^aSuperlum Diodes, SOA-372-850-SM; ^bInPhenix, IPSAD1301; ^cAmphotonix (Kamelian) C-band, preamp.



Covega semiconductor optical amplifier (SOA) for use as a booster amplifier in the O-band (around 1285 nm). The small signal gain is 27 dB and the NF is 7 dB. (*Courtesy of Thorlabs.*)



Two superluminescent light-emitting diodes; each has been pigtailed to a fiber for operation at 1310 nm. Their spectral width is 40–45 nm. (*Courtesy of InPhenix.*)

Additional Topics

4.17 SUPERLUMINESCENT AND RESONANT CAVITY LEDs: SLD AND RCLED

Superluminescent light-emitting diodes, simply called **superluminescent diodes**³⁹ (SLDs) are very bright LEDs that have an active region with gain to amplify the spontaneously emitted photons. The basic principle is the amplification of spontaneous emissions (ASE). They are not lasers because there are no lasing oscillations, and the light output is just amplified spontaneous emission, *i.e.*, incoherent photons. They operate in a domain that is between LEDs and lasers in the sense that they are based on spontaneous emission, but then the diode structure has an optical cavity with gain to amplify the spontaneously emitted light. A typical stripe geometry SLD structure is shown in Figure 4.54 (a). The structure is based on heterostructures, as in the stripe geometry LD, but there is one important difference. The back-end, shown as L_a , of the SLD shown in Figure 4.54 (a) is not pumped. The top electrode does not cover this region. Consequently the back-end of the cavity (L_a) is lossy, and there are no reflections from the rear end. Thus, laser oscillations are suppressed, but the active region, shown as L_g , is still pumped to achieve high optical gains. Spontaneous photons that are emitted toward the front facet form forward guided waves. As the waves cross the pumped active region, L_g , they are amplified. Since the optical cavity is suppressed, we can pump the LD well above its usual threshold gain g_{th} (which no longer applies as the rear-end mirror no longer reflects the waves). The front end typically has an antireflection coating to allow the photons to escape.

Another structure for the SLD is the tilted cavity shown in Figure 4.54 (b). In normal FP LDs, the cavity has cleaved crystal facets at the ends that reflect the radiation back into the cavity. However, if the cavity is tilted, as in Figure 4.54 (b), then the reflections back into the cavity are suppressed, and lasing oscillations are avoided. Tilted cavity SLDs are also commercially available.

Since the gain inside L_g in the SLD increases with the current, the increase in the output power P_o with the current is more than that in the conventional LED, due to this extra gain. Unlike conventional LEDs, the SLD does not normally suffer from saturation effects at high currents. The second most important advantage is the high optical powers that are emitted from the SLD, which leads to higher powers being launched into a fiber. SLDs have found good use in short-haul optical communications such as local area networks. The spectral width of an SLD used in

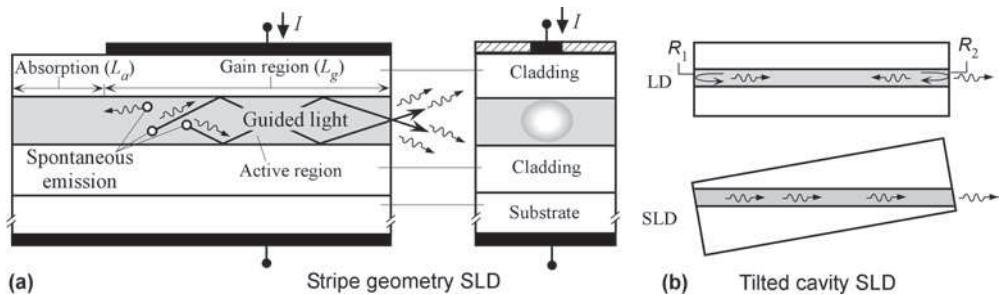


FIGURE 4.54 Superluminescent LED principles. (a) Stripe geometry SLD. (b) Tilted cavity SLD.

³⁹The superluminescent LED was invented at RCA Laboratories (Sarnoff Corporation) by Gerard Alphonse in 1986.

communications tends to be narrower than that of a corresponding conventional LED. For example, typically an SLD operating at 1310 nm would have a spectral width approximately in the 35–70 nm range, whereas the conventional LED would have 60–130 nm. However, there are also many SLDs used for their high brightness with significantly wider spectral widths than corresponding LEDs at high currents, and the spectral width increases with the injection current. The output spectrum from an SLD often has small ripples, tiny spikes, corresponding to the resonant frequencies of the optical cavity, which is not totally suppressed. Manufacturers normally quote the extent of this ripple in the output spectrum, called **gain** or **Fabry–Perot ripples** and quoted in dB.

Another important LED technology that is related to lasers is the **resonant cavity LED**⁴⁰ (RCLED). RCLED is almost identical to the VCSEL, with a structure very similar to that in Figure 4.52. The distributed Bragg reflector on the upper emission side is made to have a reflectance that is 40–60% whereas the other DBR at the lower side is nearly fully reflecting (95–99%). The poor reflectance of the upper DBR means that the threshold gain is high and the device can be operated without going into lasing oscillations. The spontaneously emitted photons are amplified by the thin active region but then they are reflected back and forth and amplified many times before being emitted. It is reminiscent to the FP SOA which operates below oscillation but has regenerative gain due to reflections from the ends feeding the radiation back into the cavity. The cavity modes matter because these are the modes that are reflected back and forth and become amplified. The emitted radiation from an RCLED therefore has a spectrum that represents a cavity mode with a spectral width that is determined by the cavity losses (or the cavity finesse), primarily the lossy upper DBR. The spectral width of the emitted radiation is broader than that of the corresponding VCSEL (where both DBRs are highly reflecting), but narrower than conventional and superluminescent LEDs; typical values being a few nanometers. Since the LED operation is based on amplifying photons emitted into a mode of the resonant cavity, the device is called a **resonant cavity LED**.

4.18 DIRECT MODULATION OF LASER DIODES

The modulation of the optical output power from a laser diode is achieved by modulating the LD drive current. The modulation can be from an analog signal or digital as in the case of LEDs. We first consider small signal AC modulation as in Figure 4.55 (a). A DC current I_1 is applied first to bias the LD and obtain a DC (steady) optical output power P_{o1} . The LD is then modulated by a sinusoidal signal $\Delta I_m \sin(2\pi ft)$ superimposed on I_1 as in Figure 4.55 (a), where ΔI_m is the maximum amplitude of the signal and f is the modulation frequency. The modulation of the diode input current about I_1 gives rise to the output optical power varying about its DC value of P_{o1} , which can be represented by $\Delta P_m \sin(2\pi ft + \phi)$, as in Figure 4.55 (b), where ΔP_m is the maximum amplitude of the modulated output and ϕ is a phase difference between the input signal (modulating current) and the output signal (modulated optical power).

Under AC operation, ΔP_m , of course, depends on ΔI_m and increases with ΔI_m . However, we are interested in the frequency response, that is, $\Delta P_m(f)$ vs. the modulation frequency f when ΔI_m is kept constant. Suppose that the output signal power at very low frequencies (ideally DC) is $\Delta P_m(0)$. What is of interest is $\Delta P_m(f)/\Delta P_m(0)$ as a function of frequency f , which is shown in Figure 4.55 (c) with some typical values. As can be seen, $\Delta P_m(f)/\Delta P_m(0)$ vs. f characteristic is

⁴⁰ RCLEDs were first realized by Fred Schubert and coworkers at AT&T Bell Laboratories, Murray Hill, NY in 1992, and reported in *Appl. Phys. Letts.*, 60, 921, 1992.

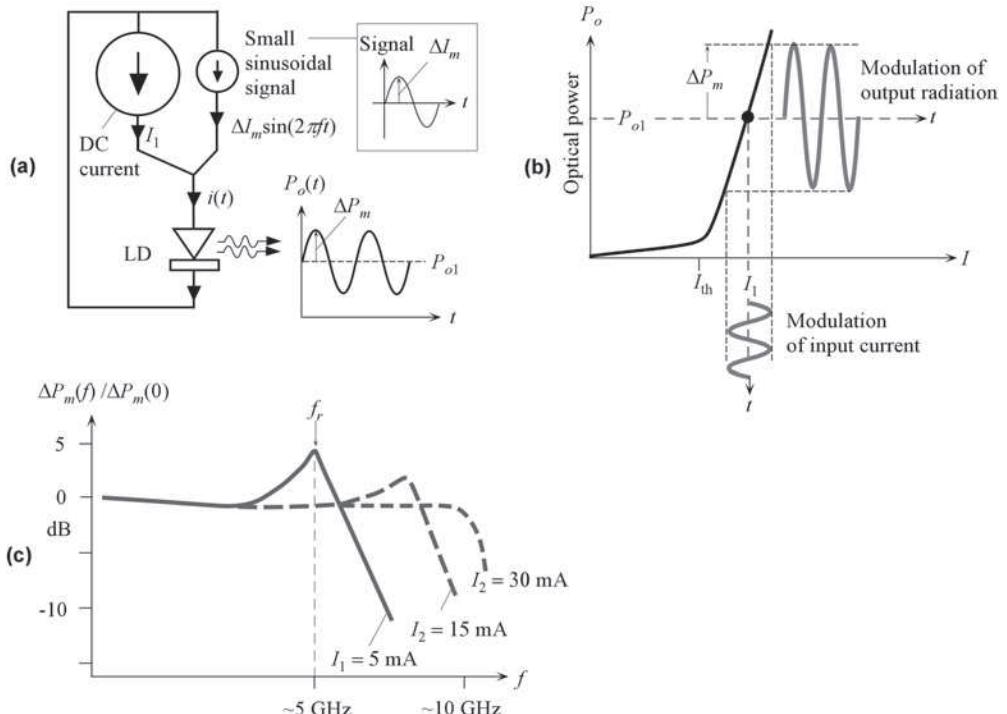


FIGURE 4.55 (a) The LD is biased by passing a DC current I_1 and a sinusoidal current modulation $\Delta I_m \sin(2\pi ft)$ is superimposed on the DC current. (b) P_o vs. I characteristics of a LD showing how a superimposed AC current on top of a DC current I_1 modulates the output optical power. The optical power output has a DC value of P_{o1} about which the power varies sinusoidally with a maximum amplitude ΔP_o . (c) The frequency response $\Delta P_m(f)/\Delta P_m(0)$ vs. frequency has a relaxation oscillation peak at a certain frequency f_r , which depends on the diode current or the output power.

relatively flat until high frequencies, and then exhibits a prominent peak at a certain frequency called the **relaxation oscillation frequency** f_r , as marked in Figure 4.55 (c). f_r depends on the LD characteristics, and is given by⁴¹

$$f_r \approx \frac{1}{2\pi(\tau\tau_{ph})^{1/2}} \left(\frac{I_1}{I_{th}} - 1 \right)^{1/2} \quad (4.18.1)$$

where τ is the effective carrier recombination time, which represents the radiative and nonradiative processes acting in parallel; τ_{ph} , as before, is the *photon cavity lifetime*, the average time it takes a photon to be lost from the laser diode cavity (by absorption inside the cavity and emission through the cavity facets); I_1 is the operating diode bias current; and I_{th} is the usual threshold current. If nonradiative recombination is negligible, then $\tau = \tau_r$, the radiative lifetime. The characteristic times τ and τ_{ph} are typically in the nano- and picosecond range so that for I_1 , roughly twice I_{th} , f_r is a few GHz. LDs are usually modulated below f_r .

The frequency response of a LD also depends on the operating current I_1 or the DC output power P_{o1} . As I_1 increases, the bandwidth increases and the magnitude of the peak at f_r decreases, as apparent in Figure 4.55 (c).

⁴¹The derivations for the equations in this section can be found in P. Bhattacharya, *Semiconductor Optoelectronic Devices*, 2nd Edition (Prentice-Hall, New York, 1993), Ch. 7. Both Eqs. (4.18.1) and (4.18.2) are reasonable approximations.

The peak in the frequency response at f_r is typical of a system that has two energy storage elements, such as an inductor L and a capacitor C that are coupled. We can view the LD as made up of two subsystems. One subsystem is the electrical subsystem that has the current and the injected electrons, and the other, the optical subsystem, that has the optical cavity and the coherent radiation. Near f_r , the energy associated with the injected electrons in the active region in the electrical subsystem is transferred back and forth to and from the radiation (photons) within the cavity in the optical subsystem. This energy transfer is mediated by stimulated emissions because both photons and electrons are involved in stimulated transitions. The electron and photon concentrations are coupled through the stimulated emission process. The phase difference ϕ between the output and the input signals arises from the delay in getting the coherent photon output keeping up with the electron injection. There is a delay associated with the recombination process, which is represented by τ , and then there is another delay associated with getting the photons out from the cavity τ_{ph} . The radiation output lags the input current.

Consider a LD that is excited by a step current pulse that changes from $I_1 (< I_{th})$ to $I_2 (> I_{th})$ as shown in Figure 4.56 (a). Assume initially I_1 is very small. The lasing radiation output begins only after a **delay time** t_d , and reaches a steady state value after a few damped oscillations as illustrated in Figure 4.56 (b). The delay time t_d is the time it takes to inject sufficient electrons into the active region to achieve population inversion, and sufficient gain to start the lasing oscillations to build up. Suppose that initially the bias on the LD is I_1 , less than I_{th} , as in Figure 4.56 (a). Thus, the diode current changes from I_1 to I_2 . The delay time t_d before any coherent radiation output appears is approximately given by

$$t_d \approx \tau \ln \left[\frac{I_2}{I_2 - (I_{th} - I_1)} \right] \quad (4.18.2)$$

Diode
laser
output
delay time

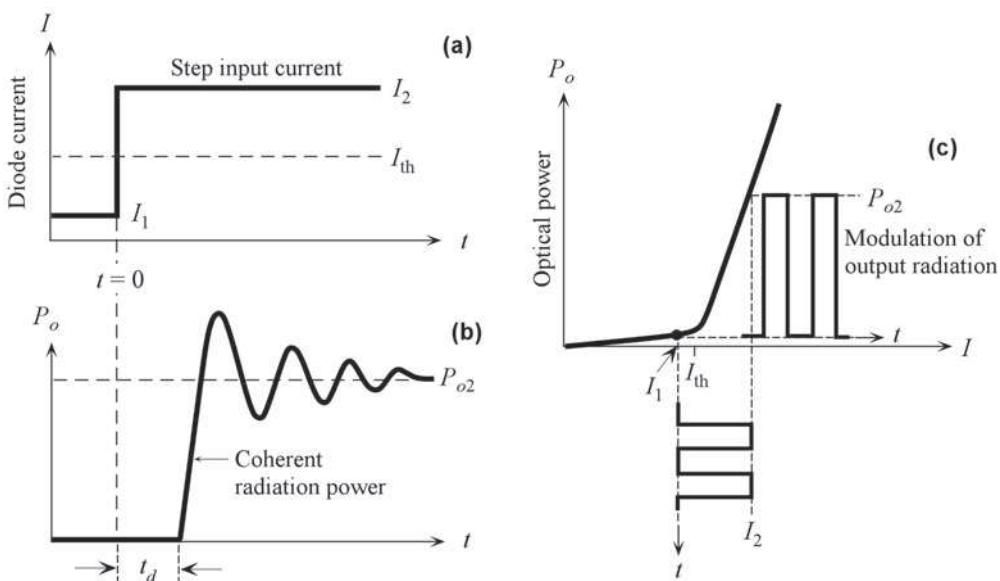


FIGURE 4.56 (a) The Diode current is applied suddenly at $t = 0$ as a step from I_1 (below I_{th}) to I_2 (above I_{th}). (b) The transient response of the output power, P_o , as a function of time. There is a delay t_d before any output appears. P_{o2} is the steady state output, after the relaxation oscillations have died out, after a few cycles. (c) The laser diode is biased at I_1 near I_{th} for faster switching. The input current is a pulse of magnitude $(I_2 - I_1)$, which generates an output P_{o2} . (Ideal response in which time delays have been neglected, as would be the case under sufficiently low repetition rates.)

where τ is the overall effective recombination time in the recombination region around the threshold operating conditions and $(I_{\text{th}} - I_1)$ is the current needed to reach I_{th} . If the LD is not biased, the delay time will be longer. We can reduce the delay time by biasing the LD at I_1 just below I_{th} , as shown in Figure 4.56 (c), so that $(I_{\text{th}} - I_1)$ is negligible, then t_d is as short as possible and roughly the same as τ , the recombination time. The damped relaxation oscillations in the output in Figure 4.56 (b) have the frequency f_r , which should not be surprising, given the above explanation for f_r in terms of two energy storage subsystems in the LD. Usually, the output reaches a steady state P_{o2} after a few oscillations.

Generally, under direct modulation, LD bandwidths are wider than conventional LEDs, and LDs are the obvious choice in high-speed optical communications, given also their much narrower spectral width. There is also a disadvantage to directly modulating LDs by the diode current. The modulation processes modifies the refractive index of the gain medium, and hence causes the radiation frequency to shift, called **frequency chirping**, during the modulation period, over the time of a digital bit. The reason is that the refractive index n , as mentioned above, depends on the injected electron concentration, which is controlled by the current. The frequency chirp can be eliminated by operating the LD as a CW laser, and using a high-speed external modulator. Some electro-optic external modulators, such as Mach-Zhender modulators, as discussed in Chapter 6, have very fast switch times and can easily operate at 40 Gb s^{-1} or above.

4.19 HOLOGRAPHY

Holography is a technique of reproducing three-dimensional optical images of an object by using a highly coherent radiation from a laser source. Although holography was invented by Denis Gabor in 1948, prior to the availability of highly coherent laser beams, its practical use and popularity increased soon after the commercial development of lasers.

The principle of holography is illustrated in Figures 4.57 (a) and (b). The object, a cat, is illuminated by a highly coherent beam as normally would be available from a laser⁴² as in Figure 4.57 (a). The laser beam has both *spatial* and *temporal coherence*. Part of the coherent wave is reflected from a mirror to form a **reference beam** E_{ref} and travels toward a fine-grained photographic plate. The waves that are reflected from the cat, E_{cat} , will have both amplitude and phase variations that represent the cat's surface (topology). If we were looking at the cat, our eyes would register the wavefront of the reflected waves E_{cat} . Moving our head around, we would capture different portions of the reflected wave and we would see the cat as a three-dimensional object.

The reflected waves from the cat (E_{cat}) is made to interfere with a reference wave (E_{ref}) at the photographic plate and give rise to a complicated *interference pattern* that depends on the magnitude and phase variation in E_{cat} . The recorded interference pattern in the photographic film (after processing) is called a **hologram**. It contains all the information necessary to reconstruct the wavefront E_{cat} reflected from the cat and hence produce a three-dimensional image.

To obtain the three-dimensional image, we have to illuminate the hologram with the reference beam E_{ref} alone as in Figure 4.57 (b). Most of the beam goes right through but some of it becomes *diffracted* by the interference pattern in the hologram. One diffracted beam is an exact replica of the original wavefront E_{cat} from the cat. The observer sees this wavefront as if the waves were reflected from the original cat and registers a three-dimensional image of the cat. This is the **virtual image**. We know that first-order diffraction from a grating has to satisfy the

⁴²A short pulse of laser light, such as from a ruby laser, will not harm the cat; this is how holograms of people are taken. The first successful three-dimensional laser holography was carried out by Emmett Leith and Juris Upatneiks (University of Michigan) in 1960. In 1979 President Jimmy Carter awarded Emmett Leith with the National Medal of Science for his contributions to optics.

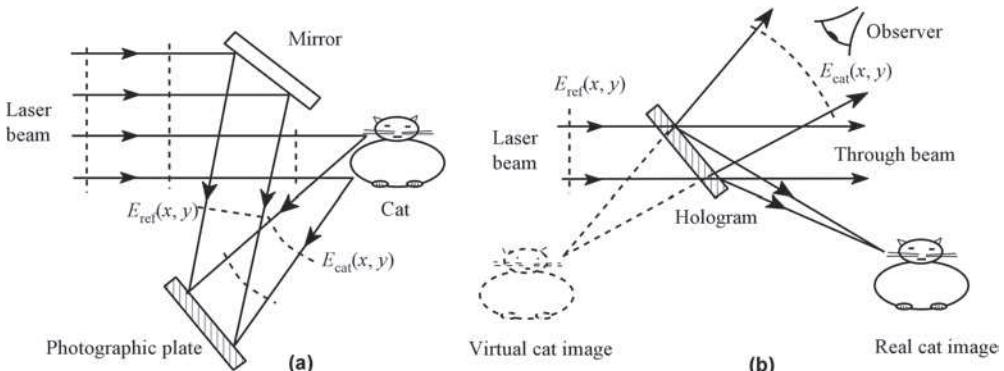


FIGURE 4.57 A highly simplified illustration of holography. (a) A laser beam is made to interfere with the diffracted beam from the subject to produce a hologram. (b) Shining the laser beam through the hologram generates a real and a virtual image.

Bragg condition, $d \sin \theta = m\lambda$ ($m = \pm 1$) where λ is the wavelength and d is the periodicity of the grating, which we took as the separation of slits in a ruled grating in Chapter 1. We can qualitatively think of a diffracted beam from one locality in the hologram as being determined by the local separation d between interference fringes in this region. Since d changes in the hologram, depending on the interference pattern produced by E_{cat} , the whole diffracted beam depends on E_{cat} and the diffracted beam wavefront is an exact scaled replica of E_{cat} . Just as normally there would be another diffracted beam on the other side of the zero-order (through) beam, there is a second image, called the **real image**, as in Figure 4.57 (b), which is of lower quality. (It may help to imagine what happens if the object consists of black and white stripes. We would then obtain periodic dark and bright interference fringes in the hologram. This periodic variation is just like a diffraction grating; the exact analysis is more complicated.)

Holography can be explained by the following highly simplified analysis. Suppose that the photographic plate is in the xy plane in Figure 4.57. Assume that the reference wave can be represented by

$$E_{\text{ref}}(x, y) = U_r(x, y)e^{j\omega t} \quad (4.19.1)$$

where $U_r(x, y)$ is its amplitude, generally a complex number that includes magnitude and phase information.

The reflected wave from the cat will have a complex amplitude that contains the magnitude and phase information of the cat's surface, that is,

$$E_{\text{cat}}(x, y) = U(x, y)e^{j\omega t} \quad (4.19.2)$$

where $U(x, y)$ is a complex number. Normally we would look at the cat and construct the image from the wavefront in Eq. (4.19.2). We thus have to reconstruct the wavefront in Eq. (4.19.2), *i.e.*, obtain $U(x, y)$.

When E_{ref} and E_{cat} interfere, the “brightness” of the photographic image depends on the intensity I and hence on

$$I(x, y) = |E_{\text{ref}} + E_{\text{cat}}|^2 = |U_r + U|^2 = (U_r + U)(U_r^* + U^*)$$

where we have used the usual definition that the square of the magnitude of a complex number is the product with its complex conjugate (indicated by an asterisk). Thus,

$$I(x, y) = UU^* + U_rU_r^* + U_r^*U + U_rU^* \quad (4.19.3)$$

This is the pattern on the photographic plate. The first and second terms are the intensities of the reflected and reference waves. It is the third and fourth terms that contain the information on the magnitude and phase of $U(x, y)$.

We now illuminate the hologram, $I(x, y)$, with the reference beam, which has an amplitude U_r as in Eq. (4.19.1), so that the transmitted light wave has a complex magnitude $U_t(x, y)$ proportional to $U_r I(x, y)$

$$U_t \propto U_r I(x, y) = U_r(UU^* + U_r U_r^* + U_r^* U + U_r U_r^*)$$

i.e.,

$$U_t \propto U_r(UU^* + U_r U_r^*) + (U_r U_r^*)U + U_r^2 U^* \quad (4.19.4)$$

or

$$U_t \propto a + bU(x, y) + cU^*(x, y)$$

where a , b , and c are appropriate constants. The first term $a = U_r(UU^* + U_r U_r^*)$ is the through beam. The second and third terms represent the diffracted beams.⁴³ Since $b = (U_r U_r^*)$ is a constant (intensity of the reference beam), the second term is a scaled version of $U(x, y)$ and represents the original wavefront amplitude from the cat, i.e., includes the magnitude and phase information. We have effectively reconstructed the original wavefront and the observer will see a three-dimensional view of the cat within the angle of original recording; $bU(x, y)$ is the **virtual image**. The third term $cU^*(x, y)$ is the **real image** and notice that its amplitude is the complex conjugate of $U(x, y)$; it is called the **conjugate image**. It is apparent that holography is a method of **wavefront reconstruction**. It should be mentioned that the observer will always see a positive image whether a positive or a negative image of the hologram is used. A negative hologram simply produces a 180° phase shift in the field of the transmitted wave and the eye cannot detect this phase shift.

Dennis Gabor (1900–1979), inventor of holography, was a Hungarian-born British physicist who published his holography invention in *Nature* in 1948 while he was at Thomson-Houston Co. Ltd, at a time when coherent light from lasers was not yet available. He was subsequently a professor of applied electron physics at Imperial College, University of London. He received the Nobel Prize in Physics in 1971 for his invention of holography. His Nobel Lecture given in 1971 provides an excellent insight into the development of holography, and is accessible from the Nobel website. (© Linh Hassel/AGE Fotostock.)



⁴³This is a simple factual statement as the mathematical treatment above has not proved this point.

Questions and Problems

- 4.1 Pumping of three-level laser systems** The 3-level Ruby laser system is shown in Figure 4.2 (a). If transitions from E_3 to E_2 are very fast, and the spontaneous decay time from E_2 to E_1 is τ_{sp} , obtain an expression for the minimum pumping power needed (the power that must be absorbed by the laser medium) for population inversion ($N_2 > N_1$).

Consider a 60 mm long ruby crystal with 6 mm diameter, doped with $2 \times 10^{20} \text{ Cr}^{3+} \text{ ions cm}^{-3}$. The spontaneous decay time is 3 ms. Using the above-obtained expression, find the minimum power that must be absorbed if the excitation is to the band of energies in the blue around ($\sim 3.1 \text{ eV}$) in Figure 4.3 (a).

- 4.2 Neodymium-doped glass laser** Consider a phosphate glass rod, with a refractive index $n = 1.50$, that has been doped with $3 \times 10^{20} \text{ Nd}^{3+} \text{ ions cm}^{-3}$. The glass rod is 10 cm long and 1 cm in diameter. We can use Figure 4.4 to represent the operation of the Nd^{3+} -glass laser. The lasing emission E_2 to E_1 is at 1054 nm. The pump wavelength is 808 nm, and takes the Nd^{3+} from E_0 to E_3 . E_1 is roughly 0.26 eV above E_0 . The linewidth of the output spectrum $\Delta\lambda$ is about 28 nm. E_2 to E_1 emission cross-section is $3 \times 10^{-20} \text{ cm}^2$ and the spontaneous decay lifetime is 300 μs . (a) Calculate the maximum possible gain coefficient in the medium. (b) What percentage of the input energy is lost as heat? (c) What is the spectral spread (linewidth) in terms of photon energy?

- 4.3 Erbium doped fiber amplifier** A glass fiber has been doped with Er^{3+} ions with doping concentration $8 \times 10^{18} \text{ Er}^{3+} \text{ ions cm}^{-3}$. Emission and absorption cross-sections of this EDFA at a wavelength of 1550 nm are found to be $3 \times 10^{-21} \text{ cm}^2$ and $2 \times 10^{-21} \text{ cm}^2$, respectively. The Er^{3+} doped fiber is pumped from a 1480 nm laser diode. Obtain the maximum gain coefficient in terms of dB m^{-1} of this medium at 1550 nm. You may assume that the maximum gain coefficient is obtained when there is full population inversion ($N_1 = 0$).

- 4.4 Erbium-doped fiber amplifier** Consider an EDFA that is pumped at 1480 nm. Let N_1 and N_2 be the concentrations of Er^{3+} at E_2 and E_1 manifolds in Figure 4.7 (b). $N_1 + N_2 = N = \text{Concentration of } \text{Er}^{3+} \text{ ions, taken as } 10^{19} \text{ cm}^{-3}$. When transparency is achieved (no net absorption nor gain), $g = 0$, and $N_1 = N_{1T}$ and $N_2 = N_{2T}$. Show that

$$N_{2T} = N\sigma_{ab}(v)/[\sigma_{ab}(v) + \sigma_{em}(v)]$$

where $\sigma_{ab}(v)$ and $\sigma_{em}(v)$ are the absorption and emission cross-sections that depend on the frequency (wavelength) of interest. The 1480-nm pump generates N_{2T} of Er^{3+} in the manifold 2 in Figure 4.7 (b). The spontaneous decay time from E_2 to E_1 is τ_{sp} . The pump must accumulate the Er^{3+} ions at E_2 before they can decay. Show that the **minimum pump power needed for transparency** is

$$P_{\text{pump}} \approx ALhv_p N\sigma_{ab}/\tau_{sp}(\sigma_{ab} + \sigma_{em})$$

where v_p is the pump frequency, A the pump core area, and L is the fiber length. Calculate the minimum pump power needed for an EDFA if $N = 10^{19} \text{ cm}^{-3}$, $\tau_{sp} \approx 10 \text{ ms}$, core diameter is $4 \mu\text{m}$, the fiber length is 7 m, and the transparency is at 1550 nm where emission and absorption cross-sections are $4 \times 10^{-21} \text{ cm}^2$ and $3 \times 10^{-21} \text{ cm}^2$, respectively. What would be the power needed for pumping all Er^{3+} ions to E_2 ?

- 4.5 Erbium-doped fiber amplifier** Consider an EDFA with an Er doping concentration of $2 \times 10^{19} \text{ cm}^{-3}$ and pumped at 1480 nm. The fiber length is 3 m. Find the small signal gain G at two wavelengths at 1550 and 1570 nm corresponding to 100% population inversion.

- 4.6 Erbium-doped fiber amplifier** Consider a 5 m long EDFA with Er^{3+} doping concentration of $1 \times 10^{19} \text{ cm}^{-3}$, core diameter of $8 \mu\text{m}$. The spontaneous decay time τ_{sp} from E_2 to E_1 is 10 ms. The fiber is pumped at 980 nm from a laser diode. The pump power coupled into the EDFA fiber is 40 mW. What is the fiber length that will absorb the pump radiation when the confinement factor is 80%? Find the small signal gain at 1550 nm corresponding to 80% population inversion.

- 4.7 Erbium-doped fiber amplifier** What is the maximum input signal power beyond which saturation effects take place for an EDFA having a small signal gain (G) of 30 dB, pumped at 80 mW at 980 nm?

- 4.8 He-Ne Laser** The He-Ne laser system energy levels can be quite complicated, as shown in Figure 4.58. There are a number of lasing emissions in the laser output in the red (632.8 nm), green (543.5 nm), orange (612 nm), yellow (594.1 nm), and in the infrared regions at $1.52 \mu\text{m}$ and $3.39 \mu\text{m}$, which give the He-Ne laser its versatility. The pumping mechanism for all these lasers operations is the same, energy transfer from excited He atoms to Ne atoms by atomic collisions in the gas discharge tube. There are two excited states for the He

atom in the configuration $1s^12s^2$. The configuration with electron spins parallel has a lower energy than that with opposite electron spins as indicated in Figure 4.58. These two He states can excite Ne atoms to either the $2p^55s^1$ or $2p^54p^1$. There is then a population inversion between these levels and the $2p^53p^1$ and between $2p^55s^1$ and $2p^54p^1$, which leads to the above lasing transitions. Generally, we do not have a single discrete level for the energy of a many-electron atom of given n, ℓ configuration. For example, the atomic configuration $2p^53p^1$ has ten closely spaced energy levels resulting from various different values of m_l and m_s that can be assigned to the sixth excited electron ($3p^1$) and the remaining electrons ($2p^5$) within quantum mechanical rules. Not all transitions to these energy levels are allowed as photon emission requires quantum number selection rules to be obeyed.

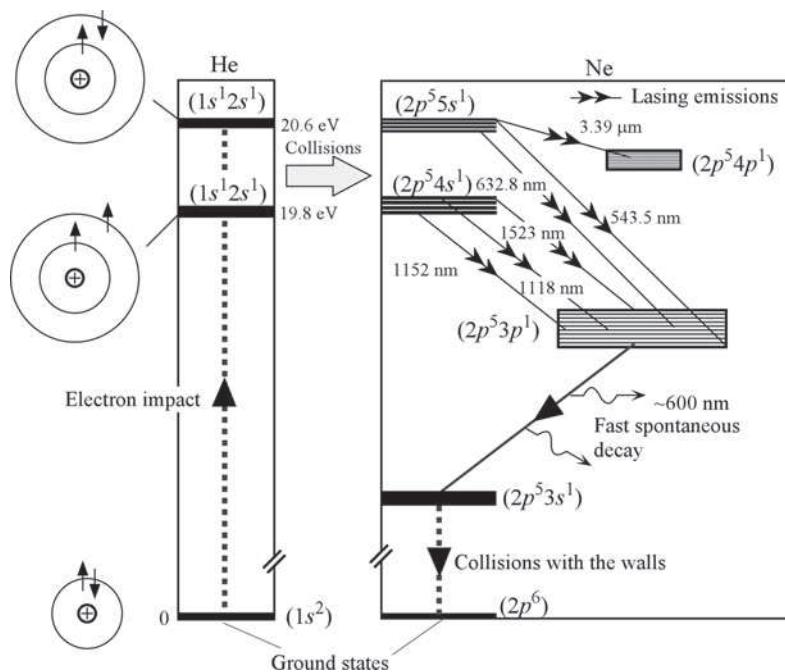


FIGURE 4.58 Various lasing transitions in the He-Ne laser.

- (a) Table 4.8 shows some typical commercial He-Ne laser characteristics (within 30–50%) for various wavelengths. Calculate the overall efficiencies of these lasers.

TABLE 4.8 Typical commercial He-Ne laser characteristic

Wavelength (nm)	543.5	594.1	612	632.8	1523
	Green	Yellow	Orange	Red	Infrared
Optical output power (mW)	1.5	2	4	5	1
Typical current (mA)	6.5	6.5	6.5	6.5	6
Typical voltage (V)	2750	2070	2070	1910	3380

- (b) The human eye is at least twice as sensitive to orange color light than for red. Discuss typical applications where it is more desirable to use the orange laser.
- (c) The 1523-nm emission has potential for use in optical communications by modulating the laser beam externally. By considering the spectral line width ($\Delta\nu \approx 1400\text{ MHz}$), typical powers, and stability, discuss the advantages and disadvantages of using a He-Ne laser over a semiconductor diode.

- 4.9 The Ar-ion laser** The argon-ion laser can provide powerful CW visible coherent radiation of several watts. The laser operation is achieved as follows: The Ar atoms are ionized by electron collisions in a high current electrical discharge. Further multiple collisions with electrons excite the argon ion, Ar^+ , to a group of $4p$ energy levels ~ 35 eV above the atomic ground state as shown in Figure 4.59. Thus a population inversion forms between the $4p$ levels and the $4s$ level which is about 33.5 eV above the Ar atom ground level. Consequently, the stimulated radiation from the $4p$ levels down to the $4s$ level contains a series of wavelengths ranging from 351.1 nm to 528.7 nm. Most of the power, however, is concentrated, approximately equally, in the 488 nm and 514.5 nm emissions. The Ar^+ ion at the lower laser level ($4s$) returns to its neutral atomic ground state via a radiative decay to the Ar^+ ion ground state, followed by recombination with an electron to form the neutral atom. The Ar atom is then ready for “pumping” again.

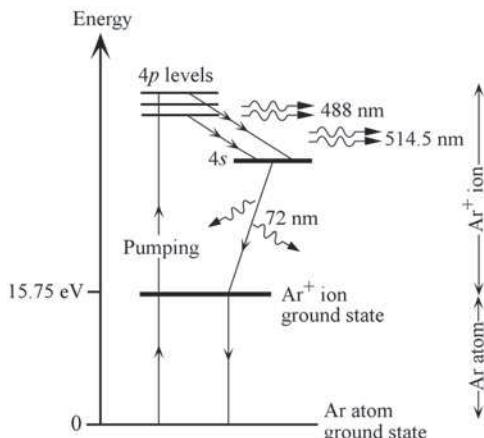


FIGURE 4.59 The Ar^+ -ion laser energy levels.

- (a) Calculate the energy drop involved in the excited Ar^+ ion when it is stimulated to emit the radiation at 514.5 nm.
- (b) The Doppler broadened linewidth of the 514.5-nm radiation is about 3500 MHz ($\Delta\nu$) and is between the half-intensity points. Calculate the Doppler broadened width $\Delta\lambda$ in the wavelength.
- (c) Estimate the operation temperature of the argon ion gas; give the temperature in $^\circ\text{C}$.
- (d) In a particular argon-ion laser the discharge tube, made of Beryllia (Beryllium Oxide), is 30 cm long and has a bore of 3 mm in diameter. When the laser is operated with a current of 40 A at 200 V DC, the total output power in the emitted radiation is 3 W. What is the efficiency of the laser?

- 4.10 He-Ne laser** A particular commercial He-Ne laser operating at 632.8 nm has a tube that is 40 cm long. The operating temperature is 130°C .

- (a) Estimate the Doppler broadened linewidth ($\Delta\lambda$) in the output spectrum.
- (b) What are the mode number m values that satisfy the resonant cavity condition? How many modes are therefore allowed?
- (c) What is the separation $\Delta\nu_m$ in the frequencies of the modes? What is the mode separation $\Delta\lambda_m$ in wavelength?
- (d) Show that if during operation, the temperature changes the length of the cavity by δL , the wavelength of a given mode changes by $\delta\lambda_m$,

$$\delta\lambda_m = (\lambda_m/L)\delta L$$

Given that typically a glass has a linear expansion coefficient $\alpha \approx 1 \times 10^{-6} \text{ K}^{-1}$, calculate the change $\delta\lambda_m$ in the output wavelength (due to one particular mode) as the tube warms up from 20°C to 130°C , and also per degree change in the operating temperature. Note that $\delta L/L = \alpha\delta T$, and $L' = L[1 + \alpha(T' - T)]$. Change in mode wavelength $\delta\lambda_m$ with the change δL in the cavity length L is called **mode sweeping**.

- (e) How do the mode separations $\Delta\nu_m$ and $\Delta\lambda_m$ change as the tube warms up from 20°C to 130°C during operation?
 - (f) How can you increase the output intensity from the He-Ne laser?
- 4.11 Doppler broadening** Consider a He-Ne laser with Doppler broadened linewidth of 2000 MHz. What should be the length of the resonator cavity so that a single mode is found to oscillate? You may assume that there is an oscillating mode at the line centre.

- 4.12 Einstein coefficients and critical photon concentration** $\rho(v)$ is the energy of the electromagnetic radiation per unit volume per unit frequency due to photons with energy $h\nu = E_2 - E_1$. Suppose that there are N_{ph} photons per unit volume. Each has an energy $h\nu$. The frequency range of emission is $\Delta\nu$. Then, $\rho(v) = (N_{\text{ph}}h\nu)/\Delta\nu$. In case of an Ar-ion laser system with the emission wavelength 488 nm and the linewidth in the output spectrum is 6×10^9 Hz (between half intensity points), estimate the photon concentration necessary to achieve more stimulated emission than spontaneous emission.
- 4.13 Photon concentration in a gas laser** The Ar-ion laser has a strong lasing emission at 488 nm. The laser tube is 1 m in length, and the bore diameter is 3 mm. The output power is 1 W. Assume that most of the output power is in the 488-nm emission. Assume that the tube end has a transmittance T of 0.1. Calculate the photon output flow (number of lasing photons emitted from the tube per unit time), photon flux (number of lasing photons emitted per unit area per unit time), and estimate the order of magnitude of the steady state photon concentration (at 488 nm) in the tube (assume that the gas refractive index is approximately 1).
- 4.14 He-Ne Laser and the photon cavity lifetime** A He-Ne laser operating at 632.8 nm has a tube that is 40 cm long and end mirrors with reflectances 99.0% and 99.9%. The tube diameter is 0.80 mm and the output power is 1 mW. The linewidth $\Delta\nu$ is 1 GHz and the loss coefficient α_s is 0.06 m^{-1} . The spontaneous emission lifetime is 125 ns. (a) Calculate the threshold gain. (b) Calculate the threshold population inversion. (c) Calculate the photon cavity lifetime. (d) Calculate the photon concentration in the cavity.
- 4.15 Threshold population Nd:YAG laser** The emission wavelength of a Nd:YAG laser is 1064 nm, the optical gain linewidth is about 100 GHz, the spontaneous lifetime (τ_{sp}) of the upper laser level is 200 μs . The loss coefficient as is 0.2 m^{-1} . A YAG rod of 10 cm in length and has its ends coated to achieve reflectances of 99.9% and 92% (emission end). What is the threshold concentration of pumped Nd-ions (Nd^{3+})?
- 4.16 Threshold gain and population inversion in an Ar-ion laser** The emission wavelength of an Ar-ion gas laser is 488 nm. The tube length $L = 80 \text{ cm}$, tube mirror reflectances are approximately 99.0% and 96.0%. The linewidth $\Delta\nu = 2 \text{ GHz}$, the loss coefficient is $\alpha_s \approx 0.15 \text{ m}^{-1}$, spontaneous decay time constant $\tau_{\text{sp}} = 1/A_{21} \approx 10 \text{ ns}$, $n \approx 1$. Estimate the threshold population inversion.
- 4.17 Fabry-Perot optical resonator in gas lasers** An idealized He-Ne laser optical cavity of length 50 cm has 99% reflectance $R_1 = R_2$ at both ends. Find out the separation $\Delta\lambda$ of the modes in wavelength and the spectral width $\Delta\lambda$ for this laser.
- 4.18 Homogeneous broadening**
- Consider the emission of radiation from an isolated atom in which an electron transits down from E_2 to E_1 . Suppose that the lifetime of the atom at E_2 is τ_2 . We will assume that the lifetime of the state E_1 is very long and neglect it. The emitted radiation's optical field is not simply sinusoidal oscillations, $\sin(\omega_o t)$ type, that would go on forever since we know this is not possible. The optical field's amplitude is decayed exponentially to give the radiation a finite temporal length, which results in a finite spectral width. Suppose that the field follows $E(t) = E_0 \exp(-t/2\tau_2) \times \cos(\omega_o t)$ for all $t > 0$, and zero for $t < 0$. In practice, $1/\tau_2 \ll \omega_o$. The Fourier transform (FT) $E_{\text{FT}}(\omega)$ of $E(t)$ is approximately given by
- $$E_{\text{FT}}(\omega) = \text{FT of } E(t) \approx \frac{E_0/4\pi}{(2\tau_2)^{-1} + j(\omega - \omega_o)}$$
- where we neglected the additional small term that has $j(\omega + \omega_o)$ in the denominator. Show that the light intensity is maximum at $\omega = \omega_o$ and that the spectrum has an FWHM width $\Delta\omega_{\text{FWHM}} = 1/\tau_2$. What is your conclusion? (The spectrum is called a *Lorentzian lineshape*.)
- Consider the two-energy-level system shown in Figure 4.5. The rate equation for the change in the concentration N_2 of atoms at E_2 from Eqs. (4.2.1) and (4.2.2) is
- $$dN_2/dt = -A_{21}N_2 - B_{21}\rho(v)N_2 - B_{12}\rho(v)N_1$$
- Since $E_2 > E_1$, we can take the equilibrium (unpumped) value of N_2 as zero. Suppose we suddenly pump the atoms to the energy level E_2 and we assume the radiation density $\rho(v)$ remains small. From the rate equation, what is the time evolution of $N_2(t)$? What does A_{21} represent? What is the time evolution of the total emitted optical power? How do your results relate to part (a)? What is your conclusion?
- 4.19 Homogeneous and inhomogeneous broadening** The expressions for homogeneous and inhomogeneous broadening in photonics are given by the *Lorentzian* and *Gaussian* lineshape functions, *i.e.*,

$$g_L(v) = \frac{\gamma/\pi}{(v - v_o)^2 + \gamma^2} \quad \text{and} \quad g_G(v) = \frac{1}{\pi^{1/2}\Gamma} \exp[-(v - v_o)^2/\Gamma^2]$$

where γ and Γ are characteristic parameters that control the width of the spectrum, for example, for lifetime or natural broadening $1/\gamma = \tau_2$, lifetime of upper state E_2 . Further, these functions are such that when integrated over all frequencies they are unity, *i.e.*, $\int g(v)dv = 1$. (a) Show that for Lorentzian broadening, the FWHM spectral width $\Delta\nu_{\text{FWHM}} = 2\gamma$. (b) Show that for Gaussian broadening, $\Delta\nu_{\text{FWHM}} = 2\Gamma[\ln(2)]^{1/2}$. (c) Suppose that widths $\Delta\nu_{\text{FWHM}}$ of the two distributions were the same, and both have the same central frequency v_o . Plot and compare the two distributions for the following: $v_o = 100\text{ THz}$ and $\Delta\nu_{\text{FWHM}} = 1\text{ GHz}$ from $v = 0$ to $v = v_o \pm 2\Delta\nu_{\text{FWHM}}$.

- 4.20 Natural broadening in gas lasers** In a He-Ne laser, the Doppler broadened linewidth of the lasing emission is 2.0 GHz. For natural broadening, the spectral width $\Delta\nu_{1/2}$ is given by $\Delta\nu_{1/2} \approx (2\pi\tau_2)^{-1}$ where τ_2 is the lifetime of the upper laser level, and we can take $A_{21} = 1/\tau_2$. Given $A_{21} = 3.5 \times 10^{-6}\text{ s}^{-1}$ for the 632.8 nm emission, calculate the natural linewidth of this emission.
- 4.21 Q-switching** The energy of the pulse in a *Q*-switched Nd^{3+} : YAG laser is 20 mJ, and the pulse duration is 10 ns. The pulse repetition rate is 200 kHz. Calculate the peak power in the pulse and the average power emitted by the laser.
- 4.22 Mode-locked lasers** (a) Consider a mode-locked Nd^{3+} :glass laser with emission wavelength 1060 nm. The spectral width of the optical gain curve is 8 THz. The refractive index of the Nd^{3+} :glass laser medium is 1.46. Find the mode-locked pulse width, and the repetition rate, for a 10 cm long the Nd^{3+} glass rod. Find the separation of the pulses in space. (b) Consider a He-Ne laser emitting at 632.8 nm, and mode locked. The tube length is 60 cm. The spectral width of the optical gain curve is 2 GHz. Find the mode-locked pulse width, and the repetition rate. What is the separation of pulses in space?
- 4.23 Photon cavity lifetime and total attenuation** Consider a semiconductor Fabry-Perot optical cavity of length L , end mirrors with reflectances R_1 and R_2 , and attenuation inside the cavity given by $\alpha_s = 20\text{ cm}^{-1}$. The total attenuation coefficient for the optical cavity is denoted by α_t . Find out the photon cavity time for an $\text{In}_{0.60}\text{Ga}_{0.40}\text{As}_{0.85}\text{P}_{0.15}$ FP optical cavity that has a refractive index of 3.7, and a length of 400 μm .
- 4.24 Population inversion in a GaAs homojunction laser diode** Consider the energy diagram of a forward biased GaAs homojunction LD as shown in Figure 4.32 (b). For simplicity we assume a symmetrical device ($n = p$) and also assume that population inversion has been just reached when the conduction band on the n^+ -side overlaps the valence band on the p^+ -side around the center of the depletion region, as illustrated in Figure 4.32 (b), and results in $E_{Fn} - E_{Fp} = E_g$. Estimate the minimum carrier concentration $n = p$ for population inversion in GaAs at 300 K. The intrinsic carrier concentration in GaAs is appropriately $2 \times 10^6\text{ cm}^{-3}$. Assume for simplicity that
- $$n = n_i \exp[(E_{Fn} - E_{Fi})/(k_B T)] \quad \text{and} \quad p = n_i \exp[(E_{Fp} - E_{Fn})/(k_B T)]$$
- (*Note:* The analysis will only be an order of magnitude as the above equations do not hold in degenerate semiconductors. A better approach is to use the Joyce-Dixon equations as can be found in advanced textbooks. The present analysis is an order of magnitude calculation.)
- 4.25 InGaAsP-InP laser** The optical cavity length is 200 μm for a given Fabry-Perot InGaAsP-InP laser diode with emission wavelength 1550 nm and the refractive index of InGaAsP is 3.7. The internal cavity loss (α_s) is 25 cm^{-1} . The optical gain bandwidth (as measured between half intensity points) will normally depend on the pumping current (diode current) but for this problem assume that it is 3 nm. (a) What is the mode integer m of the peak radiation? (b) What is the separation between the modes of the cavity? (c) How many modes are there in the cavity? (d) What is the reflectance at the ends of the optical cavity mode of the InGaAsP crystal?
- 4.26 SQW laser** Consider a SQW (single quantum well) laser which has an ultrathin active InGaAs of bandgap 0.70 eV and thickness 8 nm between two layers of InAlAs which has a bandgap of 1.44 eV. Effective mass of conduction electrons in InGaAs is about $0.04m_e$ and that of the holes in the valence band is $0.44m_e$ where m_e is the mass of the electron in vacuum. Calculate the first and second electron energy levels above E_c and the first hole energy level below E_v in the QW. What is the lasing emission wavelength for this SQW laser? What is this wavelength if the transition were to occur in bulk InGaAs with the same bandgap?
- 4.27 Semiconductor lasers and wavelength shift with temperature** Consider a Fabry-Perot semiconductor laser diode operating at 1550 nm. The active region is a III-V quaternary semiconductor alloy of InGaAsP (but quite close to being InGaAs). The LD has a cavity length of 200 μm . The refractive index of InGaAsP is approximately 3.60 and $d\eta/dT \approx 2.5 \times 10^{-4}\text{ K}^{-1}$, the linear thermal expansion coefficient is $5.8 \times 10^{-6}\text{ K}^{-1}$. The bandgap of InGaAs follows the Varshni equation $E_g = E_{go} - AT^2/(B + T)$ with $E_{go} = 0.950\text{ eV}$, $A = 5 \times 10^{-4}\text{ eV K}^{-1}$, $B = 300\text{ K}$. Find the shift in the optical gain curve, and the emission wavelength for a given mode per unit temperature change. What is your conclusion?

- 4.28 Fabry–Perot optical resonator in semiconductor lasers** Consider a semiconductor Fabry–Perot optical cavity of length 250 μm having end-mirrors of reflectance R . Calculate the cavity mode nearest to the free space wavelength of 1550 nm. If the refractive index of semiconductor is 3.67, calculate the separation of the modes. Calculate the spectral width for $R = 0.95$ and 0.8. What is your conclusion?
- 4.29 GaAs DH laser diode** Consider a GaAs DH laser diode that lases at 850 nm. It has an active layer (cavity) length L of 250 μm . The active layer thickness d is 0.2 μm and the width (W) is 5 μm . The refractive index is 3.6, and the attenuation coefficient α_s inside the cavity is 20 cm^{-1} . The radiative lifetime τ_r in the active region is 3 ns. Find the threshold gain g_{th} , carrier concentration n_{th} , current density J_{th} , and current I_{th} . Find the output optical power at $I = 2I_{\text{th}}$, and the external slope efficiency η_{slope} .
- 4.30 Threshold current and power output from a 1310-nm FP laser diode** Consider a double heterostructure InGaAsP semiconductor laser operating at 1310 nm. The cavity length $L \approx 200\text{ }\mu\text{m}$, width $W \approx 8\text{ }\mu\text{m}$, and $d \approx 0.2\text{ }\mu\text{m}$. The refractive index $n \approx 3.6$. The loss coefficient $\alpha_s \approx 25\text{ cm}^{-1}$ and the direct recombination coefficient $B \approx 2 \times 10^{-16}\text{ m}^3\text{ s}^{-1}$. Assume that the optical confinement factor is 1. Find the threshold gain g_{th} , carrier concentration n_{th} , current density J_{th} and the threshold current I_{th} . Find the output optical power at $I = 1.5I_{\text{th}}$, and the external slope efficiency η_{slope} .
- 4.31 Laser diode efficiencies** Consider a particular commercial AlGaInP power laser diode with an emission wavelength of 630 nm (red). The threshold current at 30°C is 50 mA. The typical operating voltage for this device is 2.5 V. At $I = 100\text{ mA}$. The output optical power is 40 mW.
- Calculate the external QE, external differential QE, external power efficiency, and the slope efficiency of the laser diode. What is the current required for an output of 30 mW?
 - The threshold current at 60°C is measured to be 80 mA. Assuming that $I_{\text{th}} = A \exp(T/T_o)$ where T_o is a characteristic temperature, and given I_{th} at two temperatures, find the characteristic T_o temperature parameter. What is the threshold current at 0°C?
- 4.32 Laser diode efficiencies** Consider an InGaAsP FP laser diode operating at $\lambda = 1330\text{ nm}$ for optical communications. The threshold current is 6 mA. At $I = 25\text{ mA}$, the output optical power is 6 mW and the voltage across the diode is 1.2 V. Calculate external quantum efficiency (QE), external differential QE, power conversion efficiency, and the slope efficiency of the diode. What is the forward diode current that gives an output optical power of 3 mW?
- 4.33 Laser diode efficiencies** Consider a commercial InGaAsP FP laser diode operating at $\lambda = 1550\text{ nm}$ for primary use in fiber-to-home communications. The threshold current is 8 mA. At $I = 25\text{ mA}$, the output optical power is 5 mW, and the voltage across the diode is 1.2 V. Calculate external quantum efficiency (QE), external differential QE, external power efficiency (conversion efficiency), and the slope efficiency.
- 4.34 Laser diode extraction efficiency** Consider an optical cavity with end mirrors R_1 and R_2 and an internal loss coefficient of α_s . The total loss coefficient α_t is given by

$$\alpha_t = \alpha_s + (1/2L)\ln(1/R_1) + (1/2L)\ln(1/R_2)$$

where the second term on the right represents the loss at the cavity end 1, associated with R_1 , and the third is the loss at the cavity end 2, associated with R_2 . Suppose that we are interested in how much light is coupled out from the cavity end 1. The extraction efficiency is then given by

$$\text{Extraction efficiency} = \frac{\text{Loss from cavity end 1}}{\text{Total loss}}$$

that is

$$\eta_{\text{EE}} = (1/2L)\ln(1/R_1)/\alpha_t$$

A semiconductor laser made of GaAs as its active layer has refractive index n is 3.6, the cavity length L is 300 μm , and the internal loss coefficient $\alpha_s = 20\text{ cm}^{-1}$. (a) What is η_{EE} ? (b) Explain the effect when there are no internal losses? (c) Suppose the end 2 had a 100% reflecting dielectric mirror and $\alpha_s = 20\text{ cm}^{-1}$. What is η_{EE} ? (d) Suppose in (c) you set the internal losses to 2. What is η_{EE} ?

- 4.35 Temperature dependence of laser characteristics** The threshold current of a laser diode increases with temperature because more current is needed to achieve the necessary population inversion at higher temperatures. For a particular commercial laser diode, the threshold current, I_{th} , is 60 mA at 27 °C, 45 mA at 0°C, and 90 mA at 57°C. Using these three points, explain the dependence of I_{th} on the absolute temperature. Obtain an empirical expression for it.

- 4.36 UV laser diode** Consider a GaN-based laser diode that emits at a peak wavelength of 360 nm. What is the transition energy ΔE in eV that corresponds to 360 nm? The threshold current is 50 mA and the emitted power is 25 mW at a forward current of 70 mA and a voltage of 5 V. Find the slope efficiency, EQE, EDQE, PCE. What is the power emitted at 70 mA?
- 4.37 Single DFB frequency lasers** An InGaAsP DFB laser operates at 1550 nm. Suppose that the effective refractive index $n \approx 3.5$ and the cavity length is 60 μm . What should be the corrugation period Λ for a first-order grating, $q = 1$? What is Λ for a second-order grating, $q = 2$? How many corrugations are needed for a first-order grating? How many corrugations are there for $q = 2$?
- 4.38 Single frequency DFB lasers** A DFB Laser Diode emits at 1063 nm, and has $d\lambda/dT = 0.07 \text{ nm K}^{-1}$ and $d\lambda/dI = 0.004 \text{ nm mA}^{-1}$. What should be change in the temperature, δT , that gives a wavelength shift equal to the width $\delta\lambda$ of the single-mode output spectrum? What should be the change δI in the current that gives a wavelength shift equal to $\delta\lambda$? What percentage of the operating current, 195 mA, is this change?
- 4.39 VCSEL** Consider the VCSELs in Table 4.6. (a) For each, calculate the power conversion efficiency η_{PCE} , and then average these values. What is the average PCE? (b) For each, using P_o , I , and I_{th} , calculate the slope efficiency and compare with the quoted (observed) value. (c) The emitted power from an STM (single transverse mode) VCSEL is about 1 mW. The diameter is roughly $\sim 10 \mu\text{m}$ as this is a single transverse mode device. Estimate the emitted intensity in kW cm^{-2} .
- 4.40 VCSEL and DBR reflectances** Consider a VCSELs that has 55 pairs of quarter-wave alternating layers of high and low refractive index layers for the lower DBR, and 35 pairs in the DBR on the emission side. Assume that the refractive indices of the DBR, layers are $n_1 = 3.60$ and $n_2 = 3.42$ (4% lower). Take $n_0 = 1$ and $n_3 = 3.6$, and calculate the reflectance of each DBR. (See Example 1.7.2 in Chapter 1.) *Note:* For N pairs of layers, $R = [n_1^{2N} - (n_0/n_3)n_2^{2N}]^2 / [n_1^{2N} + (n_0/n_3)n_2^{2N}]^2$; see Eq. (1.7.3).
- 4.41 SOA** Consider a TW SOA for use at 1550 nm. It has an active region that is $400 \mu\text{m}$ long. The gain of the SOA is quoted as 20 dB. Suppose that the confinement factor Γ is 0.4, and the loss coefficient α_s is 20 cm^{-1} . What is the gain coefficient g of the active region at this amplification? *Note:* The gain $G = \exp[(\Gamma g - \alpha_s)L]$, and assume perfect AR coatings at the facets of the semiconductor, that is, $R_1 = R_2 = 0$.
- 4.42 Laser diode modulation** A GaAs Laser Diode that has the following typical properties. The radiative lifetime $\tau_r \approx 3 \text{ ns}$, non radiative lifetime $\tau_{nr} \approx 60 \text{ ns}$, total attenuation coefficient $\alpha_t \approx 5000 \text{ m}^{-1}$ for the GaAs active layer of length of about $250 \mu\text{m}$ having refractive index n of 3.6. Obtain the relaxation oscillation frequency and hence the bandwidth at bias currents $I_1 = 2I_{\text{th}}$ and $3I_{\text{th}}$? What is the delay time if the LD is biased at $0.8I_{\text{th}}$ and switched by a current to $2I_{\text{th}}$? What is this delay time if the LD is not biased at all?
- 4.43 Laser diode modulation** A laser diode has the radiative recombination time $\tau_r = 3 \text{ ns}$, nonradiative lifetime $\tau_{nr} = 40 \text{ ns}$, and the photon cavity lifetime $\tau_{\text{ph}} = 6 \text{ ps}$. Find the resonance frequency for this laser diode at forward currents $I_1 = 2I_{\text{th}}$ and $3I_{\text{th}}$.



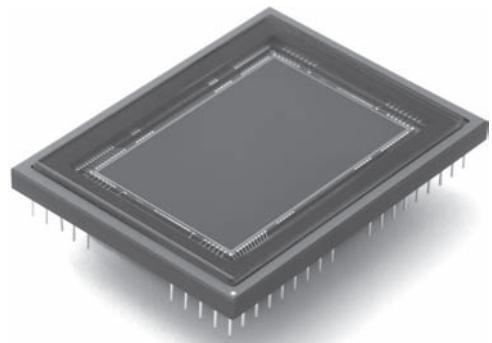
Theodore Maiman's ruby laser. The helical xenon flash surrounds the ruby crystal rod (an Al_2O_3 crystal doped with Cr^{3+} ions) and provides optical pumping of the chromium ions in the ruby rod. (Courtesy of HRL Laboratories, LLC, Malibu, California.)

The detector is like the journalist who must determine what, where, when, which, and how? What is the identity of the particle? Exactly where is it when it is observed? When does the particle get to the detector? Which way is it going? How fast is it moving?

—Sheldon L. Glashow¹



The inventors of the CCD (charge-coupled device) image sensor at AT&T Bell Labs: Willard Boyle (left) and George Smith (right). The CCD was invented in 1969, the first CCD solid state camera was demonstrated in 1970, and a broadcast quality TV camera by 1975. (W. S. Boyle and G. E. Smith, "Charge Coupled Semiconductor Devices," *Bell Syst. Tech. J.*, 49, 587, 1970; Courtesy of Alcatel-Lucent Bell Labs.)



A CCD image sensor. The FTF6040C is a full-frame color CCD image sensor designed for professional digital photography, scientific and industrial applications with 24 megapixels, and a wide dynamic range. Chip imaging area is 36 mm × 24 mm, and pixel size is 6 $\mu\text{m} \times 6 \mu\text{m}$. (Courtesy of Teledyne-DALSA.)

¹ Sheldon L. Glashow, *Interactions* (Warner Books, 1988), p. 101.

5

PHOTODETECTORS AND IMAGE SENSORS

5.1 PRINCIPLE OF THE *pn* JUNCTION PHOTODIODE

A. Basic Principles

Photodetectors convert an incident radiation to an electrical signal such as a voltage or current. In many photodetectors such as photoconductors and photodiodes this conversion is typically achieved by the creation of **free electron hole pairs** (EHPs) by the absorption of photons, that is, the creation of electrons in the conduction band (CB) and holes in the valence band (VB). In some devices, such as pyroelectric detectors, the energy conversion involves the generation of heat that increases the temperature of the device, which changes its polarization and hence its relative permittivity. We will first consider *pn* junction-based photodiode-type devices as these devices are small and have high speed and good sensitivity for use in various optoelectronics applications, the most important of which is in optical communications.

Figure 5.1 (a) shows the simplified structure of a typical *pn* junction **photodiode** that has a p^+ *n* type of junction, that is, the acceptor concentration N_a in the *p*-side is much greater than the donor concentration N_d in the *n*-side. The illuminated side has a window, defined by an annular electrode, to allow photons to enter the device. There is also an **antireflection (AR) coating**, typically Si_3N_4 , to reduce light reflections. The p^+ -side is generally very thin (less than a micron) and is usually formed by planar diffusion into an *n*-type epitaxial layer. Figure 5.1 (b) shows the net space charge distribution across the p^+ *n* junction. These charges are in the **depletion region**, or in the **space charge layer (SCL)**, and represent the exposed negatively charged acceptors in the p^+ -side and exposed positively charged donors in the *n*-side. The depletion region extends almost entirely into the lightly doped *n*-side and, at most, it is a few microns.

The photodiode is normally reverse biased. The applied reverse bias V_r drops across the highly resistive depletion layer width W and makes the voltage across W equal to $V_o + V_r$ where V_o is the built-in voltage. Normally V_r (*e.g.*, 5–20 V) is much larger than V_o (less than 1 V) and the voltage across the SCL is essentially V_r . The field E in the SCL is found by the integration of the net space charge density ρ_{net} in Figure 5.1 (b) across W subject to a voltage difference of $V_o + V_r$, *i.e.*, V_r . The field only exists in the depletion region and is not uniform. It varies across the depletion region as shown in Figure 5.1 (c), where it is maximum at the junction and

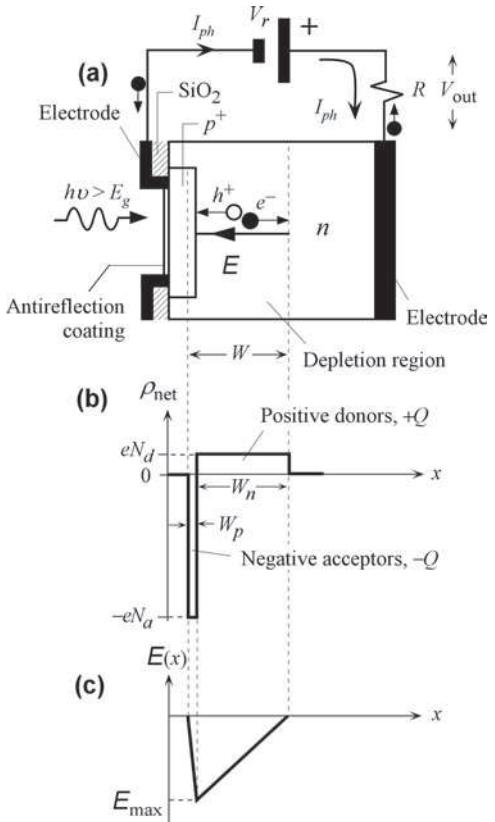


FIGURE 5.1 (a) A schematic diagram of a reverse biased pn junction photodiode. (b) Net space charge across the diode in the depletion region. N_d and N_a are the donor and acceptor concentrations in the p - and n -sides. (c) The field in the depletion region.

penetrates into the n -side. The regions outside the SCL are the **neutral regions** in which there are majority carriers.

When a photon with an energy greater than the bandgap E_g is incident, it becomes absorbed to **photogenerate** a free EHP, that is, an electron in the CB and a hole in the VB. Usually the energy of the photon is such that photogeneration takes place in the depletion layer. The field E in the depletion layer then separates the EHP and drifts them in opposite directions until they reach the neutral regions, as illustrated in Figure 5.1 (a). Drifting carriers generate a current, called the **photocurrent** I_{ph} , in the external circuit that provides the electrical signal. The photocurrent lasts for the duration it takes for the electron and hole to cross the SCL (width, W) and reach the neutral regions. As the electron drifts in the SCL toward the neutral n -side, there is an electron that has come out from n -side and is flowing around the external circuit toward the battery's positive terminal. Similarly, as the hole drifts in the SCL toward the p -side, there is an electron flowing in the external circuit from the battery's negative terminal to the p -side. When the drifting hole reaches the neutral p^+ -region, it recombines with an electron entering the p^+ -side from the external circuit. Similarly, when the drifting electron in the SCL reaches the neutral n -side, it has replenished the electron that had left the n -side for the battery. The photocurrent I_{ph} depends on the number of EHPs photogenerated and the drift velocities of the carriers while they are transiting the depletion layer. Since the field is not uniform and the absorption of photons occurs over a distance that depends on the wavelength, the time dependence of the photocurrent signal cannot be determined in a simple fashion.

As indicated in Figure 5.1 (b), the depletion layer width W_p in the *p*-side is very narrow whereas that on the *n*-side, W_n , is much wider because the amount of total charge in the SCL on the *p*-side must be the same as that on the *n*-side, that is, $N_a W_p = N_d W_n$. Thus, for the p^+n photodiode $W_p \ll W_n$ and $W \approx W_n$. It should be mentioned that the photocurrent in the external circuit is due to the flow of electrons only even though there are both electrons and holes drifting within the device. Suppose that there are N number of EHPs photogenerated. If we were to integrate the photocurrent I_{ph} to calculate how much charge has flowed, we would find an amount of charge that is due to the total number of photogenerated electrons (eN) and not due to both electrons and holes ($2eN$).

B. Energy Band Diagrams and Photodetection Modes

To obtain a better understanding of the *pn* junction detector, we need to consider the energy band diagram. Figure 5.2 (a) shows the *pn* junction reverse biased by V_r , which increases the built-in voltage to $V_o + V_r$. The field in the SCL also increases. The Fermi levels E_{Fn} and E_{Fp} on the *n*- and *p*-sides are separated by V_r . The potential hill, that is, the change in E_c from the E_c on the *n*-side to that on the *p*-side in the SCL is very steep due to the large field in the SCL. The absorption of a photon in the SCL creates an EHP. The electron and hole become separated and drifted by the field. The drift corresponds to the electron rolling down the energy hill (along E_c) toward the *n*-side, whereas the hole rolls down the energy hill toward the *p*-side. (Remember that the hole energy increases in the downward direction.) The drift creates a photocurrent I_{ph} , which is the quantity detected in the external circuit. It lasts for the duration of the drift of the electron and hole.

Photogeneration within a diffusion length to the SCL would also generate a photocurrent as illustrated in Figure 5.2 (b). An EHP is created in the *p*-side within the diffusion length L_e of an electron in this *p*-side. The electron can only reach the SCL by diffusion. The photogenerated electron diffuses (by “random walk”) to the SCL where the internal field then drifts the electron over to the *n*-side. The drift creates the photocurrent. As the electron drifts in the SCL toward the neutral *n*-side, there is an electron flowing in the external circuit toward the battery’s positive terminal; this electron came out from the *n*-side. The photogenerated hole in the *p*-side is neutralized by the flow of an electron from the external circuit into the *p*-side. The photocurrent due to photogeneration in the neutral region is weaker than that due to photogeneration in the depletion region; in the latter, the field separates and drifts the carriers immediately. Photodetector designs prefer the photogeneration process to take place in the depletion region, which is the reason for keeping the p^+ -layer as thin as possible. The photodetection in which there is an applied reverse bias V_r across the *pn* junction represents photodetection with the **photodiode reverse biased**.

Certain photodiodes are operated in the **photovoltaic mode**. There is no external battery connected to the photodiode. First, consider the *pn* junction shorted as in Figure 5.2 (c). Photogeneration in the SCL will again create an electron and a hole and these will now be driven by the built-in field E_o . The potential energy (PE) hill for the electron, change in E_c from the *n*- to the *p*-side, is no longer as steep as it was in the reverse biased *pn* junction. Nonetheless, there is a hill, and the electron will roll down the energy hill toward the *n*-side, that is, drift in the SCL. Similarly, the hole will roll down its own energy hill and drift in the opposite direction. As the electron and hole drift, they generate a photocurrent I_{ph} in the external circuit (remember that the current must be continuous). Notice that I_{ph} is, as before, a current that flows toward the *n*-side but now it is flowing in a short circuit. The situation is not far different than in Figure 5.1 (a) except there is no reverse bias V_r . If the photoexcitation is a very short pulse of light, the photocurrent magnitude is less than that in the reverse biased case (the field is lower and hence carriers

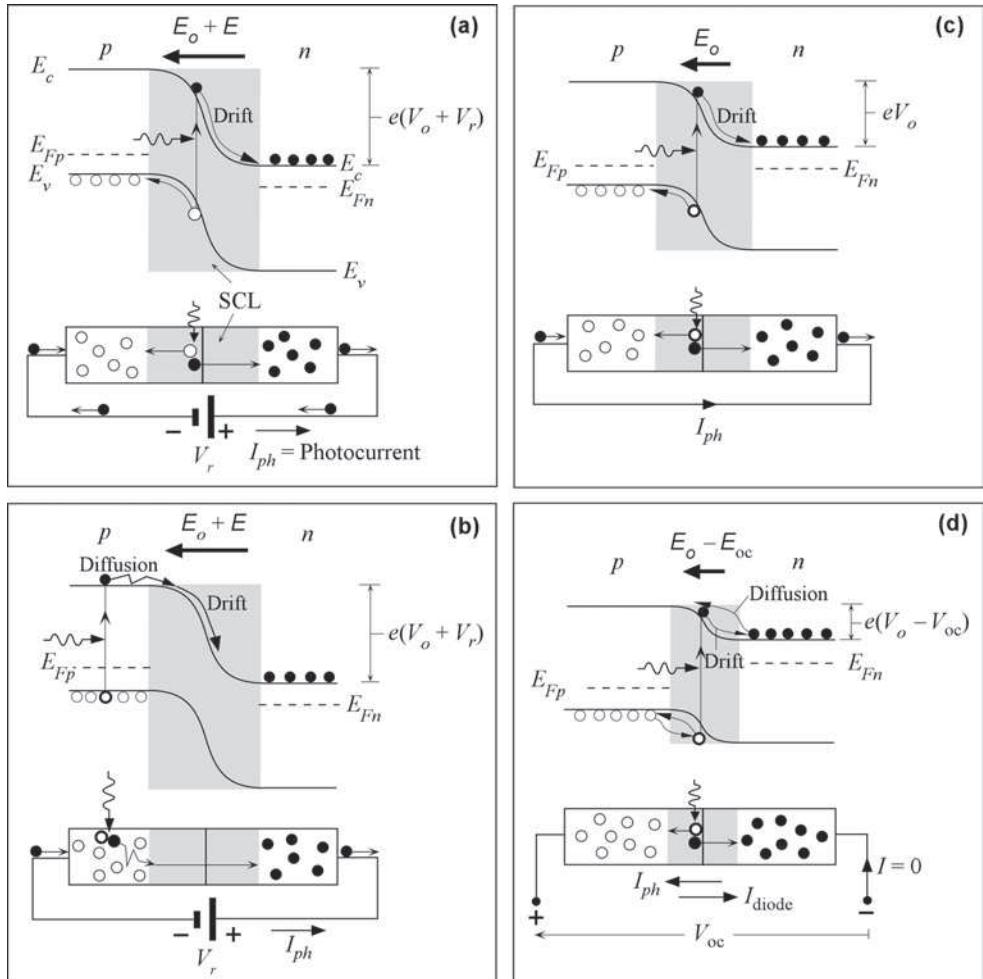


FIGURE 5.2 (a) A reverse biased pn junction. Photogeneration inside the SCL generates an electron and a hole. Both fall in their respective energy hills (electron along E_c and hole along E_o), i.e., they drift, and cause a photocurrent I_{ph} in the external circuit. (b) Photogeneration occurs in the neutral region. The electron has to diffuse to the depletion layer and then roll down the energy hill, i.e., drift across the SCL. (c) A shorted pn junction. The photogenerated electron and hole in the SCL roll down their energy hills, i.e., drift across the SCL, and cause a current I_{ph} in the external circuit. (d) The pn junction in open circuit. The photogenerated electron and hole roll down their energy hills (drift) but there is a voltage V_{oc} across the diode that causes them to diffuse back so that the net current is zero. Note that E_c and E_v are only shown in (a), and are self-apparent in (b)–(d).

drift more slowly) and the time duration of the photocurrent is longer (drift takes longer across the SCL). The total collected charge is the same, if we neglect recombination losses.

If we leave the photodiode in an open circuit, we would observe an open circuit voltage V_{oc} . The photogeneration as before creates an electron and a hole that separate and drift in the SCL as shown in Figure 5.2 (d). Extra electrons are added to the n-side and extra holes are added to the p-side. As photogeneration continues, accumulated electrons on the n-side neutralize some of the positive donor charges in the SCL close to the neutral n-side. Similarly, the accumulated holes on the p-side neutralize some of the negative acceptor charges in the SCL next to the

neutral *p*-side. First, note that the accumulation of electrons on the *n*-side and holes on the *p*-side essentially upsets the equilibrium in such a way that the extra negative charge on the *n*-side and positive charge on the *p*-side give rise to a voltage V_{oc} across the diode. We can also understand this by noting that the reduction in positive and negative donor and acceptor charges in the SCL causes the field to decrease from E_o to, say, $E_o - E_{oc}$, where E_{oc} is the field that has resulted from the effects of photogeneration in an open circuit *pn* junction. The built-in voltage must therefore also decrease from V_o to $V_o - V_{oc}$. The voltage V_{oc} appears as a forward bias across the junction. Such forward bias would inject electrons from the *n*- to the *p*-side and holes from the *p*- to the *n*-side as in normal diode operation, and hence result in a diode current I_{diode} . The current I_{diode} is actually in the opposite direction to the photocurrent I_{ph} and its magnitude is such that the total current I is zero, as it must be in an open circuit. Thus, the result is an appearance of an open circuit voltage V_{oc} across the *pn* junction. It is apparent that the **photovoltaic operation** in Figure 5.2 (c) and (d) represents the short and open circuit configurations with no external battery, and also represents the principle of operation of a **solar cell**. Normally, a load (R) would be connected to the terminals of the photodiode (or the solar cell).

C. Current-Voltage Convention and Modes of Operation

We need to establish a sign convention to be able to analyze photodiode circuits. The voltage across a *pn* junction is stated in terms of the *p*-side with respect to *n*-side and is shown as V in Figure 5.3 (a). The current direction I is taken as the current flowing out from the *n*-side or entering the *p*-side as indicated in Figure 5.3 (a). If the applied voltage is a reverse bias of 5 V, and for convenience we write it as $V_r = 5$ V, as shown in Figure 5.3 (b), then $V = -V_r = -5$ V. If we illuminate the *pn* junction, then there will be a photocurrent I_{ph} flowing from the positive battery terminal to the *n*-side as shown in Figure 5.2 (a) and (b). I_{ph} is also shown in Figure 5.3 (b),

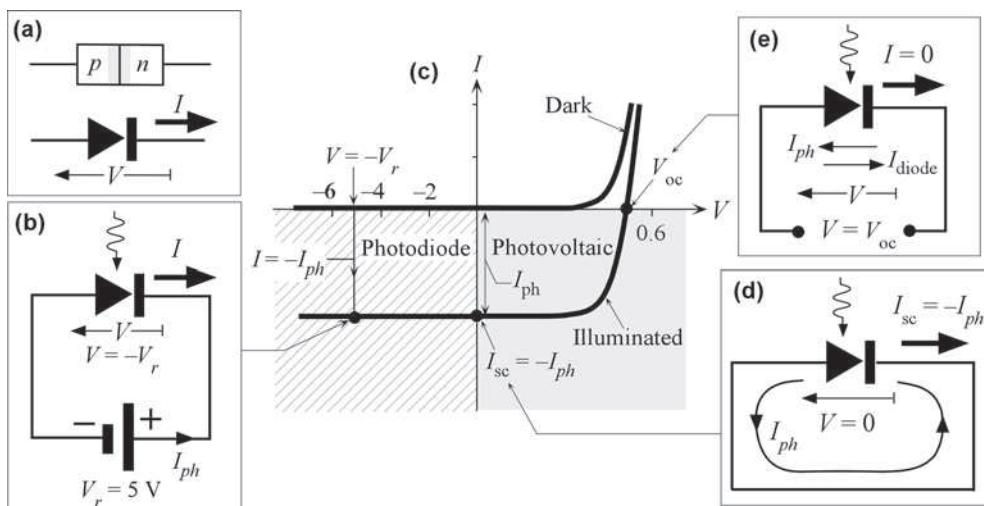


FIGURE 5.3 (a) The sign convention for the voltage V and current I for a *pn* junction. (b) If the *pn* junction is reverse biased by $V_r = 5$ V, then $V = -V_r = -5$ V. Under illumination, the *pn* junction current $I = -I_{ph}$ and is negative. (c) The I - V characteristics of a *pn* junction in the dark and under illumination. (d) A short circuit *pn* junction under illumination. The voltage $V = 0$ but there is a short circuit current so that $I = I_{sc} = -I_{ph}$. (e) An open circuit *pn* junction under illumination generates an open circuit voltage V_{oc} .

where I_{ph} is positive because it represents the actual current flow. However, the conventional diode current I is negative and written as $I = -I_{ph}$.

Figure 5.3 (c) shows the I - V characteristics of an ideal pn junction in the dark and also under illumination. In the dark, the I - V characteristic follows the usual diode equation, *i.e.*,

*pn
junction
equation*

$$I = I_o [\exp(eV/\eta k_B T) - 1] \quad (5.1.1)$$

where I and V are the diode current and voltage as in Figure 5.3 (a), I_o is the reverse saturation current, and η is the ideality factor.² For a reverse biased pn junction, $I = -I_o$ and is very small, which has been neglected in the dark characteristics in Figure 5.3 (c). We know from Figure 5.2 (c) that, in a short circuit, the photocurrent is I_{ph} , and it is in the opposite direction to the conventional pn junction current, that is, $I = I_{sc} = -I_{ph}$, where I_{sc} is the short circuit diode current. This short circuit situation is indicated in Figure 5.3 (d). We can sketch the I - V characteristics of the illuminated pn junction by simply shifting the dark I - V curve down until the current on the negative I -axis becomes $I = -I_{ph}$ as shown in Figure 5.3 (c). Notice that the pn junction I - V curve cuts the V -axis at $V = V_{oc}$. This is the **open circuit voltage** that would be generated when the pn junction is illuminated as shown in Figure 5.3 (e) in an open circuit, without any current in the external circuit. However, within the diode itself there are two currents that cancel each other exactly; one due to photogeneration (I_{ph}), and the other due to the diode current (I_{diode}) caused by V_{oc} , which acts like a forward bias across the junction.

The regions of the pn junction I - V characteristics that are bound by the positive V and negative I axes represent a **photovoltaic mode of operation**, shown as the gray region in Figure 5.3 (c). There is no applied bias and the light generates a photocurrent and a voltage across the device. When the pn junction is shorted, the current is I_{ph} , and when it is in open circuit, the voltage is V_{oc} . If there is a load of resistance R connected to the pn junction, we need to draw a load-line construction to find the operation (see Section 5.14). The region bound by the negative V -axis and the negative I -axis represents a reverse biased **photodiode mode of operation**; this is the most common mode of operation for the detection of light. However, it is not the only means of photodetection because there are several photovoltaic-type pn junction diodes used in radiation detection; two good examples are InAs and InSb infrared pn junction diodes. Although these are pn junctions, they are normally not biased due to the large dark current they have.

It should be emphasized that the dark current in the reverse biased pn junction was neglected in Figure 5.3 (c), and I_{ph} was assumed to be much greater than the magnitude of this reverse dark current. This assumption is not always true, especially under weak light conditions and in pn junctions made from narrow bandgap semiconductors that have high I_o .

5.2 SHOCKLEY-RAMO THEOREM AND EXTERNAL PHOTOCURRENT

Consider a semiconductor material with a negligible dark conductivity that is electroded and biased as shown in Figure 5.4 (a). The electrodes do not inject carriers but allow excess carriers in the sample to leave and become collected by the battery (they are termed *noninjecting electrodes*).³ The field E in the sample is uniform and it is V/L . We will later see that this situation

²The symbol η is also used later for the external quantum efficiency. One should be able to tell the difference from usage and content. Some books use n but this is used for the electron concentration.

³As will be apparent later on, the situation in Figure 5.4 (a) also represents a reverse biased *pin* photodiode, one of the most common detector elements in optoelectronics.

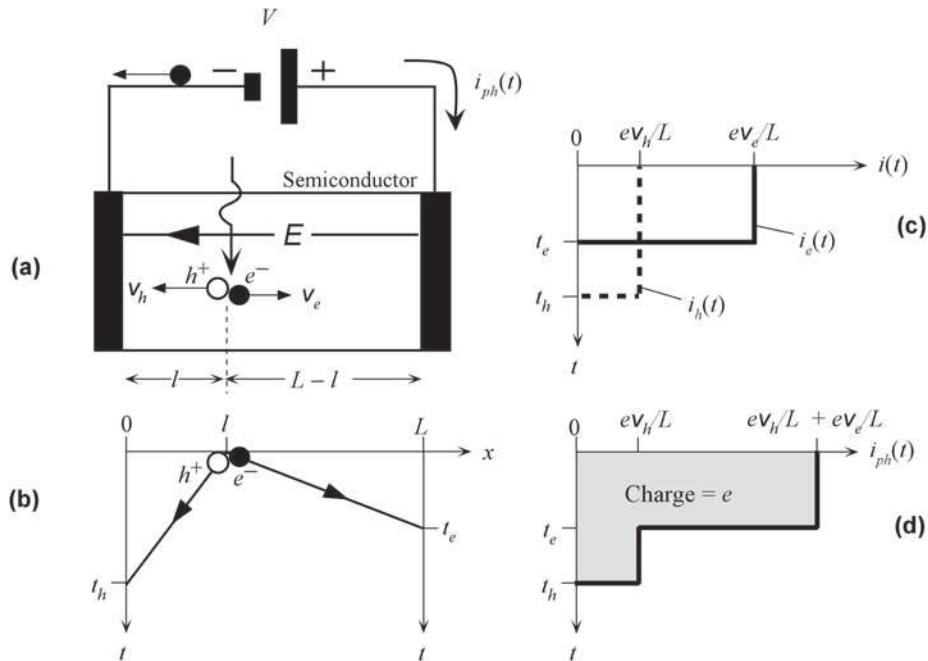


FIGURE 5.4 (a) An electron–hole pair is photogenerated at $x = l$. The electron and the hole drift in opposite directions with drift velocities v_h and v_e . (b) The electron arrives at time $t_e = (L - l)/v_e$ and the hole arrives at time $t_h = l/v_h$. (c) As the electron and hole drift, each generates an external photocurrent shown as $i_e(t)$ and $i_h(t)$. (d) The total photocurrent is the sum of hole and electron.

is almost identical to the intrinsic region of a reverse biased *pin* photodiode. Suppose that a single photon is absorbed at a position $x = l$ from the left electrode and instantly creates an electron–hole pair. The electron and the hole drift in opposite directions with respective drift velocities $v_e = \mu_e E$ and $v_h = \mu_h E$, where μ_e and μ_h are the electron and hole drift mobilities, respectively. The **transit time** of a carrier is the time it takes for a carrier to drift from its generation point to the collecting electrode. The electron and hole transit times t_e and t_h , respectively, are marked on the t vs. x diagram in Figure 5.4 (b), where

$$t_e = \frac{L - l}{v_e} \quad \text{and} \quad t_h = \frac{l}{v_h} \quad (5.2.1)$$

Electron
and hole
transit
times

Consider first only the drifting electron. Suppose that the external photocurrent due to the motion of this electron is $i_e(t)$. The electron is acted on by the force eE of the electric field. When it moves a distance dx , work must be done by the external circuit. In time dt , the electron drifts a distance dx and does an amount of work $eEdx$ which is provided by the battery in time dt as $Vi_e(t)dt$. Thus,

$$\text{Work done} = eEdx = Vi_e(t)dt$$

Using $E = V/L$ and $v_e = dx/dt$ we can find the electron photocurrent

$$i_e(t) = \frac{ev_e}{L}; \quad t < t_e \quad (5.2.2)$$

Electron
photocurrent

It is apparent that this current continues to flow as long as the electron is drifting (has a velocity v_e) in the sample. It lasts for a duration t_e at the end of which the electron reaches the battery. Thus, although the electron has been photogenerated instantaneously, the external photocurrent is *not* instantaneous and has a time spread. Figure 5.4 (c) shows the electron photocurrent $i_e(t)$.

We can apply similar arguments to the drifting hole as well, which will generate a hole photocurrent $i_h(t)$ in the external circuit given, as in Figure 5.4 (c), by

$$\text{Hole photocurrent} \quad i_h(t) = \frac{eV_h}{L}; \quad t < t_h \quad (5.2.3)$$

The total external current, called the transient or **instantaneous photocurrent**, $i_{ph}(t)$ will be the sum of $i_e(t)$ and $i_h(t)$, as shown in Figure 5.4 (d).

If we integrate the external photocurrent $i_{ph}(t)$ to evaluate the collected charge $Q_{\text{collected}}$, we would find

$$Q_{\text{collected}} = \int_0^{t_e} i_e(t) + \int_0^{t_h} i_h(t) = e \quad (5.2.4)$$

This result can be verified by evaluating the area under the $i_{ph}(t)$ curve in Figure 5.4 (d). Thus, the collected charge is not $2e$ but just one electron, as shown by the area in Figure 5.4 (d). Equations (5.2.2)–(5.2.4) constitute the **Shockley–Ramo theorem**.⁴ In general, if a charge q is being drifted with a velocity $v_d(t)$ by a field between two biased electrodes separated by L , then this motion of q generates an external current given by

$$\text{Shockley–Ramo theorem} \quad i(t) = \frac{qv_d(t)}{L}; \quad t < t_{\text{transit}} \quad (5.2.5)$$

The total external current is the sum of all currents of the type in Eq. (5.2.5) from all drifting charges between the electrodes, and the sign of q is negative for a drifting negative charge (electron).

5.3 ABSORPTION COEFFICIENT AND PHOTODETECTOR MATERIALS

The photon absorption process for photogeneration, that is, the creation of EHPs, requires the photon energy to be at least equal to the bandgap energy E_g of the semiconductor material to excite an electron from the valence band to the conduction band. The **upper cutoff wavelength** (or the threshold wavelength) λ_g for photogenerative absorption is therefore determined by the bandgap energy E_g of the semiconductor so that $h(c/\lambda_g) = E_g$ or

Cutoff wavelength and bandgap

$$\lambda_g(\mu\text{m}) = \frac{1.24}{E_g(\text{eV})} \quad (5.3.1)$$

For example, for Si, $E_g = 1.12 \text{ eV}$ and λ_g is $1.11 \mu\text{m}$, whereas for Ge $E_g = 0.66 \text{ eV}$ and the corresponding $\lambda_g = 1.87 \mu\text{m}$. It is clear that Si photodiodes cannot be used in optical communications at $1.3 \mu\text{m}$ and $1.55 \mu\text{m}$, whereas Ge photodiodes are commercially available for use

⁴Equation (5.2.5) is a simplified version of the more general treatment that examines the induced current on an electrode due to the motion of an electron. Its origins lie in tube-electronics in which engineers were interested in calculating how much current would flow into various electrodes of a vacuum tube as the electrons in the tube drifted. See W. Shockley, *J. Appl. Phys.*, 9, 635, 1938, and S. Ramo, *Proc. IRE* 27, 584, 1939.

TABLE 5.1 Bandgap energy E_g at 300 K, cutoff wavelength λ_g , and type of bandgap (D = Direct and I = Indirect) for some photodetector materials

Semiconductor	E_g (eV)	λ_g (μm)	Type
InP	1.35	0.91	D
GaAs _{0.88} Sb _{0.12}	1.15	1.08	D
Si	1.12	1.11	I
In _{0.7} Ga _{0.3} As _{0.64} P _{0.36}	0.89	1.4	D
In _{0.53} Ga _{0.47} As	0.75	1.65	D
Ge	0.66	1.87	I
InAs	0.35	3.5	D
InSb	0.18	7	D

at these wavelengths. Table 5.1 lists some typical bandgap energies and the corresponding cutoff wavelengths of various photodiode semiconductor materials.

Incident photons with wavelengths shorter than λ_g become absorbed as they travel in the semiconductor and the light intensity, which is proportional to the number of photons, decays exponentially with distance into the semiconductor. The light intensity I at a distance x from the semiconductor surface is given by

$$I(x) = I_o \exp(-\alpha x) \quad (5.3.2)$$

Light intensity and absorption coefficient

where I_o is the intensity of the incident radiation and α is the **absorption coefficient** that depends on the photon energy or wavelength λ . Absorption coefficient α is a material property. Most of the photon absorption (63%) occurs over a distance $1/\alpha$, and $1/\alpha$ is called the **penetration or absorption depth** δ . Figure 5.5 shows the α vs. λ characteristics of various semiconductors where it is apparent that the behavior of α with the wavelength λ depends on the semiconductor material.

In **direct bandgap** semiconductors such as III-V semiconductors (e.g., GaAs, InAs, InP, GaP) and in many of their alloys (e.g., InGaAs, GaAsSb) the photon absorption process is a direct

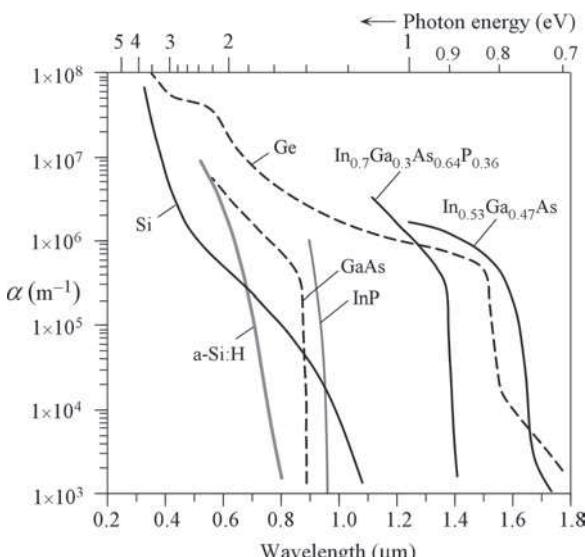


FIGURE 5.5 Absorption coefficient (α) vs. wavelength (λ) for various semiconductors. (Data selectively collected and combined from various sources.)

process which requires no assistance from lattice vibrations. The photon is absorbed and the electron is excited directly from the valence band to the conduction band without a change in its k -vector (or its crystal momentum $\hbar k$) inasmuch as the photon momentum is very small. The change in the electron momentum from the valence to the conduction band $\hbar k_{\text{CB}} - \hbar k_{\text{VB}} = \text{photon momentum} \approx 0$. This process corresponds to a vertical transition on the E - k diagram, that is, electron energy (E) vs. electron momentum ($\hbar k$) in the crystal in Figure 5.6 (a). A direct transition on the E - k diagram is a vertical transition from an initial energy E and wave vector k in the VB to a final energy E' and wave vector k' in the CB where $k' = k$, as illustrated in Figure 5.6 (a). The energy ($E' - E_c$) is the kinetic energy $(\hbar k)^2 / (2m_e^*)$ of the electron with an effective mass m_e^* , and $(E_v - E)$ is the kinetic energy $(\hbar k)^2 / (2m_h^*)$ of the hole left behind in the VB, where m_h^* is the hole effective mass. The ratio of the kinetic energies of the photogenerated electron and hole depends inversely on the ratio of their effective masses. (The absorption coefficient α is derived from the quantum mechanical transition probability from E to E' .) The absorption coefficient α of these direct bandgap semiconductors rises sharply with decreasing wavelength from λ_g as apparent for GaAs and InP in Figure 5.5.

In **indirect bandgap** semiconductors such as Si and Ge, the electron at the top of the VB has a crystal momentum $\hbar k_{\text{VB}}$ and at the bottom of the CB, its momentum is $\hbar k_{\text{CB}}$, which is very different than $\hbar k_{\text{VB}}$ as illustrated in Figure 5.6 (b). The electron therefore cannot simply absorb a photon of energy E_g to get excited to the bottom of the CB inasmuch as this will violate the conservation of momentum. The photon absorption for photon energies near E_g requires the absorption and emission of lattice vibrations, that is, **phonons**,⁵ during the photon absorption process, as shown in Figure 5.6 (b). If K is the wave vector of a lattice wave (lattice vibrations travel in the crystal), then $\hbar K$ represents the momentum associated with such a lattice vibration, that is, $\hbar K$ is a **phonon momentum**. When an electron in the valence band is excited to the conduction band there is a change in its momentum in the crystal and this change in the momentum cannot be supplied by the momentum of the incident photon, which is very small. Thus, the momentum difference must be balanced by a phonon momentum:

$$\hbar k_{\text{CB}} - \hbar k_{\text{VB}} = \text{phonon momentum} = \hbar K$$

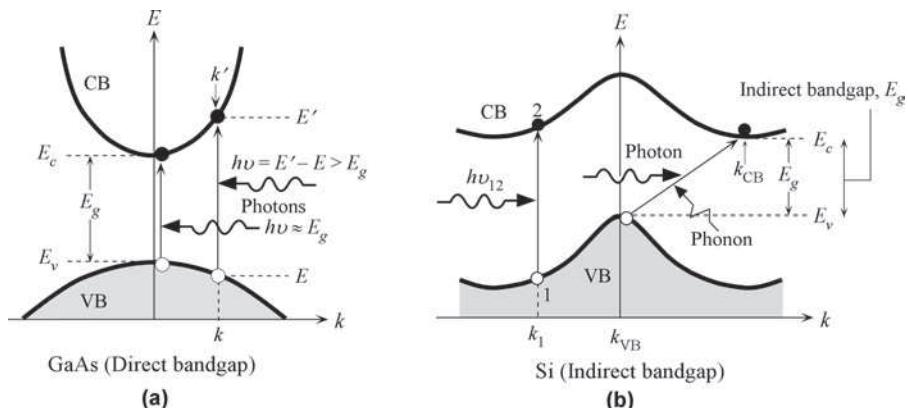


FIGURE 5.6 (a) Photon absorption in a direct bandgap semiconductor. (b) Photon absorption in an indirect bandgap semiconductor (VB, valence band; CB, conduction band).

⁵As much as an electromagnetic radiation is quantized in terms of photons, lattice vibrations in the crystal are quantized in terms of phonons.

The absorption process is said to be **indirect** as it depends on lattice vibrations which in turn depend on the temperature. Since the interaction of a photon with a valence electron needs a third body, a lattice vibration, the probability of photon absorption is not as high as in a direct transition. Furthermore, the cutoff wavelength is not as sharp as for direct bandgap semiconductors. During the absorption process, a phonon may be absorbed or emitted. The example in Figure 5.6 (b) shows the absorption of a phonon during the transition of the electron from E_g to E_c . If ϑ is the frequency of the lattice vibrations then the phonon energy is $h\vartheta$. The photon energy is $h\nu$ where ν is the photon frequency. Conservation of energy requires that

$$h\nu = E_g \pm h\vartheta$$

Thus, the onset of absorption does not exactly coincide with E_g , but typically it is very close to E_g inasmuch as $h\vartheta$ is small (<0.1 eV). The absorption coefficient initially rises slowly with decreasing wavelength from about λ_g , as apparent in Figure 5.5 for Ge and Si.

At sufficiently high photon energies, direct transitions eventually become possible, as shown by the transition 1 to 2 in Figure 5.6 (b). The 1-to-2 transition involves the absorption of a photon with energy $h\nu_{12}$. If the gradient of the E - k curve in the CB and VB are the same, then the transition probability is high.⁶ The 1-to-2 transition in Figure 5.6 (b) was chosen for this reason. If we examine α vs. λ for Si in Figure 5.5, the sharp increase for wavelengths smaller than ~ 0.5 μm is due to these direct transitions.

The choice of material for a photodiode must be such that the photon energies are greater than E_g . Further, at the wavelength of radiation, the absorption occurs over a depth covering the depletion layer so that the photogenerated EHPs can be separated and drifted by the field, and collected at the electrodes. If the absorption coefficient is too large, then absorption will occur very near the surface of the p^+ -layer which is outside the depletion layer. First, the absence of a field means that the photogenerated electron can only make it to the depletion layer to cross to the n -side by diffusion. Second, photogeneration near the surface invariably leads to rapid recombination due to surface defects that act as recombination centers. On the other hand, if the absorption coefficient is too small, only a small portion of the photons will be absorbed in the depletion region and only a limited number of EHPs can be photogenerated.

5.4 QUANTUM EFFICIENCY AND RESPONSIVITY

Not all the incident photons on a photodiode are absorbed to create *free* electron and hole pairs (EHPs) that can be collected, *i.e.*, give rise to a photocurrent. The efficiency of the conversion process of received photons to charge carriers that can be collected is measured by the **external or device quantum efficiency (QE)** η_e of the detector defined as⁷

$$\eta_e = \frac{\text{Number of collected electrons at detector terminals}}{\text{Number of incident photons}} \quad (5.4.1)$$

External quantum efficiency, QE

The measured photocurrent I_{ph} in the external circuit is due to the flow of electrons per second to the terminals of the photodiode, even though both electrons and holes would be drifting

⁶The reason is that the gradient dE/dk represents the group velocity of the electron, and photon-emitting or absorbing transitions need this group velocity to remain unchanged.

⁷“Free” implies carriers that can be collected, *i.e.*, electrons in the conduction band and holes in the valence band. Remember that the collection of one photogenerated free EHP is equivalent to the collection one electron externally.

inside the detector number of electrons collected per second is I_{ph}/e . If P_o is the incident optical power then the number of photons arriving per second is $P_o/h\nu$. The QE η_e can therefore also be defined by

*External
quantum
efficiency,
QE*

$$\eta_e = \frac{I_{ph}/e}{P_o/h\nu} \quad (5.4.2)$$

Not all of the absorbed photons may photogenerate free EHPs that can be collected. Some EHPs may disappear by recombination without contributing to the photocurrent or become immediately trapped. Further, if the semiconductor length is comparable with the penetration depth ($1/\alpha$), then not all the photons will be absorbed. There will also be photons lost by reflection at the air–semiconductor interface. The device QE is therefore always less than unity. It depends on the absorption coefficient α of the semiconductor at the wavelength of interest and on the structure of the device. QE can be increased by reducing the reflections at the semiconductor surface, increasing absorption within the depletion layer and preventing the recombination or trapping of carriers before they are collected. The QE defined in Eq. (5.4.1) is for the whole device and therefore it is known as the **external QE**. In contrast, the **internal quantum efficiency** is the number of free EHPs photogenerated per *absorbed photon* and is typically quite high for many devices. The QE definition in Eq. (5.4.1) incorporates the internal quantum efficiency as it applies to the whole device. The term QE alone normally implies external QE.

The **responsivity R** of a photodiode characterizes its performance in terms of the photocurrent generated (I_{ph}) per incident optical power (P_o) at a given wavelength

Responsivity

$$R = \frac{\text{Photocurrent (A)}}{\text{Incident optical power (W)}} = \frac{I_{ph}}{P_o} \quad (5.4.3)$$

From the definition of QE, it is clear that

*Responsivity
and QE*

$$R = \eta_e \frac{e}{h\nu} = \eta_e \frac{e\lambda}{hc} \quad (5.4.4)$$

In Eq. (5.4.4), η_e depends on the wavelength. The responsivity therefore clearly depends on the wavelength. R is also called the **spectral responsivity** or **radiant sensitivity**. The R vs. λ characteristic represents the spectral response of the photodiode and is generally provided by the manufacturer. Ideally, with a quantum efficiency of 100% ($\eta_e = 1$), R should increase with λ up to λ_g , as indicated in Figure 5.7. In practice, QE limits the responsivity to lie below the ideal photodiode line, with upper and lower wavelength limits as shown for a typical Si photodiode in Figure 5.7. The QE of a well-designed Si photodiode in the wavelength range 700–900 nm can be as high as 90–95%. Suppose that we draw a line through the origin that is a tangent to the R vs. λ curve at X as in Figure 5.7. We can show from Eq. (5.4.4) that point X represents operation under maximum QE; at the wavelength λ_1 , the QE is maximum and the responsivity is R_1 (see Example 5.4.2).

The shape of the responsivity R vs. λ curve in Figure 5.7 depends on a number of factors, the most important being the device structure, the absorption coefficient α of the semiconductor and the QE. Figure 5.8 shows a typical *pn* junction photodiode in which the *p*-side is heavily doped. The neutral *p*- and *n*-regions are marked by ℓ_p and ℓ_n , respectively. The reverse bias produces a large internal field E . The SCL has a width W , primarily on the *n*-side, and ℓ_p is much narrower than W . There are essentially three types of photogeneration processes that can contribute to the observed photocurrent.

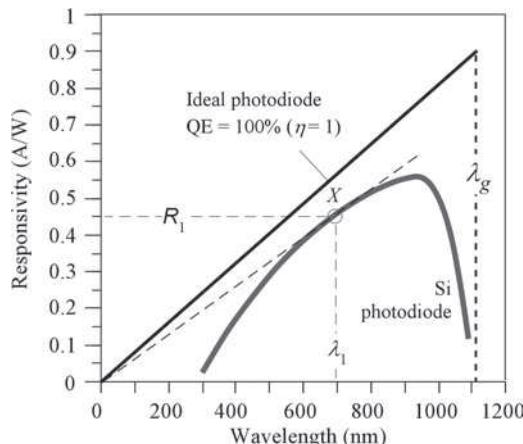


FIGURE 5.7 Responsivity (R) vs. wavelength (λ) for an ideal photodiode with QE = 100% ($\eta_e = 1$) and for a typical inexpensive commercial Si photodiode. (The exact shape of the responsivity curve depends on the device structure.) The line through the origin that is a tangent to the responsivity curve at X identifies operation at λ_1 with maximum QE.

Consider “short wavelengths,” that is, the absorption depth $1/\alpha$ is less than the width ℓ_p of the neutral region. The photogeneration takes place mainly within the neutral p -region, *i.e.*, within ℓ_p . If the photogeneration occurs within the **minority carrier diffusion length** L_e to the SCL boundary, as illustrated in Figure 5.8, then the electron can diffuse and reach the SCL, and then be drifted by the field E across the SCL over to the n -side, where it becomes collected. The photogenerated hole on the p -side is a majority carrier and is neutralized by the electron brought

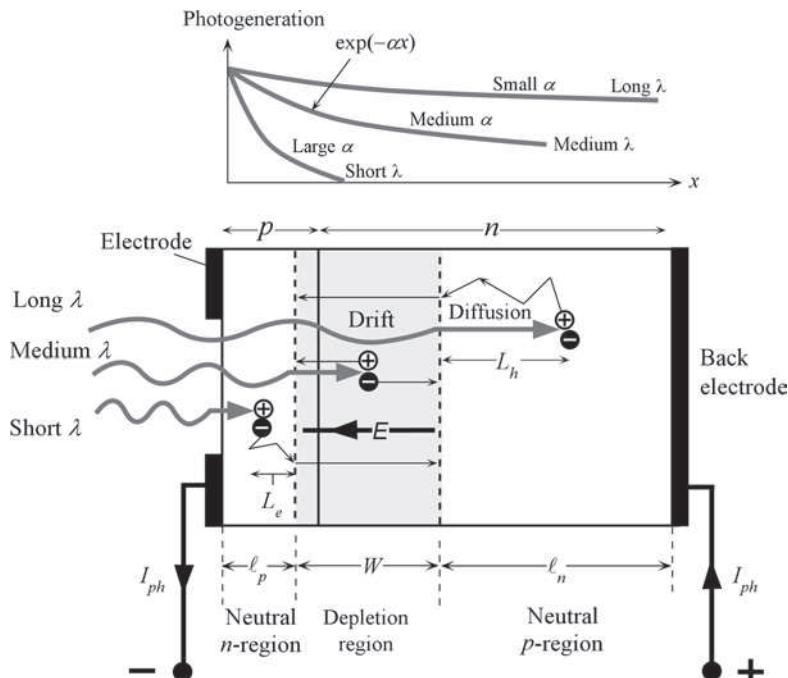
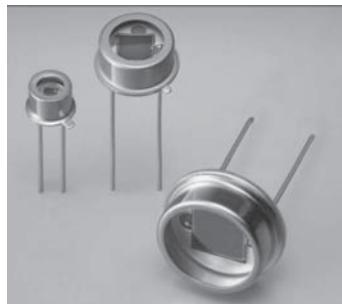


FIGURE 5.8 Different contributions to the photocurrent I_{ph} . Photogeneration profiles corresponding to short, medium, and long wavelengths are also shown.

in by the photocurrent from the negative terminal. The external photocurrent lasts until the drifting electron reaches the neutral n -region. (Although we considered the generation of one EHP, photogeneration is an ongoing process.)

Consider next “medium wavelengths,” *i.e.*, the absorption depth $1/\alpha$ is comparable to $\ell_p + W$. Significant photogeneration now takes place within the SCL, especially since ℓ_p is narrower than W . The field E in the SCL separates the EHP and drifts them in opposite directions, giving rise to a photocurrent. For “long wavelengths” for which the absorption depth is longer than $\ell_p + W$, as shown in Figure 5.8, only those holes photogenerated in the neutral n -side within the hole diffusion length L_h to the SCL can be collected. Thus, the useful photogeneration that contributes to the photocurrent takes place over $L_e + W + L_h$.

Although in this highly simplified discussion we only considered the interplay between the absorption coefficient (which depends on λ) and the device structure, the internal QE can also be important. The QE may not be uniform through the device. For very short wavelength light, the absorption depth is so small that the photogeneration occurs very close to the surface and outside the diffusion length L_e to the SCL. In Figure 5.8, these EHPs would be generated in $\ell_p - L_e$, which behaves like a “dead-zone”; the carriers essentially disappear by recombination. Short wavelength limitation of some of the pn junction photodiodes often arises from this dead zone. If ℓ_p is made shorter than L_e to overcome this problem, some carriers are still lost because they easily diffuse to the surface and disappear by recombination (surface is within the diffusion length L_e).



Si photodiodes of various sizes (S1336 series).
(Courtesy of Hamamatsu.)

EXAMPLE 5.4.1 Quantum efficiency and responsivity

Consider the photodiode shown in Figure 5.7. What is the QE at peak responsivity? What is the QE at 450 nm (blue)? If the photosensitive device area is 1 mm^2 , what would be the light intensity corresponding to a photocurrent of 10 nA at the peak responsivity?

Solution

The peak responsibility in Figure 5.7 occurs at about $\lambda \approx 940 \text{ nm}$ where $R \approx 0.56 \text{ A W}^{-1}$. Thus, from Eq. (5.4.4), that is, $R = \eta_e e \lambda / hc$, we have

$$0.56 \text{ A W}^{-1} = \eta_e \frac{(1.6 \times 10^{-19} \text{ C})(940 \times 10^{-9} \text{ m})}{(6.63 \times 10^{-34} \text{ J s})(3 \times 10^8 \text{ m s}^{-1})}, \quad \text{i.e.,} \quad \eta_e = 0.74 \quad \text{or} \quad 74\%$$

We can repeat the calculation for $\lambda = 450 \text{ nm}$, where $R \approx 0.24 \text{ A W}^{-1}$, which gives $\eta_e = 0.66$ or 66%.

From the definition of responsivity, $R = I_{ph}/P_o$, we have $0.56 \text{ A W}^{-1} = (10 \times 10^{-9} \text{ A})/P_o$, *i.e.*, $P_o = 1.8 \times 10^{-8} \text{ W}$ or 18 nW . Since the area is 1 mm^2 the intensity must be 18 nW mm^{-2} .

EXAMPLE 5.4.2 Maximum quantum efficiency

Show that a photodiode has maximum QE when

$$\frac{dR}{d\lambda} = \frac{R}{\lambda} \quad (5.4.5) \quad \begin{matrix} \text{Maximum} \\ \text{quantum} \\ \text{efficiency} \end{matrix}$$

that is, when the tangent X at λ_1 in Figure 5.7 passes through the origin ($R = 0, \lambda = 0$). Hence, determine the wavelength where the QE is maximum for the Si photodiode in Figure 5.7.

Solution

From Eq. (5.4.4) the QE is given by

$$\eta_e = \frac{hcR(\lambda)}{e\lambda} \quad (5.4.6)$$

where $R(\lambda)$ depends on λ and there is also λ in the denominator. We can differentiate Eq. (5.4.6) with respect to λ and then set it to zero to find the maximum point X . Thus

$$\frac{d\eta_e}{d\lambda} = \frac{hc}{e\lambda} \frac{dR}{d\lambda} - \frac{hcR}{e} \left(\frac{1}{\lambda^2} \right) = 0$$

which leads to Eq. (5.4.5). Equation (5.4.5) represents a line through the origin that is a tangent to the R vs. λ curve. This tangential point is X in Figure 5.7, where $\lambda_1 = 700$ nm and $R_1 = 0.45$ A W⁻¹. Then, using Eq. (5.4.6), the maximum QE is

$$\begin{aligned} \eta_e &= (6.626 \times 10^{-34} \text{ J s})(3 \times 10^8 \text{ m s}^{-1})(0.45 \text{ A W}^{-1}) / (1.6 \times 10^{-19} \text{ C})(700 \times 10^{-9} \text{ m}) \\ &= 0.80 \quad \text{or} \quad 80\% \end{aligned}$$

5.5 THE *pin* PHOTODIODE

The simple *pn* junction photodiode (Figure 5.1) has two major drawbacks. Its junction or depletion layer capacitance is not sufficiently small to allow photodetection at high modulation frequencies. This is an *RC* time constant limitation. Second, its depletion layer is at most a few microns. This means that at long wavelengths where the penetration depth is greater than the depletion layer width, the majority of photons are absorbed outside the depletion layer where there is no field to separate and drift the EHPs. The QE is correspondingly low at these long wavelengths. These problems are substantially reduced in the ***pin* (*p*-intrinsic-*n*-type) photodiode.⁸**

The *pin* refers to a semiconductor device that has the structure *p*⁺-intrinsic-*n*⁺ as schematically illustrated in the idealized structure in Figure 5.9 (a). The intrinsic layer has much smaller doping than both *p*⁺- and *n*⁺-regions and it is much wider than these regions, typically 5–50 μm depending on the particular application. In the idealized ***pin* photodiode**, we can take, for simplicity, the *i*-Si region to be truly intrinsic. When the structure is first formed, holes diffuse from the *p*⁺-side and electrons from *n*⁺-side into the *i*-Si layer where they recombine and disappear. This leaves behind a thin layer of exposed negatively charged acceptor ions in the *p*⁺-side and a thin layer of exposed positively charged donor ions in the *n*⁺-side as shown in Figure 5.9 (b).

⁸The *pin* photodiode was invented by Jun-ichi Nishizawa and his research group at Tohoku University in Japan in 1952. Among Professor Nishizawa's many distinguished awards is "The Order of Cultural Merits" from the Japanese Emperor. He is currently the President Emeritus of Tokyo Metropolitan University.

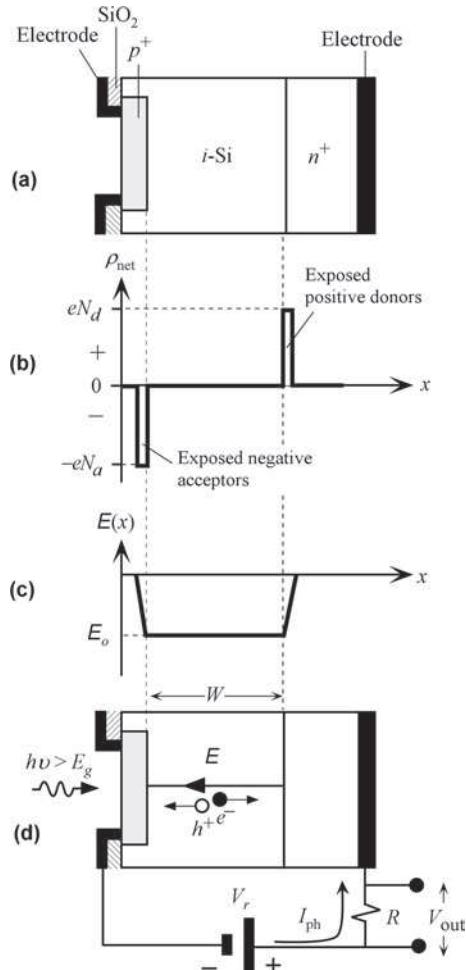


FIGURE 5.9 (a) The schematic structure of an idealized *pin* photodiode. (b) The net space charge density across the photodiode. (c) The built-in field across the diode. (d) The *pin* photodiode reverse biased for photodetection.

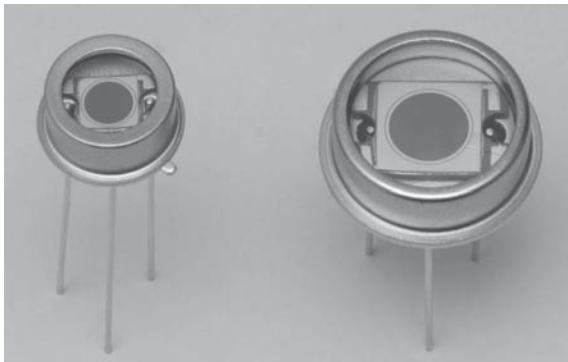
The two charges are separated by the *i*-Si layer of width W . There is a uniform built-in field E_0 in *i*-Si layer from the exposed positive ions to exposed negative ions as illustrated in Figure 5.9 (c). In contrast, the built-in field in the depletion layer of a *pn* junction is not uniform. With no applied bias, equilibrium is maintained by the built-in field E_0 which prevents further diffusion of majority carriers into the *i*-Si layer.

The separation of two very thin layers of negative and positive charges by a fixed distance, width W of the *i*-Si, is almost the same as that in a parallel plate capacitor. The **junction or depletion layer capacitance** C_{dep} of the *pin* diode is given by

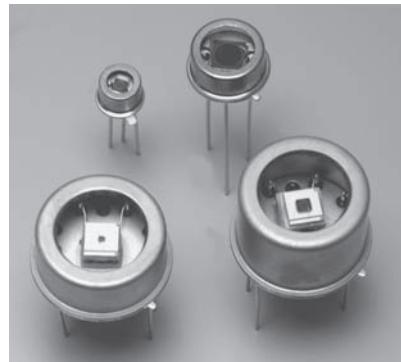
Junction
capacitance
of *pin*

$$C_{\text{dep}} = \frac{\epsilon_0 \epsilon_r A}{W} \quad (5.5.1)$$

where A is the cross-sectional area and $\epsilon_0 \epsilon_r$ is the permittivity of the semiconductor (Si), respectively, with ϵ_r being the relative permittivity. Further, since the width W of the *i*-Si layer is fixed by the structure, the junction capacitance C_{dep} does not depend on the applied voltage in contrast



Si *pin* photodiodes. Radiation receiving active device area has a diameter of 3 mm for the left and 5 mm for the right *pin* photodiode. The case diameters are about 8.1 mm and 12.4 mm. (*Courtesy of Hamamatsu.*)



InGaAs *pin* photodiodes. Active device areas have diameters from 0.3 mm (smallest) to 3 mm (largest). The bottom two are thermoelectric cooled, and have a case diameter of 15 mm. (*Courtesy of Hamamatsu.*)

to that of the *pn* junction. C_{dep} is typically of the order of a picofarad in fast *pin* photodiodes so that with a $50\text{-}\Omega$ resistor, the RC_{dep} time constant is about 50 ps.

When a reverse bias voltage V_r is applied across the *pin* device, as shown in Figure 5.9 (d), it drops almost entirely across the width of *i*-Si layer. The depletion layer widths of the thin sheets of acceptor and donor charges in the p^+ - and n^+ -sides are negligible compared with W . The reverse bias V_r increases the built-in voltage to $V_o + V_r$. The field E in the *i*-Si layer is still uniform and increases to

$$E = E_o + \frac{V_r}{W} \approx \frac{V_r}{W} \quad (V_r \gg V_o) \quad (5.5.2)$$

The *pin* structure is designed so that photon absorption occurs over the *i*-Si layer. The photogenerated EHPs in the *i*-Si layer are then separated by the field E and drifted toward the n^+ - and p^+ -sides, respectively, as illustrated in Figure 5.9 (d). While the photogenerated carriers are drifting through the *i*-Si layer they give rise to an external photocurrent which is detected as a voltage across a small sampling resistor R in Figure 5.9 (d). The **response time** of the *pin* photodiode is determined by the transit times of the photogenerated carriers across the width W of the *i*-Si layer. Increasing W allows more photons to be absorbed, which increases the QE but it slows down the speed of response as carrier transit times become longer. For a charge carrier that is photogenerated at the edge on the *i*-Si layer, the transit time or drift time t_{drift} across the *i*-Si layer is

$$t_{\text{drift}} = \frac{W}{v_d} \quad (5.5.3)$$

where v_d is its drift velocity. To reduce the drift time, that is, to increase the speed of response, we have to increase v_d and therefore increase the applied field E . At high fields v_d does not follow the expected $\mu_d E$ behavior, where μ_d is the drift mobility, but instead tends to saturate at v_{sat} , which is of the order of 10^5 m s^{-1} at fields greater than 10^6 V m^{-1} in the case of Si. Figure 5.10 shows the variation of the drift velocity of electrons and holes with the field in Si. The $v_d = \mu_d E$ behavior is only observed at low fields. At high fields, both electron and hole drift velocities saturate. For an *i*-Si layer of width $10 \mu\text{m}$, with carriers drifting at saturation velocities, the drift time

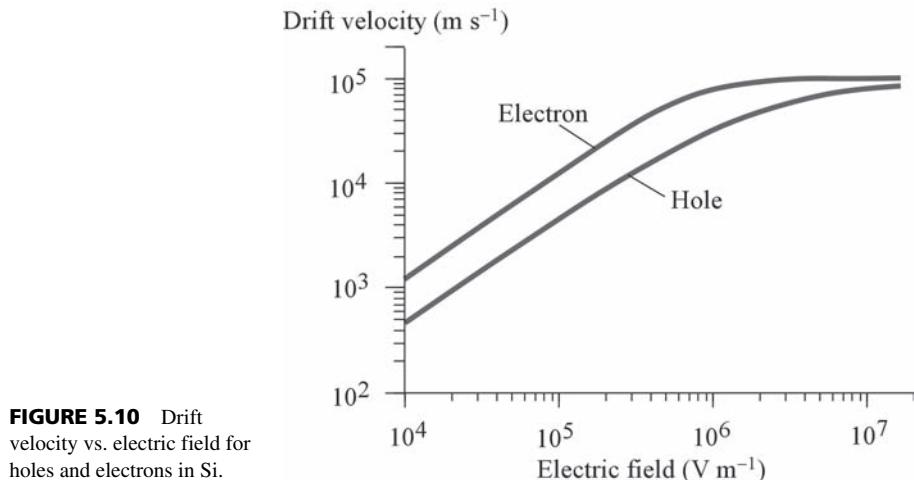


FIGURE 5.10 Drift velocity vs. electric field for holes and electrons in Si.

is about 0.1 ns which is longer than typical RC_{dep} time constants. The speed of *pin* photodiodes are invariably limited by the transit time of photogenerated carriers across the *i*-Si layer.

The *pin* photodiode structure shown in Figure 5.9 is, of course, idealized. In reality, the *i*-Si layer will have some small doping. For example, if the sandwiched layer is lightly *n*-type doped, it is labeled as a *v*-layer and the structure is p^+vn^+ . The sandwiched *v*-layer becomes a depletion layer with a small concentration of exposed positive donors. The field then is not entirely uniform across the photodiode. The field is maximum at the p^+v junction and then decreases slightly across *v*-Si to reach the n^+ -side. As an approximation we can still consider the *v*-Si layer as an *i*-Si layer.

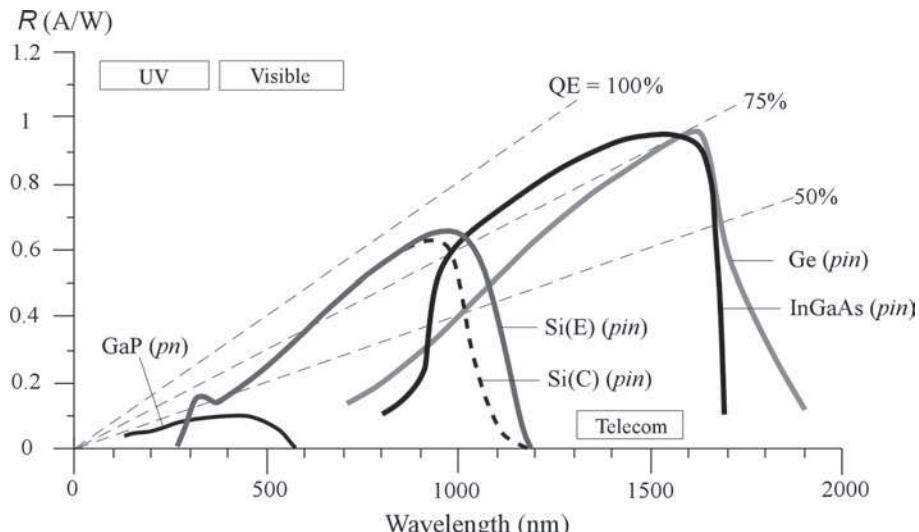


FIGURE 5.11 The responsivity of Si, InGaAs, and Ge *pin*-type photodiodes. The *pn* junction GaP detector is used for UV detection. GaP (Thorlabs, FGAP71), Si(E), IR enhanced Si (Hamamatsu S11499), Si(C), conventional Si with UV enhancement, InGaAs (Hamamatsu, G8376), and Ge (Thorlabs, FDG03). The dashed lines represent the responsivity due to $\text{QE} = 100\%$, 75% , and 50% .

As mentioned above, a distinct advantage of the *pin* photodiode is that it allows a wider spectral range to be absorbed in the SCL in which the photogeneration takes place. Consequently, the responsivity R is generally better than the simple *pn* junction photodiode and can be controlled by adjusting the width of the *i*-layer. Both Si and InGaAs *pin* photodiodes are widely available in the market, covering a range of wavelength from around 300 nm to 1700 nm. Ge *pin* photodiodes are also available but have higher dark currents, and usually have to be cooled. Figure 5.11 shows the responsivity of Ge, Si, and InGaAs *pin*, and GaP *pn* junction detectors from the UV (150 nm) to the optical communications channels. The dashed lines show the responsivity corresponding to certain quantum efficiencies, QE = 50%, 75%, and 100%. Notice that both Si and InGaAs *pin* photodiodes operate with quantum efficiencies, above 75%, and are the work horses for numerous photodetection tasks. The GaP *pn* junction is used for UV detection, though there are Si photodiodes available that can also operate in the UV. Further, notice also that the InGaAs photodiodes can be used in both 1550 nm and 1310 nm optical systems.

EXAMPLE 5.5.1 Operation and speed of a *pin* photodiode

A Si *pin* photodiode has an *i*-Si layer of width 20 μm . The p^+ -layer on the illumination side is very thin (0.1 μm). The *pin* is reverse biased by a voltage of 100 V and then illuminated with a very short optical pulse of wavelength 900 nm. What is the duration of the photocurrent if absorption occurs over the whole *i*-Si layer?

Solution

From Figure 5.5, the absorption coefficient at 900 nm is $\sim 3 \times 10^4 \text{ m}^{-1}$ so that the absorption depth is $\sim 33 \mu\text{m}$. We can assume that absorption, and hence photogeneration, occurs over the entire width W of the *i*-Si layer. The field in the *i*-Si layer is

$$E \approx V_r/W = (100 \text{ V})/(20 \times 10^{-6} \text{ m}) = 5 \times 10^6 \text{ V m}^{-1}$$

At this field the electron drift velocity v_e is very near its saturation at 10^5 m s^{-1} , whereas the hole drift velocity v_h is about $7 \times 10^4 \text{ m s}^{-1}$ as shown in Figure 5.10. Holes are slightly slower than the electrons. The transit time t_h of holes across the *i*-Si layer is

$$t_h = W/v_h = (20 \times 10^{-6} \text{ m})/(7 \times 10^4 \text{ m s}^{-1}) = 2.86 \times 10^{-10} \text{ s} \quad \text{or} \quad 0.29 \text{ ns}$$

This is the response time of the *pin* as determined by the transit time of the slowest carriers, holes, across the *i*-Si layer. To improve the response time, the width of the *i*-Si layer has to be narrowed but this decreases the quantity of photons absorbed and hence reduces the responsivity. There is therefore a trade-off between speed and responsivity.

EXAMPLE 5.5.2 Photocarrier diffusion in a *pin* photodiode

A reverse biased *pin* photodiode is illuminated with a short wavelength light pulse that is absorbed very near the surface as shown in Figure 5.12. The photogenerated electron has to diffuse to the depletion region, where it is swept into the *i*-layer and drifted across by the field in this region. What is the speed of response of this photodiode if the *i*-Si layer is 20 μm and the p^+ -layer is 1 μm and the applied voltage is 60 V? The diffusion coefficient (D_e) of electrons in the heavily doped p^+ -region is approximately $3 \times 10^{-4} \text{ m}^2 \text{ s}^{-1}$.

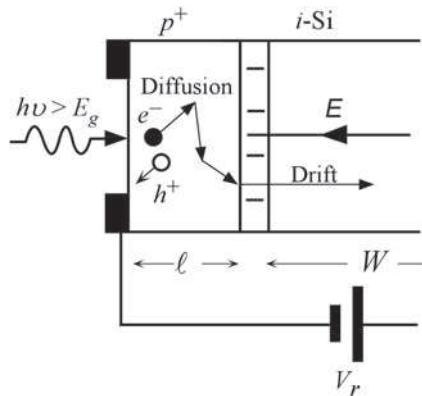


FIGURE 5.12 A reverse biased *p*-*i*-*n* photodiode is illuminated with a short wavelength light pulse that is absorbed very near the surface. The photogenerated electron has to diffuse to the depletion region where it is swept into the *i*-layer and drifted across.

Solution

There is no electric field in the p^+ -side outside the depletion region as shown in Figure 5.12. The photogenerated electrons have to make it across to the n^+ -side to give rise to a photocurrent. In the p^+ -side, the electrons move by diffusion. In time t , an electron, on average, diffuses a distance ℓ given by⁹

$$\ell = (2D_e t)^{1/2}$$

The *diffusion time* t_{diff} is the time it takes for an electron to diffuse across the p^+ -side (of length ℓ) to reach the depletion layer and is given by

$$t_{\text{diff}} = \ell^2 / (2D_e) = (1 \times 10^{-6} \text{ m})^2 / [2(3 \times 10^{-4} \text{ m}^2 \text{ s}^{-1})] = 1.67 \times 10^{-9} \text{ s} \quad \text{or} \quad 1.67 \text{ ns}$$

On the other hand, once the electron reaches the depletion region, it becomes drifted across the width W of the *i*-Si layer at the saturation drift velocity since the electric field here is $E = V_r/W = 60 \text{ V}/20 \mu\text{m} = 3 \times 10^6 \text{ V m}^{-1}$; and at this field the electron drift velocity v_e saturates at 10^5 m s^{-1} . The *drift time* across the *i*-Si layer is

$$t_{\text{drift}} = W/v_e = (20 \times 10^{-6} \text{ m}) / (1 \times 10^5 \text{ m s}^{-1}) = 2.0 \times 10^{-10} \text{ s} \quad \text{or} \quad 0.2 \text{ ns}$$

Thus, the response time of the *p*-*i*-*n* to a pulse of short wavelength radiation that is absorbed near the surface is very roughly $t_{\text{diff}} + t_{\text{drift}}$ or 1.87 ns. Notice that the diffusion of the electron is much slower than its drift. Further, in a proper analysis, we have to consider the diffusion and drift of many carriers, and we have to average $(t_{\text{diff}} + t_{\text{drift}})$ for all the electrons.

EXAMPLE 5.5.3 Responsivity of a *p*-*i*-*n* photodiode

A Si *p*-*i*-*n* photodiode has an active light-receiving area of diameter 0.4 mm. When radiation of wavelength 700 nm (red light) and intensity 0.1 mW cm^{-2} is incident, it generates a photocurrent of 56.6 nA. What is the responsivity and external QE of the photodiode at 700 nm?

Solution

The incident light intensity $I = 0.1 \text{ mW cm}^{-2}$ means that the incident power for conversion is

$$P_o = AI = [\pi(0.02 \text{ cm})^2](0.1 \times 10^{-3} \text{ W cm}^{-2}) = 1.26 \times 10^{-7} \text{ W} \quad \text{or} \quad 0.126 \mu\text{W}$$

⁹See, for example, S. O. Kasap, *Principles of Electronic Materials and Devices*, 3rd Edition (McGraw-Hill, 2006), Chs. 1 and 5 for the derivation of the root mean square diffusion distance.

The responsivity is

$$R = I_{ph}/P_o = (56.6 \times 10^{-9} \text{ A})/(1.26 \times 10^{-7} \text{ W}) = 0.45 \text{ A W}^{-1}$$

The QE can be found from

$$\eta_e = R \frac{hc}{e\lambda} = (0.45 \text{ A W}^{-1}) \frac{(6.62 \times 10^{-34} \text{ J s})(3 \times 10^8 \text{ m s}^{-1})}{(1.6 \times 10^{-19} \text{ C})(700 \times 10^{-9} \text{ m})} = 0.80 = 80\%$$

EXAMPLE 5.5.4 Steady state photocurrent in the *pin* photodiode

Consider a *pin* photodiode that is reverse biased and illuminated, as in Figure 5.9, and operating under steady state conditions. Assume that the photogeneration takes place inside the depletion layer of width W , and the neutral *p*-side is very narrow. If the incident optical power on the semiconductor is $P_o(0)$, then $TP_o(0)$ will be transmitted, where T is the transmission coefficient. At a distance x from the surface, the optical power $P_o(x) = TP_o(0)\exp(-\alpha x)$. In a small volume δx at x , the absorbed radiation power (by the definition of α) is $\alpha P_o(x)\delta x$, and the number of photons absorbed per second is $\alpha P_o(x)\delta x/h\nu$. Of these absorbed photons, only a fraction η_i will photogenerate EHPs, where η_i is the **internal quantum efficiency** IQE. Thus, $\eta_i \alpha P_o(x)\delta x/h\nu$ number of EHPs will be generated per second. We assume these will drift through the depletion region and thereby contribute to the photocurrent. The current contribution δI_{ph} from absorption and photogeneration at x within the SCL will thus be

$$\delta I_{ph} = \frac{e\eta_i \alpha P_o(x)\delta x}{h\nu} = \frac{e\eta_i \alpha T P_o(0)}{h\nu} \exp(-\alpha x)\delta x$$

We can integrate this from $x = 0$ (assuming ℓ_p is very thin) to the end of $x = W$, and assuming $W \gg L_h$ to find

$$I_{ph} \approx \frac{e\eta_i T P_o(0)}{h\nu} [1 - \exp(-\alpha W)] \quad (5.5.4)$$

Steady state photocurrent pin photodiode

where the approximate sign embeds the many assumptions we made in deriving Eq. (5.5.4). Consider a *pin* photodiode without an AR coating so that $T = 0.68$. Assume $\eta_i = 1$. The SCL width is $20 \mu\text{m}$. If the device is to be used at 900 nm , what would be the photocurrent if the incident radiation power is 100 nW ? What is the responsivity? Find the photocurrent and the responsivity if a perfect AR coating is used. What is the primary limiting factor? What is the responsivity if $W = 40 \mu\text{m}$?

Solution

From Figure 5.5, at $\lambda = 900 \text{ nm}$, $\alpha \approx 3 \times 10^4 \text{ m}^{-1}$. Further for $\lambda = 0.90 \mu\text{m}$, the photon energy $h\nu = 1.24/0.90 = 1.38 \text{ eV}$. Given $P_o(0) = 100 \text{ nW}$, we have

$$I_{ph} \approx \frac{(1.6 \times 10^{-19})(1)(0.68)(100 \times 10^{-9})}{(1.38 \times 1.6 \times 10^{-19})} [1 - \exp(-3 \times 10^4 \times 20 \times 10^{-6})] = 22 \text{ nA}$$

and the responsivity $R = 22 \text{ nA}/100 \text{ nW} = 0.22 \text{ A W}^{-1}$, which is on the low-side.

Consider next a perfect AR coating so that $T = 1$, and using Eq. (5.5.4) again, we find $I_{ph} = 32.7 \text{ nA}$ and $R = 0.33 \text{ A W}^{-1}$, a significant improvement.

The factor $[1 - \exp(-\alpha W)]$ is only 0.451, and can be significantly improved by making the SCL thicker. Setting $W = 40 \mu\text{m}$ gives $[1 - \exp(-\alpha W)] = 0.70$ and $R = 0.51$, which is close to values for commercial devices.

The maximum theoretical photocurrent would be obtained by setting $\exp(-\alpha W) \approx 0$, $T = 1$, $\eta_i = 1$, which gives $I_{ph} = 73 \text{ nA}$ and $R = 0.73 \text{ A W}^{-1}$.

5.6 AVALANCHE PHOTODIODE

A. Principles and Device Structures

Avalanche photodiodes (APDs) are widely used in optical communications due to their high speed and internal gain. They are also used in various applications where sensitivity is important. A simplified schematic diagram of a Si **reach-through APD** is shown in Figure 5.13 (a). The n^+ -side is thin and it is the side that is illuminated through a window. There are three p -type layers of different doping levels next to the n^+ -layer to suitably modify the field distribution across the diode. The first is a thin p -type layer and the second is a thick, lightly p -type doped (almost intrinsic) layer, called the π -layer, and the third is a heavily doped p^+ -layer. The diode is reverse biased to increase the fields in the depletion regions. The net space charge distribution across the diode due to exposed dopant ions is shown in Figure 5.13 (b). Under zero bias, the depletion layer in the p -region (between $n^+ p$) does not normally extend across this layer. But when a sufficient reverse bias is applied, the depletion region in the p -layer widens to *reach through* to the π -layer (and hence the name *reach-through*). The field extends from the exposed positively charged donors in the thin depletion layer in n^+ -side, all the way to the exposed negatively charged acceptors in the thin depletion layer in p^+ -side.

The electric field is given by the integration of the net space charge density ρ_{net} across the diode subject to an applied voltage V_r across the device. The variation in the field across the diode is shown in Figure 5.13 (c). The field lines start at positive ions and end at negative ions

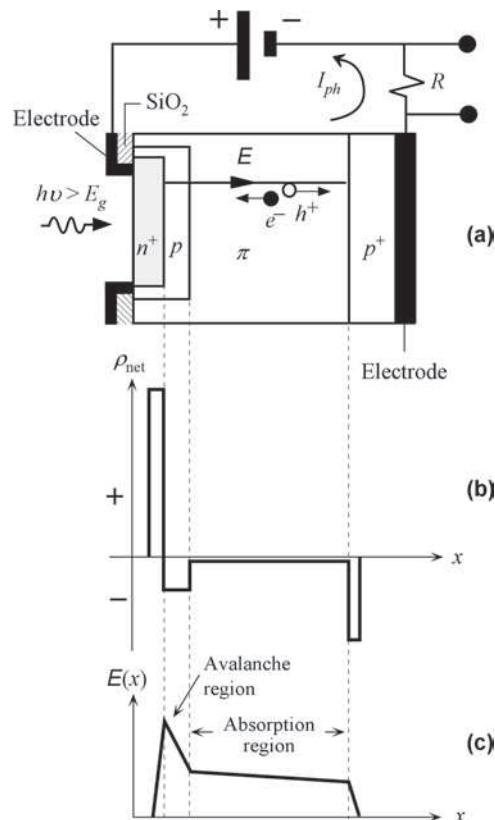


FIGURE 5.13 (a) A schematic illustration of the structure of an avalanche photodiode (APD) biased for avalanche gain. (b) The net space charge density across the photodiode. (c) The field across the diode and the identification of absorption and multiplication regions.

which exist through the p -, π -, and p^+ -layers. This means that E is maximum at the n^+p junction, then decreases slowly through the p -layer. Through the π -layer, it decreases only slightly as the net space charge density here is small. The field vanishes at the end of the narrow depletion layer in the p^+ -side.

The absorption of photons, and hence photogeneration, takes place mainly in the long π -layer. The nearly uniform field here separates the electron–hole pairs and drifts them at velocities near saturation toward the n^+ - and p^+ -sides, respectively. When the drifting electrons reach the p -layer, they experience even greater fields and therefore acquire sufficient kinetic energy (greater than E_g) to **impact-ionize** some of the Si covalent bonds and release EHPs. We can visualize the impact ionization process as shown in Figure 5.14 (a),¹⁰ where an electron entering the avalanche region (width w) gains energy from the field as it “drifts” in the opposite direction to the field, and its energy (which is kinetic energy) increases with respect to E_c . Eventually, the energy gained from the field is sufficient to excite an electron across the bandgap E_g as illustrated in Figure 5.14 (b). These impact-ionization-generated carriers are called **secondary carriers**. These secondary EHPs themselves can also be accelerated by the high fields in this region to sufficiently large kinetic energies to further cause impact ionization and release more EHPs, which leads to an **avalanche of impact ionization processes**. Thus, from a single electron entering the p -layer one can generate a large number of EHPs, all of which contribute to the observed photocurrent. The photodiode possesses an **internal gain mechanism** in that single photon absorption leads to a large number of EHPs being generated. The photocurrent in the APD in the presence of avalanche multiplication, therefore, corresponds to an effective quantum efficiency in excess of unity.

The reason for keeping the photogeneration within the π -region and reasonably separate from the avalanche p -region in Figure 5.13 (a) is that avalanche multiplication is a statistical process and hence leads to carrier generation fluctuations, which lead to **excess noise** in the avalanche-multiplied photocurrent. This is minimized if impact ionization is restricted to the carrier with the highest impact ionization efficiency, which in Si is the electron. Thus, the structure in Figure 5.13 (a) allows the photogenerated electrons to drift and reach the avalanche region but not the photogenerated holes.

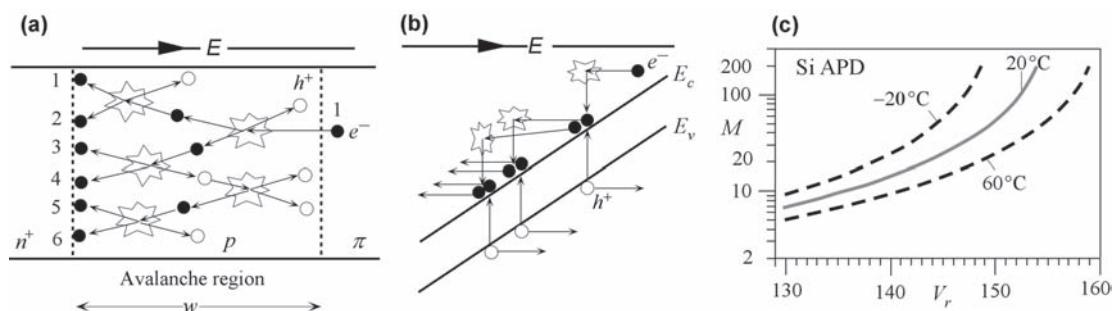


FIGURE 5.14 (a) A pictorial view of impact ionization processes releasing EHPs and the resulting avalanche multiplication. (b) Impact of an energetic conduction electron with crystal vibrations transfers the electron's kinetic energy to a valence electron and thereby excites it to the conduction band. (c) Typical multiplication (gain) M vs. reverse bias characteristics for a typical commercial Si APD, and the effect of temperature. (M measured for a photocurrent generated at 650 nm of illumination.)

¹⁰The impact of an energetic electron with a Si-Si bond to release an electron–hole pair is only an intuitive picture of the process.

The multiplication of carriers in the avalanche region depends on the probability of impact ionization, which depends strongly on the field in this region and hence on the reverse bias V_r . The overall or effective **avalanche multiplication factor M** , also known as the **gain**, of an APD is defined as¹¹

*Definition
of multi-
plication M*

$$M = \frac{\text{Multiplied photocurrent}}{\text{Primary unmultiplied photocurrent}} = \frac{I_{ph}}{I_{pho}} \quad (5.6.1)$$

where I_{ph} is the APD photocurrent that has been multiplied and I_{pho} is the **primary or unmultiplied photocurrent**, the photocurrent that is measured in the absence of multiplication, for example, under a small reverse bias V_r . The multiplication M is a strong function of the reverse bias, as shown in Figure 5.14 (c), and also the temperature. The multiplication M can be approximately described by an empirical relationship of the form¹²

*Multiplication
and Miller's
equation*

$$M = \frac{1}{1 - \left(\frac{V_r}{V_{br}}\right)^m} \quad (5.6.2)$$

where V_{br} is a parameter called the **avalanche breakdown voltage** and m is a characteristic index that provides the best fit to the experimental data. Both V_{br} and m are temperature dependent. For Si APDs M values can be as high as 1000 or more, but for many commercial Ge and InGaAs APDs M values are typically around 10–20. The multiplication M defined in Eq. (5.6.1) is also called the **gain** of the APD. Both V_r and V_{br} are positive for a reverse biased photodetector.

The speed of the reach-through APD, shown in Figure 5.13 (a), depends on three factors. First is the time it takes for the photogenerated electron to cross the absorption region (the π -layer) to the multiplication region (the p -layer). Second is the time it takes for the avalanche process to build up in the p -region and generate EHPs. The third is the time it takes for the last hole released in the avalanche process to transit through the π -region. The response time of an APD to an optical pulse is therefore somewhat longer than a corresponding *pin* structure but, in practice, the multiplicative gain often makes up for the reduction in the speed. The overall speed of a photodetector circuit also includes limitations from the electronic preamplifier connected to the photodetector. The APD requires less subsequent electronic amplification, which translates to an overall speed that can be faster than a corresponding detector circuit using a *pin* photodiode and a preamplifier.

One of the drawbacks of the simple reach-through APD structure is that the field around the n^+p junction peripheral edge reaches avalanche breakdown before the n^+p regions under the illuminated area as illustrated in Figure 5.15 (a). Ideally avalanche multiplication should occur uniformly in the illuminated region to encourage the avalanche multiplication of the primary photocurrent rather than the multiplication of the dark current (*i.e.*, the thermally generated random electron hole pairs). In a practical Si APD, an n -type doped region acting as a **guard ring** surrounds the central n^+ -region as shown in Figure 5.15 (b) so that the breakdown voltage around the periphery is now higher and avalanche is confined more to the illuminated region

¹¹This definition is for an *average* avalanche multiplication for two reasons. First is that the avalanche process is a statistical process with a mean and a standard deviation. Second, the field is not entirely uniform in the avalanche region and impact ionization probability is extremely sensitive to the electric field, which means that multiplication is not uniform over the multiplication region.

¹²This is usually called Miller's equation (S. L. Miller, *Phys. Rev.*, 99, 1234, 1955).

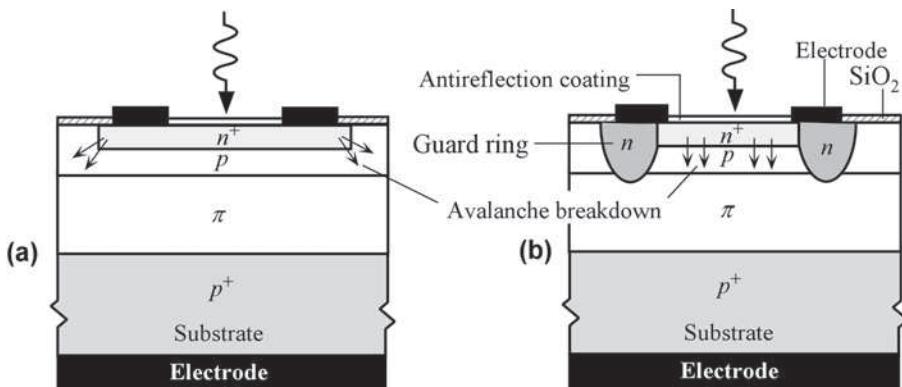


FIGURE 5.15 (a) A Si APD structure without a guard ring. (b) A schematic illustration of the structure of a more practical Si APD.

(n^+ p junction). The n^+ - and p -layers are very thin ($<2\ \mu\text{m}$) to reduce any absorption in this region; the main absorption occurs in the thick π -region.

Table 5.2 summarizes some typical characteristics of pn junction, pin , and APD photodiodes based on GaAsP, Si, Ge, InGaAs, InAs, and InSb, covering the range from the $\sim 200\ \text{nm}$ in the UV, based on GaP, to $\sim 5\ \mu\text{m}$ in the infrared (InSb). Si-based detectors, because of the well-established microelectronics technology, tend to be less expensive and more prevalent over the 200 – $1100\ \text{nm}$ range. For telecom applications, the detector must be able to respond to $1310\ \text{nm}$ and $1550\ \text{nm}$ optical signals and also have sufficient speed. These photodiodes are typically pin or avalanche photodiodes that have been carefully designed to provide a wide bandwidth detection.

TABLE 5.2 Typical characteristics of some pn junction-, pin -, and APD-type photodiodes based on GaP, GaAsP, Si, Ge, InGaAs, InAs, and InSb, covering the range from the $\sim 200\ \text{nm}$ in the UV (GaP), to $\sim 5\ \mu\text{m}$ in the infrared (InSb)

Photodiode	$\lambda_{\text{range}}\ (\text{nm})$	$\lambda_{\text{peak}}\ (\text{nm})$	R at $\lambda_{\text{peak}}\ (\text{A W}^{-1})$	Gain	I_d for $1\ \text{mm}^2$	Features
GaP pin	150–550	450	0.1	<1	1 nm	UV detection ^a
GaAsP pn	150–750	500–720	0.2–0.4	<1	0.005–0.1 nA	UV to visible, covering the human eye, low I_d
GaAs pin	570–870	850	0.4–0.5	<1	0.1–1 nA	High speed, low I_d
Si pn	200–1100	600–900	0.5–0.6	<1	0.005–0.1 nA	Inexpensive, general purpose, low I_d
Si pin	300–1100	800–1000	0.5–0.6	<1	0.1–1 nA	Faster than pn
Si APD	400–1100	800–900	0.4–0.6 ^b	$10\text{--}10^3$	1–10 nA ^c	High gains, fast
Ge pin	700–1800	1500–1580	0.4–0.7	<1	0.1–1 μA	IR detection, fast
Ge APD	700–1700	1500–1580	0.4–0.8 ^b	$10\text{--}20$	1–10 μA ^c	IR detection, fast
InGaAs pin	800–1700	1500–1600	0.7–1	<1	1–50 nA	Telecom, high speed, low I_d
InGaAs APD	800–1700	1500–1600	0.7–0.95 ^b	$10\text{--}20$	0.05–10 μA ^c	Telecom, high speed, and gain
InAs pn	2–3.6 μm	3.0–3.5 μm	1–1.5	<1	$>100\ \mu\text{A}$	Photovoltaic mode; normally cooled
InSb pn	4–5.5 μm	5 μm	3	<1	Large	Photovoltaic mode; normally cooled

Notes: I_d is the typical dark current under normal operating conditions. The dark current depends on the device area, and the values are typical for $1\ \text{mm}^2$. ^aFGAP71 (Thorlabs); ^bAt $M = 1$; ^cAt operating multiplication.

EXAMPLE 5.6.1 InGaAs APD responsivity

An InGaAs APD has a quantum efficiency (QE, η_e) of 60% at 1.55 μm in the absence of multiplication ($M = 1$). It is biased to operate with a multiplication of 12. Calculate the photocurrent if the incident optical power is 20 nW. What is the responsivity when the multiplication is 12?

Solution

The responsivity at $M = 1$ in terms of the quantum efficiency is

$$R = \eta_e \frac{e\lambda}{hc} = (0.6) \frac{(1.6 \times 10^{-19} \text{ C})(1550 \times 10^{-9} \text{ m})}{(6.626 \times 10^{-34} \text{ J s})(3 \times 10^8 \text{ m s}^{-1})} = 0.75 \text{ A W}^{-1}$$

If I_{pho} is the primary photocurrent (unmultiplied) and P_o is the incident optical power, then by definition $R = I_{pho}/P_o$ so that

$$I_{pho} = RP_o = (0.75 \text{ A W}^{-1})(20 \times 10^{-9} \text{ W}) = 1.5 \times 10^{-8} \text{ A} \quad \text{or} \quad 15 \text{ nA}$$

The photocurrent I_{ph} in the APD will be I_{pho} multiplied by M ,

$$I_{ph} = MI_{pho} = (12)(1.5 \times 10^{-8} \text{ A}) = 1.80 \times 10^{-7} \text{ A} \quad \text{or} \quad 180 \text{ nA}$$

The responsivity at $M = 12$ is

$$R' = I_{ph}/P_o = MR = (12)/(0.75) = 9.0 \text{ A W}^{-1}$$

EXAMPLE 5.6.2 Silicon APD

A Si APD has a QE of 70% at 830 nm in the absence of multiplication, that is, $M = 1$. The APD is biased to operate with a multiplication of 100. If the incident optical power is 10 nW, what is the photocurrent?

Solution

The unmultiplied responsivity is given by

$$R = \eta_e \frac{e\lambda}{hc} = (0.70) \frac{(1.6 \times 10^{-19} \text{ C})(830 \times 10^{-9} \text{ m})}{(6.626 \times 10^{-34} \text{ J s})(3 \times 10^8 \text{ m s}^{-1})} = 0.47 \text{ A W}^{-1}$$

The unmultiplied primary photocurrent from the definition of R is

$$I_{pho} = RP_o = (0.47 \text{ A W}^{-1})(10 \times 10^{-9} \text{ W}) = 4.7 \text{ nA}$$

The multiplied photocurrent is

$$I_{ph} = MI_{pho} = (100)(4.67 \text{ nA}) = 470 \text{ nA} \quad \text{or} \quad 0.47 \mu\text{A}$$

B. Impact Ionization and Avalanche Multiplication

The impact ionization process illustrated in Figure 5.14 (a) involves impact ionization in the avalanche region that has a width w . Intuitively, if we make the width longer, there should be more impact ionization processes, and hence M should be larger. Although the exact treatment can be quite complicated, we can at least derive some simple relationships between M and the width w and the

reverse bias V_r through semiquantitative argument. The probability that an electron in the avalanche region (w) causes impact ionization per unit distance is called the **electron ionization coefficient** α_e . We can view $1/\alpha_e$ as the average distance an electron travels in w before it causes impact ionization. Put differently, it is the average separation between two consecutive impact ionizations. Similarly, the probability that a hole causes impact ionization per unit distance would be the **hole ionization coefficient** α_h , and $1/\alpha_h$ is the average distance between two hole-initiated impact ionizations. The ratio k of these ionization coefficients is important in APD designs, and is defined as

$$k = \frac{\text{Ionization coefficient for holes}}{\text{Ionization coefficient for electrons}} = \frac{\alpha_h}{\alpha_e} \quad (5.6.3)$$

Ionization coefficient ratio

For Si, $\alpha_e > \alpha_h$ and k is less than unity. The ionization coefficients increase sharply with the field E in an exponential-like manner and can be written as¹³

$$\alpha_e = A \exp(-B/E) \quad (5.6.4)$$

Ionization coefficient and field

where A and B are material-specific constants, different for holes and electrons. α_e and E in Eq. (5.6.4) are at the same point inside the avalanche region. The increase in the field leads to large increases in α_e and α_h and hence M . Consider a small volume of narrow width dx , within w , in which there are N number of electrons. First, we will ignore the impact ionization by holes. The increase dN in N depends on the number electrons we have, N , and the probability of impact ionization in this volume, $\alpha_e dx$, so that $dN = N\alpha_e dx$. Put differently $(1/N)dN = \alpha_e dx$. We can easily integrate this from the left to the right end of the avalanche width w in Figure 5.14 (a), from N_1 to N_2 electrons, to obtain the increase in the number of electrons and hence the multiplication factor $M = N_2/N_1$. The integration is simple and gives an exponential dependence for M ,

$$M = \exp(\alpha_e w) \quad (5.6.5)$$

Multiplication by electrons only

Inasmuch as α_e depends on the field by virtue of Eq. (5.6.4), we end up with M having a very strong dependence on the field. In a proper treatment we would need to include the hole-initiated ionization events as well in accounting for multiplication in w . Equation (5.6.5) must be modified, and final result is¹⁴

$$M = \frac{1 - k}{\exp[-(1 - k)\alpha_e w] - k} \quad (5.6.6)$$

Multiplication by electrons and holes

Clearly, if $k = 0$ (holes do not ionize), Eq. (5.6.6) reduces to Eq. (5.6.5). In our APD model in Figure 5.14 (a), we allowed only the electrons to enter the avalanche region (not holes). We need to use Eq. (5.6.6) to properly account for the impact ionized holes (secondary holes) avalanching in w as well, as shown in Example 5.6.3. If electrons cannot impact ionize in this region, that is, when $k = \infty$, then $M = 1$, even if the hole ionization coefficient is very large. Equation (5.6.6) is easily modified to account for multiplication when holes are allowed to drift into the avalanche region. We would invert the definition of k and use α_h for α_e .

¹³Equation (5.6.4) is called the Chyweth's law. More generally α_e would be given by $\alpha_e = A \exp[-(B/E)^n]$ where n is a material-specific index. Since impact ionization coefficient measurements usually have some scatter (due to experimental uncertainties), many researchers simply use Eq. (5.6.4) with $n = 1$.

¹⁴See, for example, B. E. A. Saleh and M. C. Teich, *Fundamental of Photonics*, 2nd Edition (Wiley, 2007), Ch. 18, pp. 769–773.

Notice that M in Eq. (5.6.6) approaches infinity when the denominator $\exp[-(1 - k)\alpha_e w] - k \approx 0$. This occurs at sufficiently high fields when α_e and k satisfy this equation, and corresponds to a runaway multiplication in the avalanche region, *i.e.*, both electrons and holes, and their secondary carriers, are all heavily involved in avalanche multiplication. There is no such runaway, $M \approx \infty$, when $k = 0$, and we have only electrons ionizing, that is, only one type of carrier is involved. Further, in the presence of both electron and hole impact ionization, there is higher excess noise generated by the avalanche process as discussed below.

EXAMPLE 5.6.3 Avalanche multiplication in Si APDs

The electron and hole ionization coefficients α_e and α_h in silicon are approximately given by Eq. (5.6.4) with¹⁵ $A \approx 0.740 \times 10^6 \text{ cm}^{-1}$, $B \approx 1.16 \times 10^6 \text{ V cm}^{-1}$ for electrons (α_e), and $A \approx 0.725 \times 10^6 \text{ cm}^{-1}$ and $B \approx 2.2 \times 10^6 \text{ V cm}^{-1}$ for holes (α_h). Suppose that the width w of the avalanche region is $0.5 \mu\text{m}$. Find the multiplication gain M when the applied field in this region reaches $4.00 \times 10^5 \text{ V cm}^{-1}$, $4.30 \times 10^5 \text{ V cm}^{-1}$, and $4.38 \times 10^5 \text{ V cm}^{-1}$. What is your conclusion?

Solution

At the field of $E = 4.00 \times 10^5 \text{ V cm}^{-1}$, from Eq. (5.6.4)

$$\begin{aligned}\alpha_e &= A \exp(-B/E) \\ &= (0.740 \times 10^6 \text{ cm}^{-1}) \exp[-(1.16 \times 10^6 \text{ V cm}^{-1})/(4.00 \times 10^5 \text{ V cm}^{-1})] = 4.07 \times 10^4 \text{ cm}^{-1}\end{aligned}$$

Similarly using Eq. (5.6.4) for holes, $\alpha_h = 2.96 \times 10^3 \text{ cm}^{-1}$. Thus $k = \alpha_h/\alpha_e = 0.073$. Using this k and α_e above in Eq. (5.6.6) with $w = 0.5 \times 10^{-4} \text{ cm}$,

$$M = \frac{1 - 0.073}{\exp[-(1 - 0.073)(4.07 \times 10^4 \text{ cm}^{-1})(0.5 \times 10^{-4} \text{ cm})] - 0.073} = 11.8$$

Note that if we had only electron avalanche without holes ionizing, then the multiplication would be

$$M_e = \exp(\alpha_e w) = \exp[(4.07 \times 10^4 \text{ cm}^{-1})(0.5 \times 10^{-4} \text{ cm})] = 7.65$$

TABLE 5.3 Avalanche multiplication in a Si APD and breakdown

$E (\text{V cm}^{-1})$	$\alpha_e (\text{cm}^{-1})$	$\alpha_h (\text{cm}^{-1})$	k	M	M_e	Comment
4.00×10^5	4.07×10^4	2.96×10^3	0.073	11.8	7.65	M and M_e not too different at low E
4.30×10^5	4.98×10^4	4.35×10^3	0.087	57.2	12.1	7.5% increase in E , large difference between M and M_e
4.38×10^5	5.24×10^4	4.77×10^3	0.091	647	13.7	1.9% increase in E

We can now repeat the calculations for $E = 4.30 \times 10^5 \text{ V cm}^{-1}$ and again for $E = 4.38 \times 10^5 \text{ V cm}^{-1}$. The results are summarized in Table 5.3 for both M and M_e . Notice how quickly M builds up with the field and how a very small change at high fields causes an enormous change in M that eventually leads to a breakdown. (M running away to infinity as V_r increases.) Notice also that in the presence of only electron-initiated ionization, M_e simply increases without a sharp runaway to breakdown.

¹⁵ Values from Table 12.12 in S. Adachi, *Properties of Group IV, III-V and II-VI Semiconductors* (Wiley, UK, 2005).

5.7 HETEROJUNCTION PHOTODIODES

A. Separate Absorption and Multiplication APD

Various III-V compound-based APDs have been developed for use at the communications wavelengths $1.3\text{ }\mu\text{m}$ and $1.55\text{ }\mu\text{m}$. We will consider a few examples. As in the reach-through Si APD, the **absorption or photogeneration region** is separated from the **avalanche or multiplication region**, which allows the multiplication to be initiated by one type of carrier. Figure 5.16 is a simplified schematic diagram of the structure of an InGaAs-InP APD with a **separate absorption and multiplication (SAM) regions**. InP has a wider bandgap than InGaAs and the *p*- and *n*-type doping of InP is indicated by capital letters *P* and *N*. The main depletion layer is between P^+ -InP and *N*-InP layers and it is within the *N*-InP. This is where the field is greatest and therefore it is in this *N*-InP layer where avalanche multiplication takes place. With sufficient reverse bias, the depletion layer in the *N*-InP reaches (or punches) through to the *n*-InGaAs layer. Eventually, both *N*-InP and *n*-InGaAs are depleted. The field in the depletion layer in *n*-InGaAs is not as great as that in *N*-InP. The variation of the field across the device is also shown in Figure 5.16 under an applied reverse bias that has caused reach-through. Although the long-wavelength photons are incident onto the InP side, they are not absorbed by InP since the photon energy is less than the bandgap energy of InP ($E_g = 1.35\text{ eV}$). Photons pass through the InP layer and become absorbed in the *n*-InGaAs layer. The field in the *n*-InGaAs layer drifts the holes to the multiplication region where impact ionization multiplies the carriers.

There are a number of practical features which are not shown in the highly simplified diagram in Figure 5.16. Photogenerated holes drifting from *n*-InGaAs to *N*-InP become trapped at the interface because there is a sharp increase in the bandgap and a sharp change ΔE_v in E_v (valence band edge) between the two semiconductors, and holes cannot easily surmount the potential energy barrier ΔE_v as illustrated in Figure 5.17 (a). This problem is overcome by using thin layers of *n*-type InGaAsP with intermediate bandgaps to provide a graded transition from InGaAs to InP as illustrated in Figure 5.17 (b). Effectively ΔE_v has been broken up into two or more steps. The hole has sufficient energy to overcome the first step and enter the InGaAsP layer. It drifts and accelerates in the InGaAsP layer to gain sufficient energy to surmount the second step. These devices are called **separate absorption, grading and multiplication**.

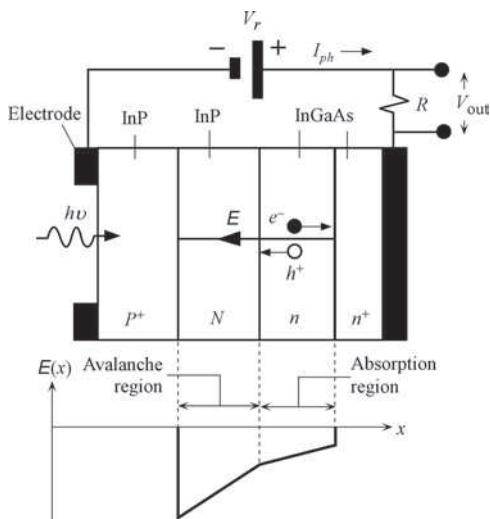


FIGURE 5.16 Simplified schematic diagram of a separate absorption and multiplication APD using a heterostructure based on InGaAs-InP. *P* and *N* refer to *p*- and *n*-type wider bandgap semiconductor.

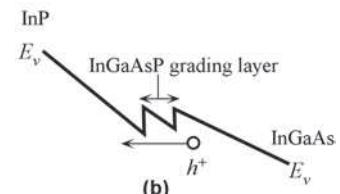
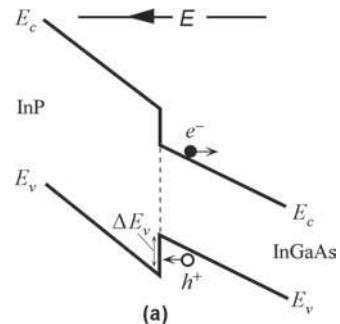


FIGURE 5.17 (a) Energy band diagrams for a SAM detector with a step junction between InP and InGaAs. There is a valence band step ΔE_v from InGaAs to InP that slows hole entry into the InP layer. (b) An interposing grading layer (InGaAsP) with an intermediate bandgap breaks ΔE_v and makes it easier for the hole to pass to the InP layer for a detector with a graded junction between InP and InGaAs. This is the SAGM structure.

(SAGM) APDs. Both the InP layers are grown epitaxially on an InP substrate. The substrate itself is not used directly to make the $P-N$ junction to prevent crystal defects (*e.g.*, dislocations) in the substrate appearing in the multiplication region and hence deteriorating the device performance. The schematic diagram of a more practical SAGM APD is illustrated in Figure 5.18.

The current vs. voltage characteristics for an InGaAs reach-through APD are shown in Figure 5.19. The dark current without any illumination is I_d and, like the photocurrent, it is also multiplied by the gain. The photodiode current has both the dark current I_d and the photocurrent I_{ph} . Initially the depletion region in the N -InP does not cover this entire region. The photogenerated hole in the n -InGaAs layer therefore has to diffuse some distance to the depletion region to find the field to be drifted. Thus, initially, as the reverse bias increases, the photocurrent I_{ph} and the dark current I_d increase because the reverse bias widens the depletion region further in the N -InP layer. The holes have to diffuse less and less distance as the voltage is increased, and the current increases until reach-through is attained, and the whole of the N -InP is depleted. From then onward, there is a field in the whole of the depleted N -layer and the photogenerated holes become drifted toward P^+ -InP. Impact ionization starts soon after reach-through, and the gain

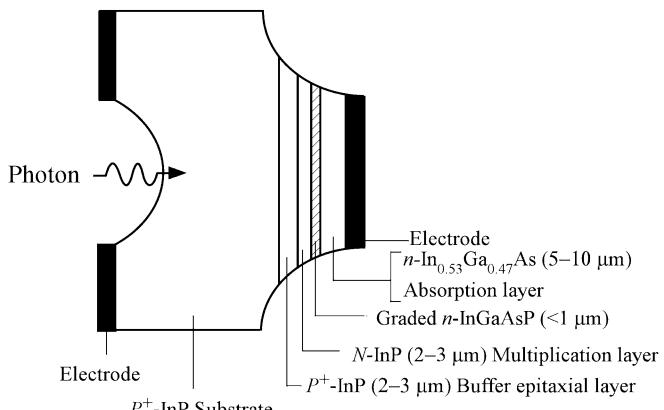


FIGURE 5.18 Simplified schematic diagram of a more practical mesa-etched SAGM layered APD.

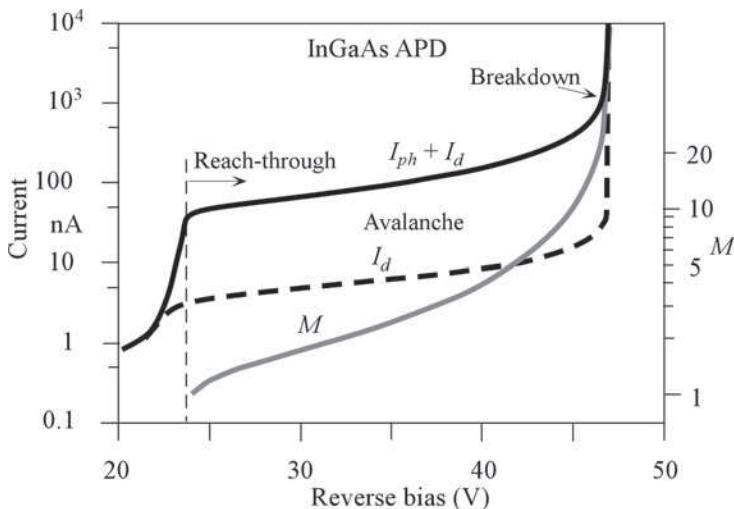


FIGURE 5.19 Typical current and gain (M) vs. reverse bias voltage for a commercial InGaAs reach-through APD. I_d and I_{ph} are the dark current and photocurrent, respectively. The input optical power is ~ 100 nW. The gain M is 1 when the diode has attained reach-through and then increases with the applied voltage. (Source: The data extracted selectively from Voxtel Catalog, Voxtel, Beaverton, OR 97006.)

M increases as the bias voltage increases as shown in Figure 5.19. Typical gains for commercial InGaAs APD are in the 10–20 range. At a sufficiently high reverse bias, the device will eventually break down with a runaway current; this voltage is the **breakdown voltage**.

B. Superlattice APDs

As mentioned earlier, APDs exhibit excess noise in the photocurrent (above the expected shot noise) due to inherent statistical variations in the avalanche multiplication process. This excess avalanche noise is reduced to minimum when only one type of carrier—for example, the electron in the case of Si APDs—is involved in impact ionizations. One method of achieving single carrier multiplication is by fabricating multilayer devices that have alternating layers of different bandgap semiconductors, as in **multiple quantum well** (MQW) devices discussed in Chapters 3 and 4.

The multilayered structure consisting of many alternating layers of different bandgap semiconductors is called a **superlattice**. Figure 5.20 (a) shows a highly simplified energy band diagram of a superlattice APD. There are thin alternating layers of narrow bandgap (E_{g1}) and wider bandgap (E_{g2}) semiconductor. Their widths are not the same. The wider bandgap semiconductors form the **barrier layers** surrounding the narrower bandgap semiconductors. The bandgap difference $E_{g2} - E_{g1}$ is taken up by the discontinuities ΔE_c and ΔE_v . For example, for GaAs/AlGaAs superlattice, ΔE_c to ΔE_v ratio is roughly 3 to 2. For InGaAs/InAlAs

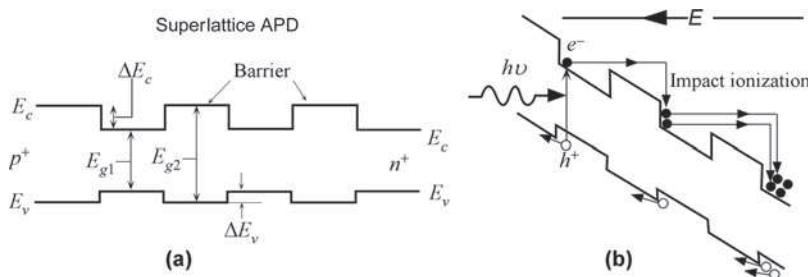


FIGURE 5.20
(a) Energy band diagram of an MQW superlattice APD. (b) Energy band diagram with an applied field and impact ionization.

superlattices, useful in the optical communications wavelengths 1.3–1.55 μm , ΔE_c to ΔE_v ratio is roughly 2 to 1.

When a large field E is applied, the whole band diagram is bent as shown in Figure 5.20 (b) to account for the potential energy of the electron (charge \times voltage). An electron in the conduction band (CB) in the E_{g1} -layer accelerates, passes into the E_{g2} -layer, continues to accelerate, and when it arrives and enters the E_{g1} -layer again it has high kinetic energy, and also an additional energy ΔE_c above E_c in this E_{g1} -layer. With its high kinetic energy and the additional ΔE_c , it can easily impact ionize the E_{g1} -semiconductor as shown in Figure 5.20 (b). Of course, the hole accelerates as well but ΔE_v is small. Thus, the electron impact ionization is enhanced much more than the hole impact ionization, a favorable outcome in terms of avalanche multiplication noise. In fact, with proper design, hole impact ionization in this structure can be prevented.

Another important superlattice structure is the **staircase superlattice APD** in which the bandgap is graded within each layer as shown in Figure 5.21 (a) without any bias. The bandgap in each layer changes from a minimum E_{g1} to a maximum E_{g2} which is more than twice E_{g1} . There is a step change ΔE_c in the conduction band edge between two neighboring graded layers that is greater than E_{g1} .

In very simple terms, as illustrated in Figure 5.21 (b), when a bias voltage is applied, that is a field E , the bands bend down. The photogenerated electron initially drifts in the graded layer conduction band. When the electron drifts into the neighboring layer, it now has a kinetic energy ΔE_c above E_c in this layer. It therefore enters the neighboring layer as a highly energetic electron and loses the excess energy ΔE_c by impact ionization. The process repeats itself from layer to layer, leading to an avalanche multiplication of the photogenerated electron. Since the impact ionization is primarily achieved as a result of transition over ΔE_c , the device does not need the high fields typical of avalanche multiplication in bulk semiconductors; it can operate at lower fields. Further, the impact-ionized holes experience only a small ΔE_v which is insufficient to lead to multiplication. Thus, effectively, only electrons are multiplied and the device behaves as if it were a **solid state photomultiplier**.

Such staircase superlattice APDs are difficult to fabricate and typically involve varying the composition of a quaternary III–V semiconducting alloy to obtain the necessary bandgap grading. Superlattice structures that are simply alternating layers of low and high bandgap semiconductor layers are easier to fabricate. Molecular beam epitaxy (MBE) is typically used to fabricate such multilayer structures.

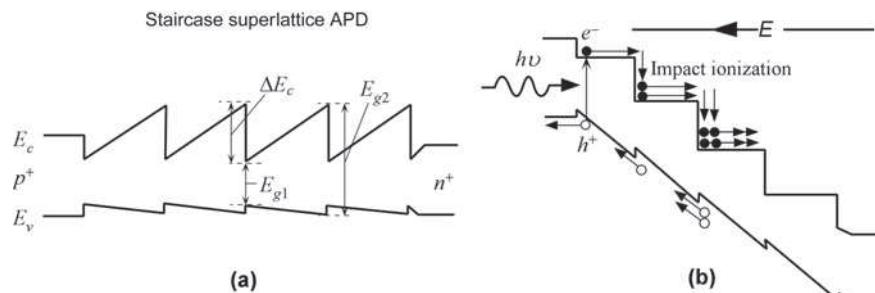


FIGURE 5.21 (a) Energy band diagram of a staircase superlattice APD without an applied bias, and (b) with an applied bias and impact ionization taking place.

5.8 SCHOTTKY JUNCTION PHOTODETECTOR

We consider the junction formed when a metal and an n -type semiconductor (n -SC) are brought into contact. In practice this process is frequently carried out by the evaporation of a metal onto the surface of a semiconductor crystal in vacuum. The energy band diagrams for the metal and the semiconductor are shown in Figure 5.22 (a). The work function, denoted as Φ , is the energy difference between the vacuum level and the Fermi level. Vacuum level defines the energy where the electron is free from that particular solid, and where the electron has zero KE.

We assume that the work function of the metal is greater than that of the semiconductor, $\Phi_m > \Phi_n$. When the two solids come into contact, the more energetic electrons in the CB of the n -SC can readily tunnel into the metal in search of lower empty energy levels (just above E_{Fm}) and accumulate near the surface of the metal as illustrated in Figure 5.22 (a). Electrons tunneling from the n -SC leave behind an electron-depleted region of width W in which there are exposed positively charged donors, in other words, net positive space charge. A **built-in potential** V_o therefore develops between the metal and the n -SC. There is obviously also a **built-in electric field** E_o from the positive charges to the negative charges on the metal surface as shown in Figure 5.22 (b). Eventually this built-in potential reaches a value that prevents further accumulation of electrons at the metal surface, and equilibrium is reached. The depletion region has been depleted of free carriers (electrons) and hence contains the exposed positive donors. The situation is actually similar to the pn junction. This region thus constitutes a **space charge layer** in which there is a nonuniform internal field directed from the semiconductor to the metal surface. The maximum value of this built-in field is denoted as E_o and occurs at the metal–semiconductor junction (this is where there are maximum number of field lines from positive to negative charges).

Once in contact, the Fermi level throughout the whole solid, the metal and n -SC in contact, must be uniform in equilibrium. Otherwise a change in the Fermi level, ΔE_F , going from one end to the other end will be available to do external (electrical) work. Thus, E_{Fm} and E_{Fn} line up as shown in Figure 5.22 (b). The SCL, however, has been depleted of electrons so in this region $E_c - E_{Fn}$

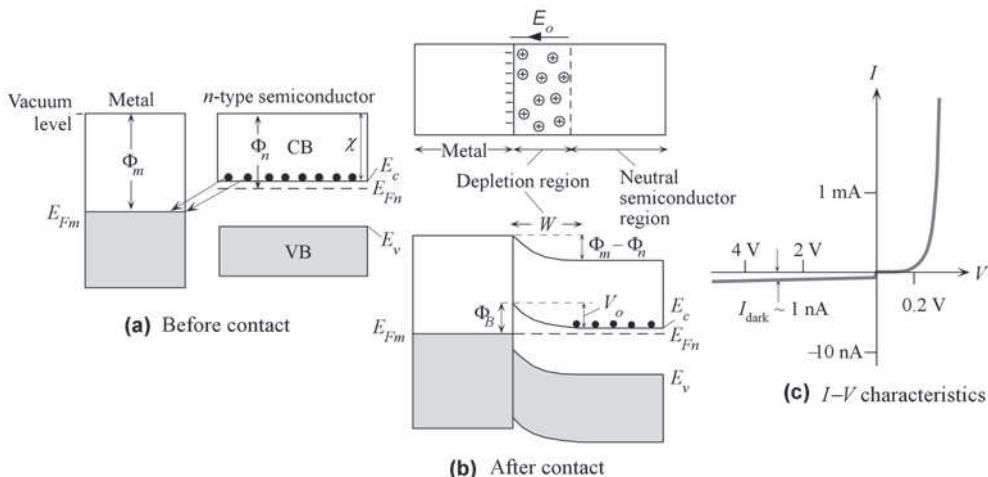


FIGURE 5.22 (a) Metal and an n -type semiconductor before contact. The metal work function Φ_m is greater than that of the n -type semiconductor. (b) A Schottky junction forms between the metal and the semiconductor. There is a depletion region in the semiconductor next to the metal and a built-in field E_o . (c) Typical I – V characteristics of a Schottky contact device.

must increase toward the junction so that n decreases. The bands must bend to increase $E_c - E_{Fn}$ toward the junction as illustrated in Figure 5.22 (b). Far away from the junction we, of course, still have an n -type semiconductor. The bending is just enough for the vacuum level to be continuous and changing by $\Phi_m - \Phi_n$ from the semiconductor to the metal as this much energy is needed to take an electron across from the semiconductor to the metal. The built-in potential energy barrier eV_o is simply $\Phi_m - \Phi_n$, given this is the amount of band bending, that is, the change in E_{Fn} to line it up with E_{Fm} . The PE barrier for electrons moving from the metal to the semiconductor is called the **Schottky barrier height**, Φ_B , which is greater than eV_o .

The $I-V$ characteristic of the Schottky junction is shown in Figure 5.22 (c), and is very similar to the pn junction.¹⁶ As apparent from Figure 5.22 (b), the barrier against electron injection from the n -SC to the metal is eV_o . The barrier against electron injection from the metal into the n -SC is Φ_B . In equilibrium, these two injection rates (they depend exponentially on the barrier heights) are small and just balance each other. Under forward bias, with the positive terminal connected to the metal and negative to the n -SC, the applied voltage V drops across the SCL, which reduces the built-in voltage to $V_o - V$. The barrier Φ_B remains unaltered. Since the injection rate depends on the Boltzmann factor $\exp[-e(V_o - V)/k_BT]$, there is an increase in this rate by a factor of $\exp(eV/k_BT)$, which leads to a very large rate of injection from the n -SC to the metal. The forward current is therefore large and depends exponentially on V .

Under reverse bias V_r , as shown in Figure 5.23 (a), V_o increases to $V_o + V_r$. The electron injection rate from n -SC to the metal vanishes and the current is dominated by the small rate of injection from the metal to n -SC over Φ_B and depends on $\exp(-\Phi_B/k_BT)$. The reverse current in the dark I_d depends on the nature of the metal to semiconductor contact (Φ_B) and the device area. It is smaller for wider bandgap semiconductors; some values are shown in Table 5.4 where I_d ranges from a few femtoamps to microamps per mm^2 of device area.

As a photodiode, the Schottky junction (SJ) is reverse biased so that the field in the depletion region is large. If we illuminate the depletion region of the SJ with photons of energy greater than E_g the photogeneration will take place within the SCL, where there is a strong field. The electrons and holes will roll down their respective energy hills as shown in Figure 5.23 (b), that is, they drift in this depletion region as in Figure 5.23 (c). For photon energies less than E_g the device can still respond, providing that the $h\nu$ can excite an electron from E_{Fm} in the metal over the PE barrier Φ_B into the CB, from where the electron will roll down toward the neutral

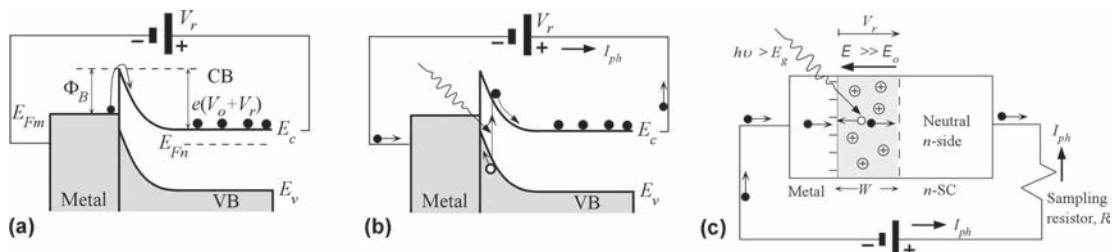


FIGURE 5.23 (a) Reverse biased Schottky junction and the dark current due to the injection of electrons from the metal into the semiconductor over the barrier Φ_B . (b) Photogeneration in the depletion region and the resulting photocurrent. (c) The Schottky junction photodetector.

¹⁶A straightforward derivation of the Schottky junction current–voltage characteristics can be found in S. O. Kasap, *Principles of Electronic Materials and Devices*, 2nd Edition (McGraw-Hill, 2006), Ch. 5.

TABLE 5.4 Schottky junction-based photodetectors and some of their features.
 τ_R and τ_F are the rise and fall times of the output of the photodetector for an optical pulse input

Schottky junction	λ range (nm)	R_{peak} (at peak) (A W^{-1})	J_{dark} (per mm^2)	Features with typical values
GaAsP	190–680	0.18 (610 nm)	5 pA	UV to red, $\tau_R = 3.5 \mu\text{s}$ (G1126 series ^a)
GaP	190–550	0.12 (440 nm)	5 pA	UV to green, $\tau_R = 5 \mu\text{s}$ (G1961 ^a)
AlGaN	220–375	0.13 (350 nm)	1 pA	Measurement of UV; blind to visible light (AG38S ^b)
GaAs	320–900	0.2 (830 nm)	$\sim 1 \text{nA}$	Wide bandwidth $> 10 \text{GHz}$, $\tau_R < 30 \text{ps}$ (UPD-30-VSG-P ^c)
InGaAs MSM	850–1650	0.4 (1300 nm)	5 μA	Optical high speed measurements, $\tau_R = 80 \text{ps}$, $\tau_F = 160 \text{ps}$ (G7096 ^a)
GaAs MSM	450–870	0.3 (850 nm)	0.1 nA	Optical high speed measurements, $\tau_R = 30 \text{ps}$, $\tau_F = 30 \text{ps}$ (G4176 ^a)

Note: The rise and fall times represent the times required for the output to rise from 10% to 90% of its final steady state value and to fall from 90% to 10% of its value before the optical pulse is turned off. ^aHamamatsu (Japan); ^bsglux (Germany); ^cAlphalas (Germany).

n-region. In this case hv must only be greater than Φ_B . Note also that the SJ photodiode can also be operated in the *photovoltaic mode*, *i.e.*, with no applied bias, since the built-in field E_o can easily separate and drift the photogenerated carriers in the SCL.

One distinct advantage of a SJ photodiode is that the SCL is right next to the metal contact. The short wavelength light that would be absorbed in the neutral region next to the electrode in a *pn* or *pin* photodiode is now absorbed in the SCL in the SJ photodiode. Consequently, the EHPs photogenerated can be separated immediately, drifted, and collected, whereas the minority carriers have to diffuse to the SCL in the *pn* and *pin* photodiodes. In fact, in *pn* and *pin* photodiodes, many of the EHPs diffuse to the surface and disappear by recombination. Clearly, SJ photodiodes are well suited for detecting short-wavelength light that is absorbed very close to the metal–semiconductor interface; and there are many SJ photodiodes for use in the UV region as summarized in Table 5.4. Further, the SJ diodes tend to have a wide spectral responsivity, *e.g.*, from the UV to the red.

Another advantage of SJ photodiodes is that they can be significantly faster than *pn* or *pin* photodiodes. Injected carriers in both forward and the reverse bias are electrons, which are majority carriers and therefore do not suffer from the limitations of minority carrier recombination time. In high-speed applications, such as in optical communications or optical measurements, the SJ photodiodes are used in pairs in a so-called interdigital electrode structure on the surface of the semiconductor as shown in Figure 5.24 (a) and (b). The light is absorbed in the semiconductor between the electrodes. We consider two of the digital electrodes, *A* and *B*, which would have been connected to positive and negative terminals of the battery as in Figure 5.24 (a). The Schottky junctions face each other, which means that whatever the polarity of the applied bias, one would always be reverse biased, *e.g.*, *A*, and the other, *B*, forward biased. Without any bias, the energy band diagram is symmetric as shown in Figure 5.24 (c). The SCLs for *A* and *B* are shown as SCL_1 and SCL_2 , respectively. The built-in voltages of *A* and *B* junctions are V_{o1} and V_{o2} , respectively. If an EHP is photogenerated in the semiconductor, they cannot drift and become collected as in the single SJ device. They are stuck between the two potential barriers (eV_{o1} and eV_{o2}) as indicated in Figure 5.24 (c).

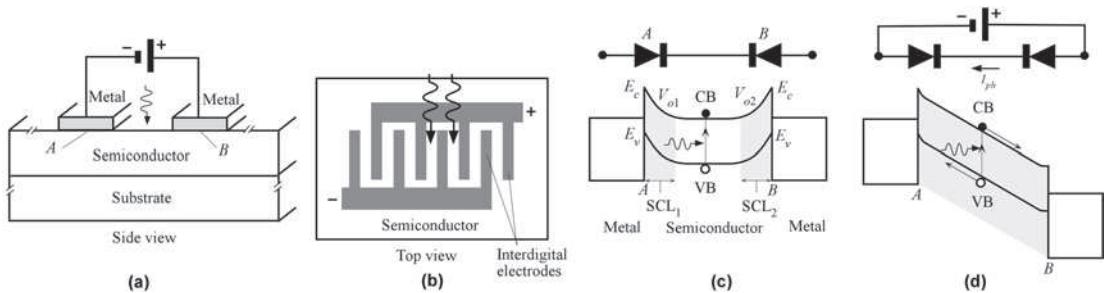


FIGURE 5.24 (a) The metal electrodes are on the surface of the semiconductor crystal (which is grown on a suitable substrate). (b) The electrodes are configured to be interdigital and on the surface of the crystal. (c) Two neighboring Schottky junctions are connected end-to-end, but in opposite directions as shown for A and B. The energy band diagram without any bias is symmetrical. The gray areas represent the SCL₁ and SCL₂ at A and B. (d) Under a sufficiently large bias, the SCL₁ from A extends and meets that from B so that the whole semiconductor between the electrodes is depleted. There is a large field in this region, and the photogenerated EHPs become separated and then drifted, which results in a photocurrent.

In contrast, in a single SJ, the EHP photogenerated inside the SCL can be collected as they become separated and drifted.

When a voltage is applied as in Figure 5.24 (d), V_{o1} is increased and V_{o2} is reduced since A is reverse and B is forward biased. The dark current is very small since A is reverse biased. However, we must bend the right-hand side (B) of the band down by eV_r to account for the applied voltage as shown in Figure 5.24 (d), because when the electron from A reaches B its PE must have dropped by eV_r . The applied bias extends the width of SCL₁ in the semiconductor and shrinks SCL₂, which means that further voltage increases will drop more across SCL₁ than SCL₂. With a sufficiently large voltage, the reverse bias at A will extend SCL₁ so much that it will punch through, or *reach-through*, onto the SCL₂, and the whole semiconductor between the electrodes becomes depleted. The corresponding bias voltage is called the **reach-trough voltage** (also known as the **flat-band voltage**). The advantage is that there is now a strong field in this depletion region and photogeneration would create an electron and a hole that can roll down their respective energy hills, *i.e.*, drift toward the electrodes, and give rise to a photocurrent. The photodiode we have considered is based on the basic **metal–semiconductor–metal** (MSM) structure, as in Figure 5.24 (a), that has two Schottky junctions facing each other as in Figure 5.24 (c). These devices are marketed as **MSM photodiodes**; two examples based on GaAs and InGaAs are listed in Table 5.4 and are capable of high-speed operation.



GaAsP Schottky junction photodiode for 190–680 nm detection, from UV to red. (*Courtesy of Hamamatsu.*)

GaP Schottky junction photodiode for 190–550 nm detection. (*Courtesy of Hamamatsu.*)

Schottky junction–type metal–semiconductor–metal (MSM)–type photodetectors. (*Courtesy of Hamamatsu.*)

AlGaN Schottky junction photodiode for UV detection. (*Courtesy of sglux, Germany.*)

5.9 PHOTOTRANSISTORS

The **phototransistor** is a bipolar junction transistor (BJT) that operates as a photodetector with a photocurrent gain. The basic principle is illustrated in Figure 5.25. In an ideal device, only the depletion regions, or the space charge layers, contain an electric field. The base terminal is normally open and there is a voltage applied between the collector and emitter terminals just as in the normal operation of a common emitter BJT. An incident photon is absorbed in the SCL between the base and collector to generate an electron-hole pair. The field E in the SCL separates the electron hole pair and drifts them in opposite directions. This is the **primary photocurrent** and it effectively constitutes a base current even though the base terminal is open circuit (current is flowing into the base from the collector rather than from the external base terminal). When the drifting electron reaches the collector, it becomes collected (and thereby neutralized) by the battery. On the other hand, when the hole enters the neutral base region, it can only be neutralized by injecting a large number of electrons into the base, as explained below. It effectively forces a large number of electrons to be injected from the emitter. Typically the electron recombination time in the base is very long compared with the time it takes for the electron to diffuse across the base. This means that only a small fraction of electrons injected from the emitter can recombine with photogenerated holes that have entered the base. Thus, the emitter has to inject a large number of electrons to neutralize any extra hole in the base. These electrons (except one) diffuse across the base and reach the collector and thereby constitute an amplified photocurrent.

Alternatively, in a more intuitive way, one can argue that the photogeneration of EHPs in the collector SCL decreases the resistance of this region, which decreases the voltage V_{CB} across the base-collector junction. Consequently the base-emitter voltage V_{BE} must increase inasmuch as $V_{BE} + V_{CB} = V_{CC}$ (Figure 5.25). This increase in V_{BE} acts as if it were a forward bias across the base-emitter junction and injects electrons into the base due to the transistor action, that is, $I_E \propto \exp(eV_{BE}/k_B T)$.

Since the photon-generated primary photocurrent I_{pho} is amplified as if it were a base current (I_B), the photocurrent flowing in the external circuit is

$$I_{ph} \approx \beta I_{pho} \quad (5.9.1)$$

Photo-current in
a phototransistor

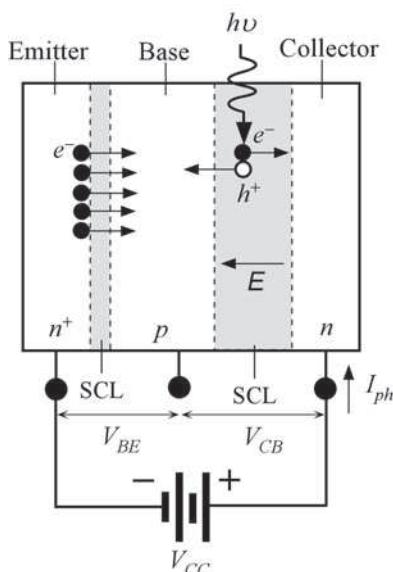


FIGURE 5.25 The principle of operation of the phototransistor. SCL is the space charge layer or the depletion region. The primary photocurrent acts as a base current and gives rise to a large photocurrent in the emitter-collector circuit.

where β is the current gain (or h_{FE}) of the transistor. The phototransistor construction normally allows the incident radiation to be absorbed in the base–collector junction SCL. Most commercial phototransistors only have two terminals, the emitter and collector, and generally do not have an external base terminal.

It is possible to construct a **heterojunction phototransistor** that has different bandgap materials for the emitter, base, and collector. For example, if the emitter in Figure 5.25 is InP ($E_g = 1.35$ eV) and the base is an InGaAsP alloy (e.g., $E_g \approx 0.85$ eV) then photons with energies less than 1.35 eV but more than 0.85 eV will pass through the emitter and become absorbed in the base. This means the device can be illuminated through the emitter.

5.10 PHOTOCONDUCTIVE DETECTORS AND PHOTOCONDUCTIVE GAIN

Photoconductive detectors have the simple structure schematically shown in Figure 5.26. Two electrodes are attached to a semiconductor that has the desired absorption coefficient and quantum efficiency over the wavelengths of interest. A bias voltage V is applied to the electrodes. Incident photons become absorbed in the semiconductor and photogenerate electron–hole pairs (EHPs). The result is an increase in the conductivity of the semiconductor and hence an increase in the external current, which constitutes the photocurrent I_{ph} as illustrated in Figure 5.26.

The actual response of the detector depends on whether the contacts to the semiconductor are ohmic or blocking (e.g., a reverse biased Schottky junction that does not inject carriers), and on the nature of carrier recombination kinetics. We will consider a photoconductor with ohmic contacts, that is, the contacts do not limit the current flow. With ohmic contacts, the photoconductor exhibits **photoconductive gain**, that is, *the external photocurrent is due to more than one electron flow per absorbed photon*, as illustrated in Figure 5.27 and explained below.

An absorbed photon photogenerates an EHP, which drift in opposite directions as shown in Figure 5.27 (a). The electron drifts much faster than the hole and therefore leaves the sample quickly. The sample, however, must be neutral which means another electron must enter the sample from the negative electrode as in (b) (the electrode is ohmic). This new electron also drifts across quickly as in (b) and (c) to leave the sample while the hole is still drifting slowly in the sample. Thus, another electron must enter the sample to maintain neutrality as in (d) and (e),

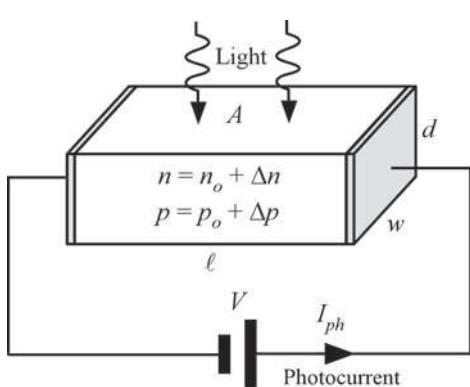
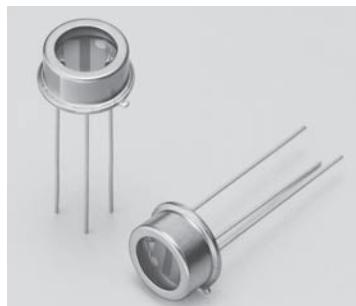


FIGURE 5.26 A semiconductor slab of length ℓ , width w , and depth d is illuminated with light of wavelength λ .



PbS (lead sulfide) photoconductive detectors for the detection of IR radiation up to $2.9\text{ }\mu\text{m}$. They are typically used in such applications as radiation thermometers, flame monitors, water content and food ingredient analyzers, and spectrophotometers (P9217 series). (Courtesy of Hamamatsu.)

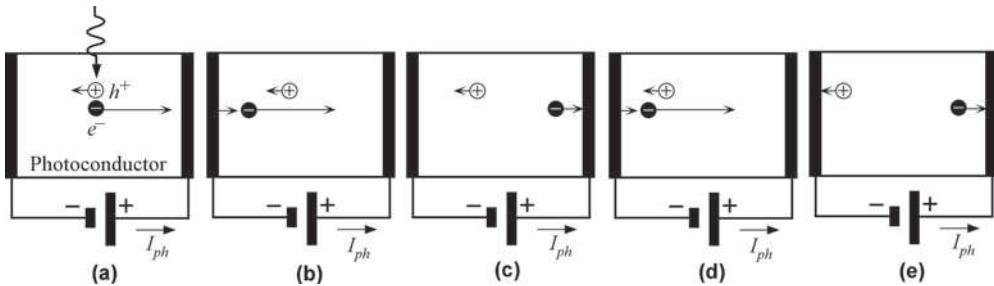


FIGURE 5.27 A photoconductor with ohmic contacts (contacts not limiting carrier entry) can exhibit gain. As the slow hole drifts through the photoconductors, many fast electrons enter and drift through the photoconductor because, at any instant, the photoconductor must be neutral. Electrons drift faster which means that as one leaves, another must enter. (a) An electron and hole are photogenerated. The hole drifts slowly but the electron drifts quickly and leaves the sample. (b) To maintain neutrality, an electron is injected from the negative electrode, which drifts quickly. (c) The electron reaches the positive electrode and leaves the sample while the hole is still slowly drifting. Figures (d) and (e) show the process in (b) and (c) repeated, except that the hole has drifted closer to the negative electrode. Eventually, the hole reaches the negative electrode, shortly after (e), and the photocurrent ceases. The hole can also recombine with one of the injected electrons during its drift.

and so on, until either the hole reaches the negative electrode or recombines with one of these electrons entering the sample. The external photocurrent therefore corresponds to the flow of many electrons per absorbed photon which represents a gain, called **photoconductive gain**. The gain depends on the drift time of the carriers and their recombination lifetime.

Suppose that the photoconductor is suddenly illuminated by a step light as shown in Figure 5.26. If Φ_{ph} is the number of photons arriving per unit area per unit second (the photon flux), then $\Phi_{ph} = I/h\nu$, where I is the light intensity (radiation energy flowing per unit area per second) and $h\nu$ is the photon energy. If all incident photons are absorbed then the number of EHPs generated per unit volume per second, i.e., the photogeneration rate per unit volume g_{ph} , is given by

$$g_{ph} = \frac{\eta_i A \Phi_{ph}}{Ad} = \frac{\eta_i \left(\frac{I}{h\nu}\right)}{d} = \frac{\eta_i I \lambda}{hcd} \quad (5.10.1)$$

where A is the illuminated surface area (ℓw) and η_i is the internal quantum efficiency.

Suppose that at any instant the electron concentration is n (includes photogenerated electrons) and the thermal equilibrium concentration (in the dark) is n_o . Then $\Delta n = n - n_o$ is the **excess electron concentration**. For photogeneration, $\Delta n = \Delta p$ inasmuch as we need to maintain charge neutrality in the sample. At any instant in the sample we have,

$$\left[\begin{array}{l} \text{The rate of increase in} \\ \text{the excess electron concentration} \end{array} \right] = \left[\begin{array}{l} \text{Rate of photogeneration} \\ \text{of excess electrons} \end{array} \right] - \left[\begin{array}{l} \text{Rate of recombination} \\ \text{of excess electrons} \end{array} \right]$$

If τ is the mean recombination time of excess electrons, then

$$\frac{d\Delta n}{dt} = g_{ph} - \frac{\Delta n}{\tau} \quad (5.10.2)$$

It is clear from Eq. (5.10.2) that Δn increases exponentially from the instant the light is turned on until a steady state is reached when

$$\frac{d\Delta n}{dt} = g_{ph} - \frac{\Delta n}{\tau} = 0$$

so that

$$\Delta n = \tau g_{ph} = \frac{\tau \eta_i \lambda}{hcd} \quad (5.10.3)$$

The conductivity of a semiconductor is given by $\sigma = e\mu_e n + e\mu_h p$, so that the change in the conductivity, called **photoconductivity**, is

$$\Delta\sigma = e\mu_e \Delta n + e\mu_h \Delta p = e\Delta n(\mu_e + \mu_h) \quad (5.10.4)$$

since electrons and holes are generated in pairs, $\Delta n = \Delta p$. Thus, substituting for Δn in the $\Delta\sigma$ expression above we get

$$\Delta\sigma = \frac{e\eta_i \lambda \tau (\mu_e + \mu_h)}{hcd} \quad (5.10.5)$$

The photocurrent density is simply

$$J_{ph} = \Delta\sigma \frac{V}{\ell} = \Delta\sigma E \quad (5.10.6)$$

where E is the applied field. The number of electrons flowing in the external circuit can be found from the photocurrent because

$$\text{Rate of electron flow} = \frac{I_{ph}}{e} = \frac{wdJ_{ph}}{e} = \frac{w\eta_i \lambda \tau (\mu_e + \mu_h) E}{hc} \quad (5.10.7)$$

However, the rate of electron (*i.e.*, EHP) photogeneration is

$$\text{Rate of electron generation} = (\text{Volume})g_{ph} = (wd\ell)g_{ph} = w\ell \frac{\eta_i \lambda}{hc}$$

The photoconductive gain is then simply

$$G = \frac{\text{Rate of electron flow in external circuit}}{\text{Rate of electron generation by light absorption}} = \frac{\tau(\mu_e + \mu_h)E}{\ell} \quad (5.10.8)$$

Equation (5.10.8) can be simplified further by noting that the drift velocities of the electrons and holes in the photoconductor are $\mu_e E$ and $\mu_h E$ respectively, so that their corresponding transit times (time to cross the semiconductor) are

$$t_e = \ell/(\mu_e E) \quad \text{and} \quad t_h = \ell/(\mu_h E)$$

Using these transit times in Eq. (5.10.8) we can obtain

$$G = \frac{\tau}{t_e} + \frac{\tau}{t_h} = \frac{\tau}{t_e} \left(1 + \frac{\mu_h}{\mu_e}\right) \quad (5.10.9)$$

The photoconductive gain can be quite high if τ/t_e is kept large, which means a long recombination time and a short transit time. The transit time can be made shorter by applying a

greater field but this will also lead to an increase in the dark current and more noise. The speed of response of the device is limited by the recombination time of the injected carriers. A long τ means a slow device.

5.11 BASIC PHOTODIODE CIRCUITS

The reverse biased photodiode (PD) mode of operation is shown in Figure 5.28 (a). We will assume steady state operation. There is usually a resistor R_L in the reverse bias circuit, which may be a load or a sampling resistor across which the signal is taken, or simply a current limiting resistor to reverse bias the PD. The definitions for positive I and V , by convention, are for the case in which the PD is forward biased; they are also shown in Figure 5.28 (a). The I - V characteristics of the photodiode are shown in Figure 5.28 (b), where both I and V have negative values representing reverse current and voltage. In the dark, there is a small reverse current, which usually increases with the reverse bias. (Ideally, it should be simply constant and equal to the reverse saturation current.) Upon illumination, a photocurrent I_{ph} is generated in the photodiode; I_{ph} is proportional to the incident light power P_o through the responsivity R . If the photodiode terminals were shorted, the short circuit current I_{sc} under illumination would be $-I_{ph}$ as indicated in Figure 5.28 (b).

The current through R and the voltage across it obey Ohm's law. However, care must be taken to properly account for the current and voltage signs. From Figure 5.28 (a), the current through R_L is

$$I = [(-V) - (+V_r)]/R_L = -(V + V_r)/R_L$$

This equation is called the **load line** because it represents the load R on the same axes I and V as the photodiode.¹⁷ For $I = 0$, open circuit $V = -V_r$, and this is a point on the load line. For $V = 0$, $I = -V_r/R$, and this is also a point on the load line. The load line can thus be quickly drawn to pass through these two points as shown in Figure 5.28 (b). It has a negative slope

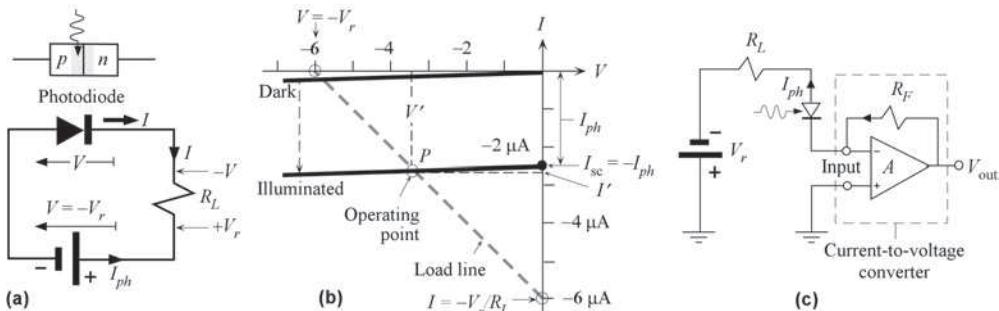


FIGURE 5.28 (a) The photodiode is reverse biased through R_L and illuminated. Definitions of positive I and V are shown as if the photodiode were forward biased. (b) I - V characteristics of the photodiode with positive I and V definitions in (a). The load line represents the behavior of the load R . $V_r = 6$ V and $R = 1 \text{ M}\Omega$ have been used. The operating point is P , where the current and voltage are I' and V' . (c) A simple circuit for the measurement of the photocurrent I_{ph} by using a current–voltage converter or a transimpedance amplifier. (The reverse bias V_r is a positive number.)

¹⁷It is left as an exercise to show that if we move the resistor R_L from the right to the left side of the circuit, we would get the same load line equation.

of $-1/R_L$. It cuts the PD characteristics at P , which represents the operating point of the circuit; the current and voltage will be I' and V' as indicated in Figure 5.28 (b). The particular example in Figure 5.28 (b) is for $V_r = 6\text{ V}$ and $R = 1\text{ M}\Omega$.

It is apparent from Figure 5.28 (b) that as long as the PD is reverse biased, the magnitude of the operating current I' under illumination is very close to the generated photocurrent I_{ph} (the short circuit current), i.e., $I' \approx -I_{ph}$. This is not the case in the photovoltaic mode in which there is a load resistor right across the PD without any V_r . Further, we can redirect the photocurrent into a current-to-voltage converter to convert I_{ph} to an output voltage V_{out} as in Figure 5.28 (c), in which the output $V_{out} = R_F I_{ph}$. The feedback resistor R_F determines the transimpedance gain V_{out}/I_{ph} . Notice that V_{out} increases with increasing I_{ph} , that is, the gain is positive.

There are many applications for which we need to represent the PD by an equivalent circuit. The basic operation of a reverse biased PD is shown in Figure 5.29 (a) where A and B represent the external terminals of the PD. Suppose the EHP generation occurs in the SCL; then the drift of electrons and holes will give rise to a photocurrent I_{ph} . We can represent this effect with a current generator in the equivalent circuit as shown in Figure 5.29 (b). The points A' and B' represent the terminals of the ideal pn junction diode that generates the photocurrent I_{ph} . The origin of the capacitance C_t and resistances R_s and R_p in the equivalent circuit are explained below. A and B are the actual external terminals of the PD. The definitions for positive current and voltage are shown as I and V . I flows into A and V is the voltage at A with respect to B . Notice that, due to a voltage drop across R_s , the voltage that appears across the pn junction (between A' and B') is V_1 .

If we are not limited by the drift and diffusion times of the photogenerated carriers, then I_{ph} will follow the input optical power modulation. Under reverse bias, $V = -V_r$, and the voltage across the ideal pn junction V_1 is nearly $-V_r$ because R_s is usually small. There will also be some dark current I_d from the normal diode action (or pn junction behavior), i.e., with reverse bias, $I_{diode} = I_{so}[\exp(eV_1/k_B T) - 1] = -I_{so}$, reverse saturation current. This reverse current is represented by an ideal diode D in parallel with the current generator as shown in Figure 5.29 (b). I_{diode} is the ideal diode current and along the positive current direction for D . The dark current I_d is quoted as a positive number in data sheets so that $I_d = -I_{diode} = I_{so}$.

Since the PD is reverse biased, there will be a depletion region capacitance C_{dep} appearing across the PD terminals as in Figure 5.29 (a). We also need to consider any stray capacitances, including terminal to terminal capacitance (so-called packaging capacitances), all of

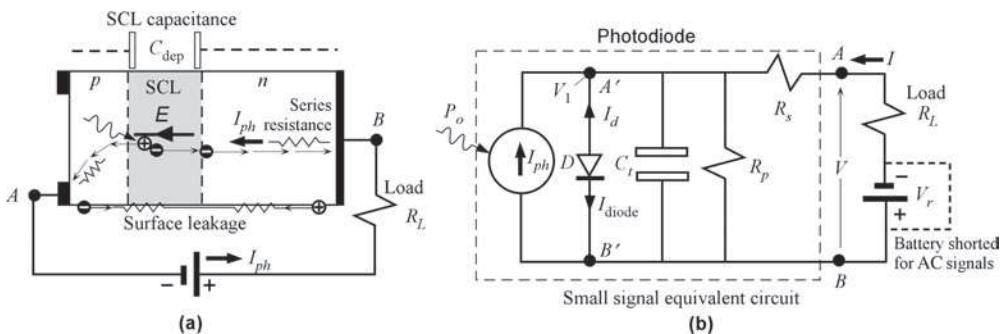
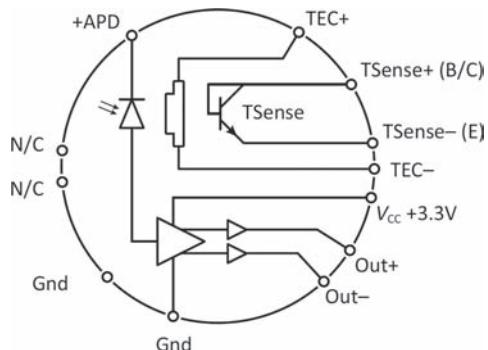
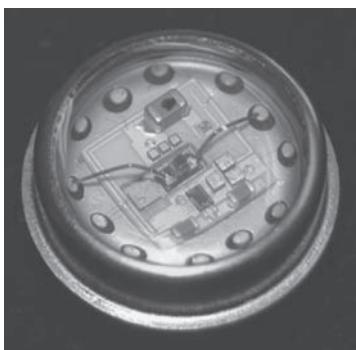


FIGURE 5.29 (a) A real photodiode has series and parallel resistances R_s and R_p and an SCL capacitance C_{dep} . A and C represent anode and cathode terminals. (b) The equivalent circuit of a photodiodes. For AC (or transient) signals, the battery can be shorted since AC signals will simply pass through the battery.

which will add to C_{dep} , and can be lumped into a single total capacitance called the **terminal capacitance** C_t . In Figure 5.29 (b), C_t appears across the photocurrent generator. Because of C_{dep} , C_t decreases with increasing reverse bias, and is quoted at a certain bias voltage in data sheets. It depends on the diode area, and can be anything from a few picofarads (small area PDs) to a few nanofarads.

The photocurrent inside the PD also has to flow through the neutral *p*- and *n*-sides, that is, through bulk semiconductor regions that will have a finite resistance (not zero) as shown in Figure 5.29 (a). We can represent the bulk semiconductor regions by introducing a **series resistance** R_s into the equivalent circuit as shown in Figure 5.29 (b). There will also be surface leakage currents which represent the flow of carriers on the surfaces of the crystal. The net resistance of all these surface paths are represented by using a **shunt resistance** or a **parallel resistance** R_p across the current generator as shown in Figure 5.29 (b). The load R_L is connected to the external terminals *A* and *B* of this equivalent circuit. Although Figure 5.29 (b) has also included the battery (V_r) to indicate reverse bias, this would be shorted in using the circuit for AC signals. The equivalent circuit shown in Figure 5.29 (b) is probably one of the simplest but quite useful in predicting the output from a PD. It can also be used for photovoltaic behavior (without a battery but just a load across *A* and *B*).



A photoreceiver that has an InGaAs APD and peripheral electronics (ICs) to achieve high gain and high sensitivity. There is also a thermoelectric cooler (TEC) and a temperature sensor (TSense). (Courtesy of Voxel Inc.)

Photodiodes are widely used in optoelectronics to measure modulated light signals, which may be digital, in the form of optical pulses, or analog-type, in which the amplitude of the light wave is modulated. As a rough estimate of the PD speed, we can assume R_p is very large, and can be ignored. Further, R_s and the load R_L in Figure 5.29 (b) can be lumped into one resistance, but R_s is typically much smaller than R_L . We then have C_t and R_L in parallel with a time constant $R_L C_t$. The cutoff frequency f_c or the bandwidth of the PD will then be determined by $R_L C_t$, i.e., $f_c = 1/(2\pi R_L C_t)$. Manufacturers usually quote C_t and the bandwidth (the cutoff frequency) for a given load (R_L). Suppose that we make $R_L C_t$ as small as possible, then the bandwidth will be determined by the transit and diffusion times of carriers inside the PD as discussed above.

We can easily build a simple and fast PD circuit as shown in Figure 5.30 (a). The PD is reverse biased through R_B , and C_B shorts both R_B and the battery for transient signals; it is an AC short. There is a small load or a sampling resistor R_L . The transient photocurrent generated by an optical pulse develops a transient voltage across R_L , which can be coupled into a fast buffer amplifier.

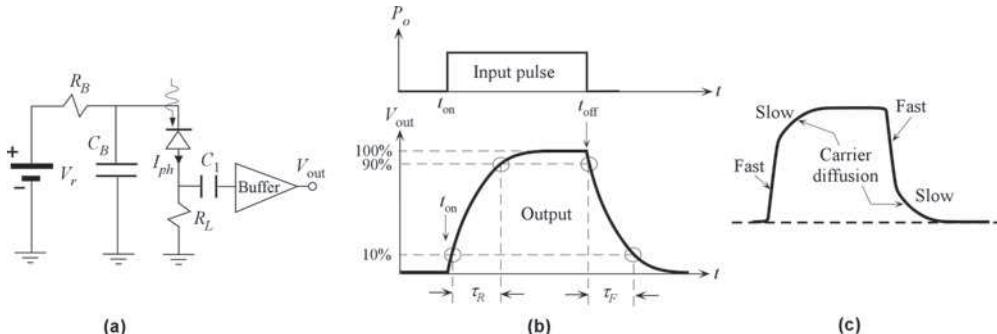


FIGURE 5.30 (a) The photodiode is reverse biased by V_r through R_B . C_B shorts both the R_B and the battery for transient signals. C_1 is a signal coupling capacitor to pass the photocurrent signal into a fast buffer amplifier. (b) Definitions of rise and fall times in the PD output in response to a pulse optical excitation. (c) Carrier diffusion, which is a slow process, results in slow regions in the output pulse shape.

The time constant $R_L C_t$ must be smaller than the optical pulse to be measured. C_t in the circuit of Figure 3.30 (a) should include the capacitance of the buffer amplifier as well, which maybe comparable to that of the PD. Instead of using a sampling resistor R_L , one can, of course, feed the output directly into the input of a fast current to voltage converter amplifier as in Figure 5.28 (c). The critical time constant is then $R_F C_r$.

Manufacturers' data sheets typically provide rise and fall times τ_R and τ_F in the output waveform when the PD is excited from an optical pulse. τ_R is time required for the output to rise from its 10% to 90% of final value. τ_F is the time needed for the output to fall from its 90% to 10% value after it has reached a stated state before the turn-off. Both definitions are shown in Figure 5.30 (b). When the $R_t C_t$ time constant is very small, the rise and fall times are limited by the drift and diffusion of photogenerated carriers in the PD. If the photogeneration occurs in the depletion region the rise and fall times would be quite short and limited by transit times across the depletion width W .

The exact output waveform depends on the excitation wavelength and the PD structure. For long wavelength excitation, for which the absorption depth is long, photogeneration will occur both in the SCL (width, W) and within diffusion lengths L_e and L_h from the SCL as shown in Figure 5.8. The output waveform will have an initial fast and then a slow region in both the rise and fall parts of the waveform as shown in Figure 5.30 (c). Initially, the photogenerated carriers in the SCL will drift very quickly through W and give a fast rise in the output. Those carriers generated outside the SCL have to slowly diffuse to the SCL. Upon arriving at the SCL, they will be drifted quickly. The prolonged arrival of these diffusing carriers is responsible for the slow rise part of the waveform. As soon as the optical pulse is turned off, those carriers in W can be drifted and collected quickly, and give rise to a fast initial fall. However, those carriers that are generated outside W have to diffuse to the SCL before they can be drifted and collected, and this gives rise to the slowly decaying final part in the waveform.

5.12 NOISE IN PHOTODETECTORS

A. The *pn* Junction and *pin* Photodiodes

The lowest signal that a photodetector can detect is determined by the extent of random fluctuations in the current through the detector and the voltage across it as a result of various statistical processes in the device. When a *pn* junction is reverse biased, there is still a dark

current I_d present, which is mainly due to the thermal generation of electron–hole pairs in the SCL and within diffusion lengths to the SCL. If the dark current were absolutely constant with no fluctuations, then any change in the PD current, however small (even a tiny fraction of I_d), due to an optical signal could be easily detected by blocking or removing I_d . The dark current however exhibits **shot noise** or fluctuations about I_d , as shown in Figure 5.31. This shot noise is due to the fact that electrical conduction is by *discrete charges*, which means that there is a statistical distribution in the transit times of the carriers across the photodiode. Carriers are collected as discrete amounts of charge (e) that arrive at random times and not continuously. It is unlike a continuous flow of water through a pipe but rather like rolling ball bearings down a pipe at random times. There will be fluctuations in the outflow rate of ball bearings at the collection end.

The *root mean square* (rms) value of the fluctuations in the dark current represents the shot noise current $i_{n\text{-dark}}$,

$$i_{n\text{-dark}} = [2eI_dB]^{1/2} \quad (5.12.1)$$

Dark current shot noise

where B is the frequency bandwidth of the photodetector. The photocurrent signal must be greater than this shot noise in the dark current.

The photodetection process involves the interaction of discrete photons with valence electrons. The discrete nature of photons means that there is an unavoidable random fluctuation in the rate of arrival of photons even if we did our best to keep the rate constant. The quantum nature of the photon therefore gives rise to a statistical randomness in the EHP photogeneration process. This type of fluctuation is called **quantum noise** (or **photon noise**) and it is equivalent to shot noise as far as its effects are concerned. Thus, the photocurrent will always exhibit fluctuations about its mean value due to quantum noise. If I_{ph} is the mean photocurrent, the fluctuations about this mean has an rms value that is called **shot noise current due to quantum noise**

$$i_{n\text{-quantum}} = [2eI_{ph}B]^{1/2} \quad (5.12.2)$$

Quantum noise

Generally the dark current shot noise and quantum noise are the main sources of noise in *pn* junction– and *pin*-type photodiodes. The total shot noise i_n generated by the photodetector is not a simple sum of Eqs. (5.12.1) and (5.12.2) because the two processes are due to independent

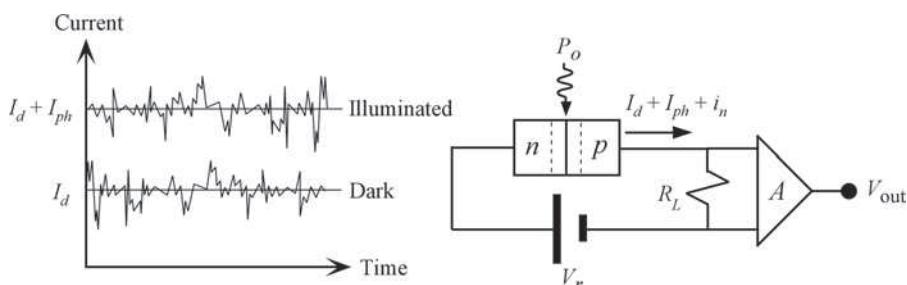


FIGURE 5.31 In *pn* junction and *pin* devices, the main source of noise is shot noise due to the dark current and photocurrent.

random fluctuations. We need to sum the power in each or add the mean squares of the shot noise currents, *i.e.*,

$$i_n^2 = i_{n\text{-dark}}^2 + i_{n\text{-quantum}}^2$$

so that the rms total shot noise current is

*Shot noise
in pn
junction*

$$i_n = [2e(I_d + I_{ph})B]^{1/2} \quad (5.12.3)$$

In Figure 5.31 the photodetector current, $I_d + I_{ph} + i_n$, flows through a load resistor R_L , which acts as a sampling resistor for measuring the current. The voltage across R_L is amplified. In considering the noise of the receiver we also have to include the thermal noise in the resistor R_L and the noise in the input stage of the amplifier. **Thermal noise** is random voltage fluctuations across any conductor due to random motions of the conduction electrons; it is also called **Johnson noise**. In receiver design we are often interested in the **signal to noise ratio**, SNR or S/N , which is defined as the ratio of signal power to noise power,¹⁸

*Signal
to noise
ratio*

$$\text{SNR} = \frac{\text{Signal power}}{\text{Noise power}} \quad (5.12.4)$$

For the photodetector alone, SNR is simply the ratio of I_{ph}^2 to i_n^2 . SNR for the receiver must include the noise power generated in the sampling resistor R_L (thermal noise) and in the input elements of the amplifier (*e.g.*, resistors and transistors).

The **noise equivalent power** (NEP) is an important property of a photodetector that is frequently quoted. NEP is defined as the required optical input power to achieve a SNR of 1 within a bandwidth of 1 Hz. Stated differently, it is the optical signal power required to generate a photocurrent signal (I_{ph}) that is equal to the total noise current (i_n) in the photodetector at a given wavelength and within a bandwidth of 1 Hz. Suppose that B is the bandwidth of the detector and P_1 is the incident power that results in $\text{SNR} = 1$, then by definition

*NEP
definition*

$$\text{NEP} = \frac{\text{Input power for } \text{SNR} = 1}{\sqrt{\text{Bandwidth}}} = \frac{P_1}{B^{1/2}} \quad (5.12.5)$$

If R is the responsivity and P_o is the monochromatic incident optical power, then the generated photocurrent is

$$I_{ph} = RP_o \quad (5.12.6)$$

Suppose that the photogenerated current I_{ph} is equal to the noise current i_n in Eq. (5.12.3), when the incident optical power P_o is P_1 . Then

$$RP_1 = [2e(I_d + I_{ph})B]^{1/2}$$

¹⁸Some authors define SNR as the ratio of the actual signal magnitude to the rms noise, that is, the square root of the definition in Eq. (5.12.4). The definition used here is common in optical communications. On the other hand, imaging applications tend to define SNR based on the signal magnitude and rms noise.

From this we can find the optical power per square root of bandwidth and hence the NEP from Eq. (5.12.5) as

$$\text{NEP} = \frac{P_1}{B^{1/2}} = \frac{1}{R} [2e(I_d + I_{ph})]^{1/2} \quad (5.12.7)$$

Noise equivalent power

It is clear that if we put $B = 1 \text{ Hz}$, we obtain numerically $\text{NEP} = P_1$, that value of P_o which makes I_{ph} equal to the total noise current i_n . From Eq. (5.12.7), the units for NEP are $\text{W Hz}^{-1/2}$. It is clear from the definition of NEP that we need to lower the NEP for better photodetection.

We can easily extend the NEP definition to photoconductive (PC) detectors as well. In this case, the noise added by the detector is the thermal noise of its resistance and that from the bias or load resistor. The NEP definition stays the same as in Eq. (5.12.5), though Eq. (5.12.7) would not be applicable as we need to consider the thermal noise generated by the photoconductor and its load resistor.

The **detectivity** D is defined as the reciprocal of noise equivalent power (NEP), $D = 1/\text{NEP}$; it is a measure of detector's sensitivity taking into account various noise contributions. However, for a proper comparison of detectors, we should consider all detectors to have the same photosensitive area. If A is the photosensitive area receiving the radiation, then **specific detectivity** D^* is defined by

$$D^* = \frac{A^{1/2}}{\text{NEP}} \quad (5.12.8)$$

Specific detectivity

The units are $\text{m Hz}^{1/2} \text{ W}^{-1}$. However, $\text{cm Hz}^{1/2} \text{ W}^{-1}$ is more common, and is called Jones. Typically, the detectivity increases as the temperature of the detector is lowered due to the lowering of the dark current. Table 5.5 lists typical NEP and D^* values for a range of photodetectors. Photoconductive detectors such as PbS, PbSe, and InSb are based on measuring their photoconductivity, that is, the change in the resistance upon illumination. They are used for the detection of radiation at long wavelengths, *e.g.*, longer than $\sim 2 \mu\text{m}$. Notice also that these photoconductive detectors have lower detectivity than wider bandgap photodiodes used at shorter wavelengths, and have to be cooled.

According to Eq. (5.12.7), the NEP should depend on the dark current as $I_d^{1/2}$. Figure 5.32 shows a plot of NEP vs. I_d on a log-log scale for a collection of commercial Si and InGaAs *pin*, Ge *pn* and GaAsP Schottky junction photodiodes. NEP does indeed increase with I_d , very roughly as $I_d^{1/2}$ for a given class of photodiodes. The InGaAs *pin* photodiode case covers the widest range of dark currents, and involves room temperature, -10°C and -20°C values.

TABLE 5.5 Typical noise characteristics of a few selected commercial photodetectors

Photodiode	GaP Schottky	Si <i>pin</i>	Ge <i>pin</i>	InGaAs <i>pin</i>	PbS (PC) -10°C	PbSe (PC) -10°C	InSb (PC) -10°C
$\lambda_{\text{peak}} (\mu\text{m})$	0.44	0.96	1.5	1.55	2.4	4.1	5.5
I_d or R_d	10 pA	0.4 nA	3 μA	5 nA	0.1–1 M Ω	0.1–1 M Ω	1–10 k Ω
NEP ($\text{W Hz}^{-1/2}$)	5.4×10^{-15}	1.6×10^{-14}	1×10^{-12}	4×10^{-14}			–
D^* ($\text{cm Hz}^{1/2} \text{ W}^{-1}$)	1×10^{13}	1×10^{12}	1×10^{11}	5×10^{12}	1×10^9	5×10^9	1×10^9

Notes: PC means a photoconductive detector, whose photoconductivity is used to detect light. For PC detectors, what is important is the dark resistance R_d , which depends on the temperature.

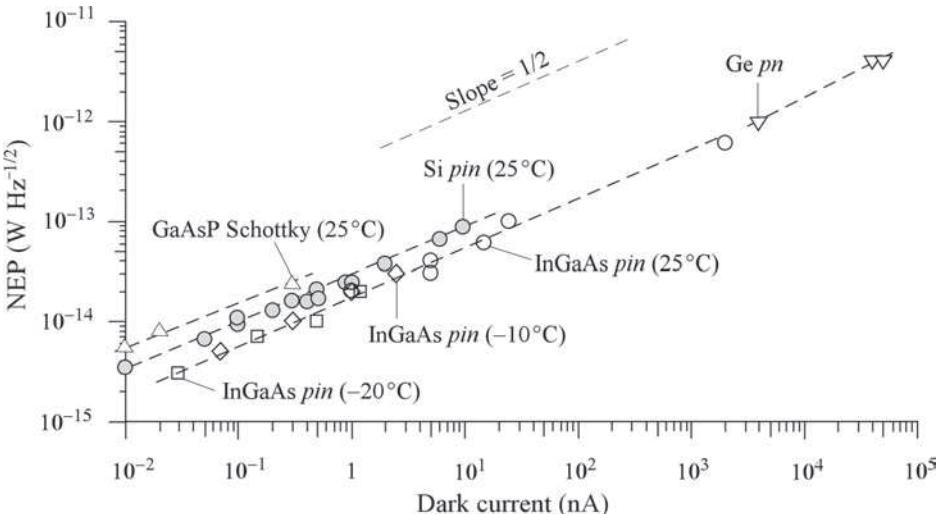


FIGURE 5.32 The dependence of $\text{NEP}(\text{W Hz}^{-1/2})$ on the photodetector dark current I_d for Si and InGaAs *pin*, Ge *pn* junction, and GaP Schottky photodiodes. Dashed lines indicate observed trends. Filled circle, Si *pin*; open circle, InGaAs *pin* at 25°C ; open diamond at -10°C ; open square at -20°C ; inverted triangle, Ge *pn*; triangle, GaAsP Schottky. (Source: Data extracted from the datasheets of 35 commercial photodiodes.)

I_d decreases with decreasing temperature, and so does NEP, roughly along the $\text{NEP}-I_d^{1/2}$ line. In fact, at -20°C , InGaAs *pin* photodiodes NEP is quite low, $\sim 3 \times 10^{-13} \text{ W Hz}^{-1/2}$, which is the reason that it is cooled to achieve higher detectability.

EXAMPLE 5.12.1 NEP of a Si *pin* photodiode

A Si *pin* photodiode has a quoted NEP of $1 \times 10^{-13} \text{ W Hz}^{-1/2}$. What is the optical signal power it needs for a signal to noise ratio (SNR) of 1 if the bandwidth of operation is 1 GHz?

Solution

By definition, NEP is that optical power per square root of bandwidth which generates a photocurrent equal to the noise current in the detector.

$$\text{NEP} = P_1/B^{1/2}$$

Thus,

$$P_1 = \text{NEP}B^{1/2} = (10^{-13} \text{ W Hz}^{-1/2})(10^9 \text{ Hz})^{1/2} = 3.16 \times 10^{-9} \text{ W} \quad \text{or} \quad 3.16 \text{ nW}$$

EXAMPLE 5.12.2 Noise of an ideal photodetector

Consider an ideal photodiode with $\eta_e = 1$ (QE = 100%) and no dark current, $I_d = 0$. Show that the minimum optical power required for a signal to noise ratio (SNR) of 1 is

$$P_1 = \frac{2hc}{\lambda} B \quad (5.12.9)$$

Calculate the minimum optical power for a SNR = 1 for an ideal photodetector operating at 1300 nm with a bandwidth of 1 GHz. What is the corresponding photocurrent?

Solution

We need the incident optical power P_1 that makes the photocurrent I_{ph} equal to the noise current i_n , so that SNR = 1. The photocurrent (signal) is equal to the noise current when

$$I_{ph} = i_n = [2e(I_d + I_{ph})B]^{1/2} = (2eI_{ph}B)^{1/2}$$

since $I_d = 0$. Solving the above, $I_{ph} = 2eB$

From Eqs. (5.4.3) and (5.4.4), the photocurrent I_{ph} and the incident optical power P_1 are related by

$$I_{ph} = \frac{\eta_e e P_1 \lambda}{hc} = 2eB$$

Thus,

$$P_1 = \frac{2hc}{\eta_e \lambda} B$$

For an ideal photodetector, $\eta_e = 1$, which leads to Eq. (5.12.9). We note that for a bandwidth of 1 Hz, NEP is numerically equal to P_1 or $\text{NEP} = 2hc/\lambda$.

For an ideal photodetector operating at 1.3 μm and at 1 GHz

$$\begin{aligned} P_1 &= 2hcB/\eta_e \lambda = 2(6.63 \times 10^{-34} \text{ J s})(3 \times 10^8 \text{ m s}^{-1})(10^9 \text{ Hz})/(1)(1.3 \times 10^{-6} \text{ m}) \\ &= 3.1 \times 10^{-10} \text{ W} \quad \text{or} \quad 0.31 \text{ nW} \end{aligned}$$

This is the minimum signal for an SNR = 1. The noise current is due to quantum noise. The corresponding photocurrent is

$$I_{ph} = 2eB = 2(1.6 \times 10^{-19} \text{ C})(10^9 \text{ Hz}) = 3.2 \times 10^{-10} \text{ A} \quad \text{or} \quad 0.32 \text{ nA}$$

Alternatively, we can calculate I_{ph} from $I_{ph} = \eta_e e P_1 \lambda / hc$ with $\eta_e = 1$.

EXAMPLE 5.12.3 SNR of a receiver

Consider an InGaAs *pin* photodiode used in a receiver circuit as in Figure 5.31 with a load resistor of 10 kΩ. The photodiode has a dark current of 2 nA. The bandwidth of the photodiode and the amplifier together is 1 MHz. Assuming that the amplifier is noiseless, calculate the SNR when the incident optical power generates a mean photocurrent of 5 nA (corresponding to an incident optical power of about 6 nW since R is about 0.8–0.9 nA/nW at the peak wavelength of 1550 nm).

Solution

The noise generated comes from the photodetector as shot noise and from R_L as thermal noise. The mean thermal noise power in the load resistor R_L is $4k_B TB$. If I_{ph} is the photocurrent and i_n is the shot noise in the photodetector then

$$\text{SNR} = \frac{\text{Signal power}}{\text{Noise power}} = \frac{I_{ph}^2 R_L}{i_n^2 R_L + 4k_B TB} = \frac{I_{ph}^2}{[2e(I_d + I_{ph})B] + 4k_B TB/R_L}$$

The term $4k_B TB/R_L$ in the denominator represents the mean square of the thermal noise current in the resistor. We can evaluate the magnitude of each noise current by substituting, $I_{ph} = 5 \text{ nA}$, $I_d = 2 \text{ nA}$, $B = 1 \text{ MHz}$, $R_L = 10^4 \Omega$, $T = 300 \text{ K}$.

$$\text{Shot noise current from the detector} = [2e(I_d + I_{ph})B]^{1/2} = 0.047 \text{ nA}$$

$$\text{Thermal noise current from } R_L = \left(\frac{4k_B TB}{R_L} \right)^{1/2} = 1.29 \text{ nA}$$

Thus, the noise contribution from R_L is greater than that from the photodiode. The SNR is

$$\text{SNR} = \frac{(5 \times 10^{-9} \text{ A})^2}{(0.047 \times 10^{-9} \text{ A})^2 + (1.29 \times 10^{-9} \text{ A})^2} = 15.0$$

Generally SNR is quoted in decibels. We need $10\log(\text{SNR})$, or $10\log(15.0)$, i.e., 11.8 dB. Clearly, the load resistance has a dramatic effect on the overall noise performance.

B. Avalanche Noise in the APD

In the avalanche photodiode, both photogenerated and thermally generated carriers entering the avalanche zone are multiplied. The shot noise associated with these carriers are also multiplied. If I_{do} and I_{pho} are the unmultiplied ($M = 1$) dark current and photocurrent (primary photocurrent), respectively, in the APD then the shot noise current (as an rms value) in the APD should be

$$i_{n-\text{APD}} = M[2e(I_{do} + I_{pho})B]^{1/2} = [2e(I_{do} + I_{pho})M^2B]^{1/2} \quad (5.12.10)$$

The APDs exhibit excess avalanche noise that is above this multiplied shot noise of the photocurrent and dark current. This excess noise is due to the randomness of the impact ionization process in the multiplication region. Some carriers travel far and some short distances within this zone before they cause impact ionization. Furthermore, the impact ionization does not occur uniformly over the multiplication region but is more frequent in the highest field zone. Thus, the multiplication M fluctuates about a mean value. The result of the statistics of impact ionization is an *excess noise* contribution, called **avalanche noise**, to the multiplied shot noise. The noise current in an APD is then given by

APD noise current

$$i_{n-\text{APD}} = [2e(I_{do} + I_{pho})M^2FB]^{1/2} \quad (5.12.11)$$

where F is called the **excess noise factor** and is a function of M and the impact ionization probabilities (called coefficients). Generally, F is approximated by the relationship $F \approx M^x$ where x is an index that depends on the semiconductor, the APD structure, and the type of carrier that initiates the avalanche (electron or hole). For Si APDs x is 0.3–0.5, whereas for Ge and III–V (such as InGaAs) alloys it is 0.7–1.

EXAMPLE 5.12.4 Noise in an APD

Consider an InGaAs APD with $x \approx 0.7$ which is biased to operate at $M = 10$. The unmultiplied dark current is 10 nA and bandwidth is 700 MHz.

- (a) What is the APD noise current per square root of bandwidth?
- (b) What is the APD noise current for a bandwidth of 700 MHz?
- (c) If the responsivity (at $M = 1$) is 0.8 A W^{-1} what is the minimum optical power for a SNR of 10?

Solution

- (a) In the absence of any photocurrent, the noise in the APD comes from the dark current. If the unmultiplied dark current is I_{do} then using $F = M^x$ in Eq. (5.12.11), the noise current (rms) is

$$i_{n\text{-dark}} = [2eI_{do}M^{2+x}B]^{1/2} \quad (5.12.12)$$

Thus,

$$\begin{aligned} \frac{i_{n\text{-dark}}}{\sqrt{B}} &= \sqrt{2eI_{do}M^{2+x}} = \sqrt{2(1.6 \times 10^{-19} \text{ C})(10 \times 10^{-9} \text{ A})(10)^{2+0.7}} \\ &= 1.27 \times 10^{-12} \text{ A Hz}^{-1/2} \quad \text{or} \quad 1.27 \text{ pA Hz}^{-1/2} \end{aligned}$$

- (b) In a bandwidth B of 700 MHz, the noise current is

$$i_{n\text{-dark}} = (700 \times 10^6 \text{ Hz})^{1/2}(1.27 \text{ pA Hz}^{-1/2}) = 3.35 \times 10^{-8} \text{ A} \quad \text{or} \quad 33.5 \text{ nA}$$

- (c) The SNR with a primary photocurrent I_{pho} in the APD is

$$\text{SNR} = \frac{\text{Signal power}}{\text{Noise power}} = \frac{M^2 I_{pho}^2}{[2e(I_{do} + I_{pho})M^{2+x}B]} \quad (5.12.13) \quad \text{SNR APD}$$

Rearranging to obtain I_{pho} we get

$$(M^2)I_{pho}^2 - [2eM^{2+x}B(\text{SNR})]I_{pho} - [2eM^{2+x}B(\text{SNR})I_{do}] = 0$$

This is a quadratic equation in I_{pho} with defined coefficients since M , x , B , I_{do} , and SNR are given. Solving this quadratic with a $\text{SNR} = 10$ for I_{pho} we find

$$I_{pho} \approx 1.76 \times 10^{-8} \text{ A} \quad \text{or} \quad 17.6 \text{ nA}$$

While it may seem odd that I_{pho} is less than the dark noise current (33.5 nA) itself, the actual photocurrent I_{ph} is 176 nA, since it is multiplied by M . Further, the total noise current $i_{n\text{-APD}} = [2e(I_{do} + I_{pho})M^{2+x}B]^{1/2}$ is 55.7 nA, so that one can easily check that $\text{SNR} = I_{ph}^2/i_{n\text{-APD}}^2$ is indeed 10.

By the definition of responsivity, $R = I_{pho}/P_o$, we find

$$P_o = I_{pho}/R = (1.76 \times 10^{-8} \text{ A})/(0.8 \text{ A W}^{-1}) = 2.2 \times 10^{-8} \text{ W} \quad \text{or} \quad 22 \text{ nW}$$

5.13 IMAGE SENSORS

A. Basic Principles

An image sensor, as illustrated in Figure 5.33 (a), is an integrated circuit chip made up of an array of photosensitive elements, that is able to capture an image and provide an output in the form of an electrical signal such as current, charge, or voltage. The output is usually put through a multiplexer and an analog-to-digital converter to obtain a digital form of the image as shown in Figure 5.33 (b). The sensor usually consists of an array of elements, called **pixels** (*pixel* from “picture element”), in N rows and M columns, where each element or pixel has a detector and is able to provide an electrical signal that is proportional to the light intensity received at this pixel. A lens forms an image of the object on the image sensor, and the light intensity $I(X,Y)$ at each point of this image (on the sensor) excites the corresponding pixel at the X,Y location. The image point X,Y excites the pixel at that location, that is, the pixel in row i and column j , and the

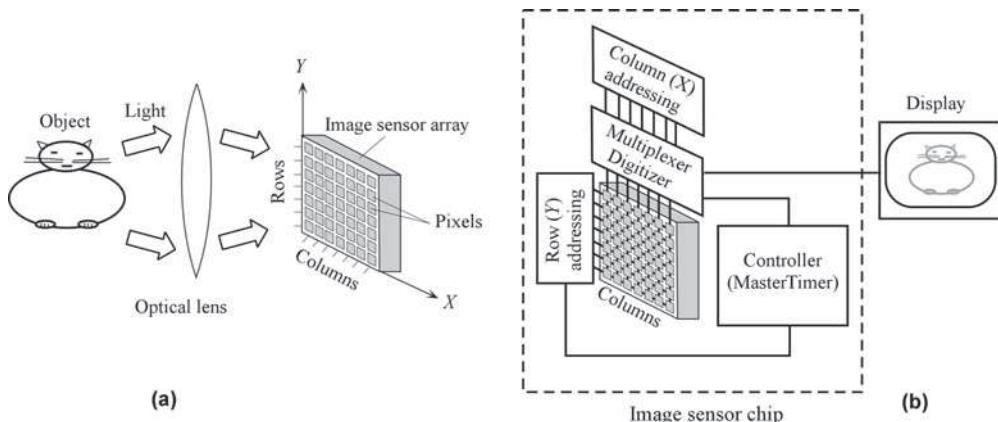


FIGURE 5.33 (a) The basic image-sensing operation using an array of photosensitive pixels. (b) The image sensor chip that incorporates the auxiliary electronics that run the sensor array (CMOS technology).

pixel (i,j) has an electrical signal (e.g., charge) proportional to $I(X,Y)$. Thus, each pixel (i,j) carries a piece of the image information, a pixel of it.

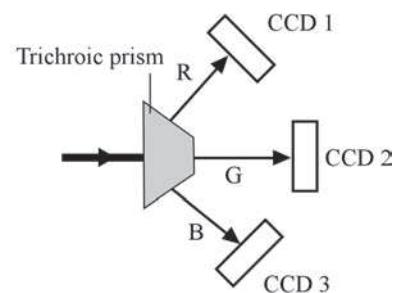
The imaging array shown in Figure 5.33 (b) is addressed from a master timer to read out the signals (charges) at the pixels and hence provide an electronically recorded image. There are two major solid state technologies: **charge-coupled device** (CCD) and **complementary metal oxide semiconductor** (CMOS) sensors, each with its own advantages and drawbacks as discussed in Sections B and C below. CMOS refers to the technology that is used to fabricate the sensor. In the CCD technology, the incident light at a pixel generates charge that is stored in the CCD pixel (called a well) at this location (and discussed below). This charge (signal) is read out by a sequential clocking mechanism, a characteristic of the CCD operation, to shift the charges to an eventual shift register for read-out. The CMOS sensor is more like the image sensor chip shown in Figure 5.33 (b). Each pixel has MOS transistors and a photodiode. The electrical signals at pixels are read out line by line. A row i is addressed by the controller, and all the pixels in this row provide their signals to the columns ($j = 1$ to M). The parallel data are multiplexed and digitized to provide a digital image. The image sensor quantizes the image because it breaks into $N \times M$ number of pixels. The pixel size imposes a resolution limit because the image is sampled by the pixels, which are of finite size.

Color rendering with image sensors can be done in three ways. The most common is to use red, green, and blue filters to separate out the colors at a point onto three different adjacent pixels,

FIGURE 5.34 (a) Color imaging by using two different techniques. Bayer filtering uses red (R), green (G), and blue (B) filters on three pixels for capturing the R, G, and B information. (b) Three CCD sensor uses a trichroic prism to separate the image colors into red, green, and blue, and uses three CCD chips for each.

R	G	R	G
G	B	G	B
R	G	R	G
G	B	G	B

Bayer filtering



Three CCD color image sensor

(a)

(b)

such as in the **Bayer color image sensor** shown in Figure 5.34 (a). In **three CCD imaging**, as shown in Figure 5.34 (b), a trichroic prism assembly is used to separate the light from the object into its red, green, and blue components, and three different CCDs are used to detect each component. A third technique uses different absorption depths for red, green, and blue light in silicon. The red, green, and blue light have different penetration depths into silicon so that, in principle, we can use photodiodes at different depths to distinguish between the colors. The Bayer filtering technique is the most common among color image sensors.

B. Active Matrix Array and CMOS Image Sensors

Many image sensors, as well as various flat panel TV technologies, make use of active matrix arrays (AMAs). An **active matrix array** is a two-dimensional array of pixels in which each pixel has a thin film transistor (a field effects transistor that has been fabricated by using thin film technologies) that can be externally addressed to read a signal from a sensor located at that pixel. Depending on the application, an AMA can have few pixels or millions of pixels. The TFT AMA technology was pioneered by Peter Brody using CdSe TFTs in early 1970s. A simplified diagram is shown in Figure 5.35 (a) for the array and in (b) for each pixel. Each pixel has a photodiode and a capacitor C_{px} . When a pixel receives light from a particular point on the illuminated object, the photodiode generates a current I_{signal} which charges the pixel capacitor C_{px} . The signal is the charge Q_{signal} stored on C_{px} . The array of pixels therefore has the image stored as charges on pixel capacitances. All we have to do is read the charges out.

Each pixel is identical with its TFT gate connected to a particular address line and the source to a particular data line as shown in Figure 5.35 (b). When row a is addressed, all pixels and hence all signals on that row are read out, that is, the data are read out onto the data lines as *parallel data*, which can be converted into serial data. We can scan the whole image row by row (line by line) by starting from the top row and sequentially activating one row after another. Thus, each time we activate and access a row, we extract the signals (Q_{signal} on each C_{px}) on that row as parallel data.

A **CMOS image sensor** is basically an AMA in which each pixel has a photodiode or a photogate and one or more CMOS transistors to read and amplify the electrical signal that has been

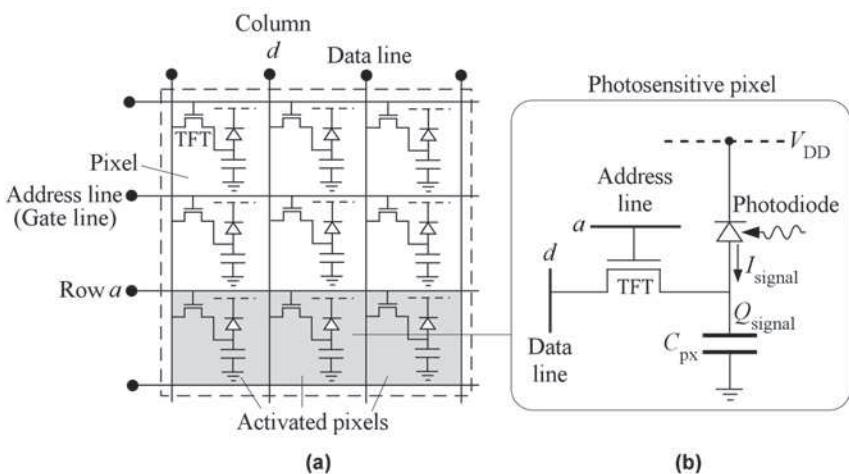


FIGURE 5.35 (a) An active matrix array (AMA). (b) A basic photosensitive pixel structure for detecting the photons arriving at the pixel defined by row a and column d .

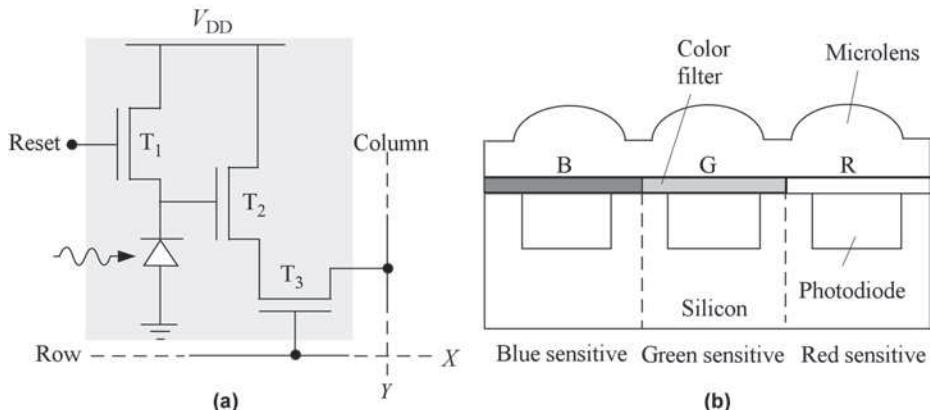
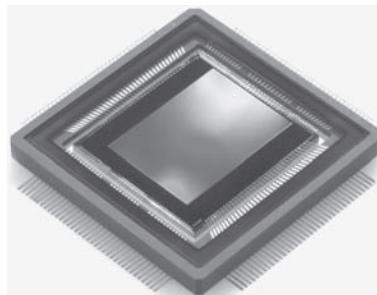


FIGURE 5.36 (a) The pixel architecture in a CMOS image sensor. (b) A cross-section(s) of a CMOS imager with microlenses and color filters (B = blue, G = green, R = red) for color imaging.

generated by the light incident on this pixel. The pixel read-out operation is essentially the same as that in the AMA. The pixel architecture, however, can change from application to application.

CMOS is a well-established silicon fabrication technology, which means that CMOS image sensors can be fabricated relatively inexpensively. If the pixel is active with gain, then the imager is called an **active pixel image sensor**, or **active pixel sensor**. In a **passive pixel sensor** there is only a switching transistor to read the charge out when the transistor switch is addressed. Nearly all modern CMOS imagers are active pixel sensors. One possible active pixel architecture using a photodiode is shown in Figure 5.36 (a). T₁ is the reset transistor, T₂ works as a source follower (*i.e.*, a buffer), and T₃ is a pixel switch transistor. With T₁ off, the photocurrent charges the self-capacitance of the photodiode to a certain voltage. When the row X receives a signal, T₃ is turned on, and the signal voltage on the photodiode is transferred through the buffer T₂ to the column Y. The pixel is then reset, and T₁ connects the photodiode to V_{DD} to clear the accumulated charge.



Four Megapixel CMOS image sensor.
(Courtesy of Teledyne-DALSA.)

CMOS imagers have a number of distinct advantages. Standard and well-established CMOS fabrication, and hence lower cost, mean on-chip integration such as a **camera on a chip**. A camera on a chip has a microlens at each pixel—a red, green, or blue filter at each pixel for color imaging—as illustrated in Figure 5.36 (b), analog signal processing after the read-out of the image, and analog-to-digital conversion. Active pixel architecture can include pixel amplification as well to improve the overall efficiency. CMOS low power consumption leads to a longer battery life.

C. Charge-Coupled Devices

Charge-coupled devices (CCDs) are commonly used as image sensors in professional and consumer television cameras and camcorders, and as image sensors in digital still cameras. However, the term CCD in general does not imply an image sensor but a chip that is able to store and transfer signals in the form of charge. The CCD chip is an integrated circuit made from crystalline silicon and has a large number of pixels (each is a detector element); for example, a $2.5\text{ cm} \times 2.5\text{ cm}$ CCD chip may have 1024×1024 or 2048×2048 pixels on its surface. The basic pixel structure is a MOS (metal-oxide-semiconductor) or a MIS (metal-insulator-semiconductor) device as shown in Figure 5.37. Notice that the structure is based on a *p*-type Si, an oxide layer, and a metal electrode, which is usually transparent. There is a depletion region inside the *p*-type semiconductor. The EHPs are generated inside the depletion region by illumination either from the top surface or from the backside as shown in Figure 5.37. In back-thinned CCD, light enters the depletion region not from the gate side but from the “substrate” side, which has been thinned to allow the light to pass through.

When a positive voltage $+V$ is applied to the gate V_G in Figure 5.37 ($V_G = +V$), the photo-generated electrons in the depletion region are collected in a layer near the interface as shown in Figure 5.37. (With no gate voltage, photogenerated electrons and holes disappear by recombination.) These electrons are trapped inside a potential energy well, introduced by $+V$ on the gate. Their total charge is proportional to the total light exposure. This charge constitutes the electrical signal. The objective is to read all these charges stored at the illuminated pixels.

In a **three-phase CCD** read-out, there are three line voltages V_1 , V_2 , V_3 to which the gates are connected in an alternating fashion: G_1 to V_1 , G_2 to V_2 , and G_3 to V_3 , G_4 to V_1 again and so on as shown in Figure 5.38. V_1 , V_2 , and V_3 are appropriately clocked to shift the charges from pixel to pixel to a register located at the end of the chip. If initially (time $t = t_1$) $V_1 = +V$ and $V_2 = V_3 = 0$, then photogenerated charges will be stored under G_1 , G_4 , etc. as Q_a , Q_b , etc. Later ($t = t_2$) we can make $V_1 = +V$, $V_2 = +V$, and $V_3 = 0$. The charge Q_a is shared between the wells under G_1 and G_2 . Even later ($t = t_3$), we can bring V_1 down to zero, the charge Q_a must go into the available potential well, which is under G_2 . Thus, by toggling gate voltages, Q_a has been shifted from G_1 to G_2 , and similarly Q_b from G_4 to G_5 and so on. The charges are therefore clocked progressively along the gates, from pixel to pixel, left to right, until they reach the end of the array where there is a register. The CCD read-out therefore functions like a *shift register* in that clock pulses shift the information along the chain; they are often termed *CCD shift registers*.

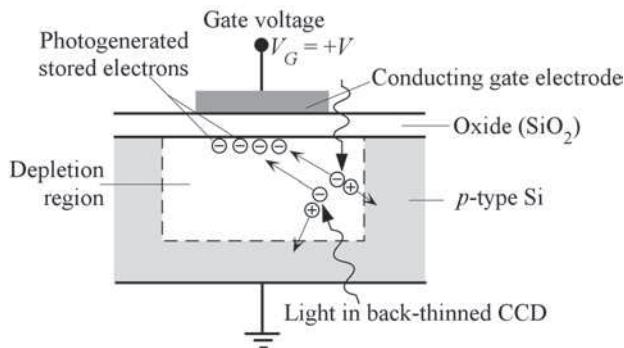


FIGURE 5.37 One element of a CCD imaging sensor, which is a MOS (metal-oxide-semiconductor) device. The absorbed photon actually creates an electron and a hole. The hole simply drifts to the *p*-side.



Various CCD chips. (Courtesy of Teledyne-DALSA.)

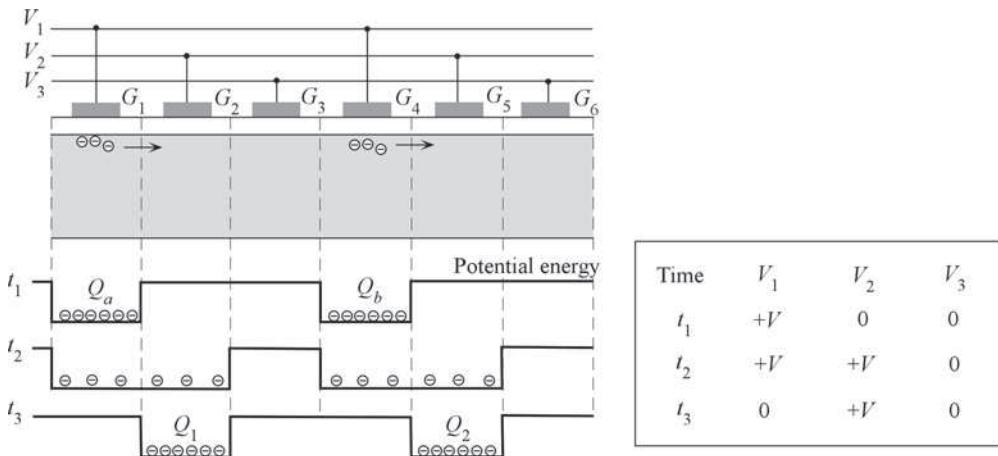


FIGURE 5.38 Transfer of charge from one well to another by clocking the gate voltages. The table shows the gate voltage sequences in a three-phase CCD.

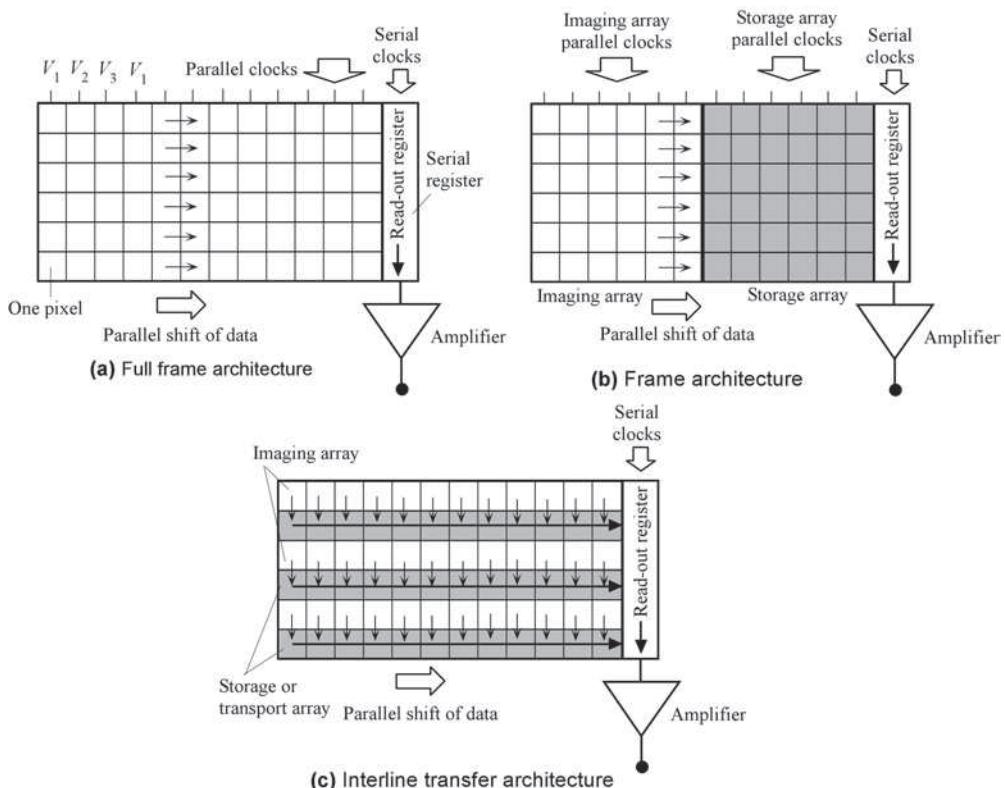


FIGURE 5.39 Three CCD architectures: (a) Full frame architecture. (b) Frame architecture. (c) Interline transfer architecture.

There are three main types of CCD read-out techniques and hence CCD architectures called full frame, frame, and interline transfer architectures. In a **full frame architecture**, the sensor consists of parallel CCD shift registers, a serial CCD shift register as a *read-out register*, and an amplifier as shown in Figure 5.39 (a). The columns of data are fed into the serial register starting soon after the projection of the image onto the array, as the clock signals start transferring the charges pixel to pixel toward the read-out register. A short build-up or exposure time is necessary to build up the charges in the imaging arrays. The process continues until all the columns of data have been transferred. This architecture has the highest resolution and is the densest.

The **frame architecture**, as shown in figure 5.39 (b), is similar to the full frame except that half the array is used as data storage to increase the speed. The storage array is blocked from light. The idea is to shift the data away from the image sensors quickly so that they can capture the next image soon; these have higher frame rates at the expense of resolution.

The **interline transfer architecture** uses every other row of the CCD array for *storage* and *transport* right next to the imaging array as illustrated in figure 5.39 (c). The transport array is blocked from light, and the data from the imaging array are transferred in parallel to the transport array after the image has been built-up in the imaging array. The data on the transport array are shifted, just as in the normal CCD.

Additional Topics

5.14 PHOTOVOLTAIC DEVICES: SOLAR CELLS

A. Basic Principles

A **photovoltaic device** or a **solar cell** converts the incident radiation energy into electrical energy.¹⁹ Incident photons are absorbed to photogenerate charge carriers, which then pass through an external load to do electrical work. Photovoltaic devices may be metal–semiconductor Schottky junctions, *pn* junctions, or *pin* devices. For example, a crystalline Si *pn* junction solar cell may have a thin *n*-type semiconductor layer on a thick *p*-type substrate as shown in Figure 5.40 (a). The electrodes attached to the *n*-side must allow the light to enter the device and at the same time result in a small series resistance. They are deposited on to *n*-side to form an array of **finger electrodes** on the surface as illustrated in Figure 5.40 (a). There is a thin film antireflection (AR) coating (one or two layers) on the surface outside the electrodes. The AR coating significantly reduces the reflectance of the bare semiconductor–air interface.

The *n*-side of the *pn* junction is very narrow to allow most of the photons to be absorbed within the depletion region (*W*) and within the neutral *p*-side. The photogeneration of electron–hole pairs occurs mainly in these regions. In Si, the electron diffusion length L_e is longer than the hole diffusion length L_h , which is the reason for having the *p*-layer as the main photon absorbing layer. EHP photogeneration then occurs in a bigger volume defined by *W* and L_e in the *p*-side as indicated in Figure 5.40 (a). Remember that only those electrons in the *p*-side within L_e to the SCL can contribute the photocurrent. EHPs photogenerated in *W* in Figure 5.40 (a) are

¹⁹ Although the terms *photovoltaic device* and *solar cell* are often used interchangeably, solar cells would convert solar radiation into electrical work, whereas a photovoltaic device could also be used as a photodetector, operating in the photovoltaic mode; their designs would be somewhat different.

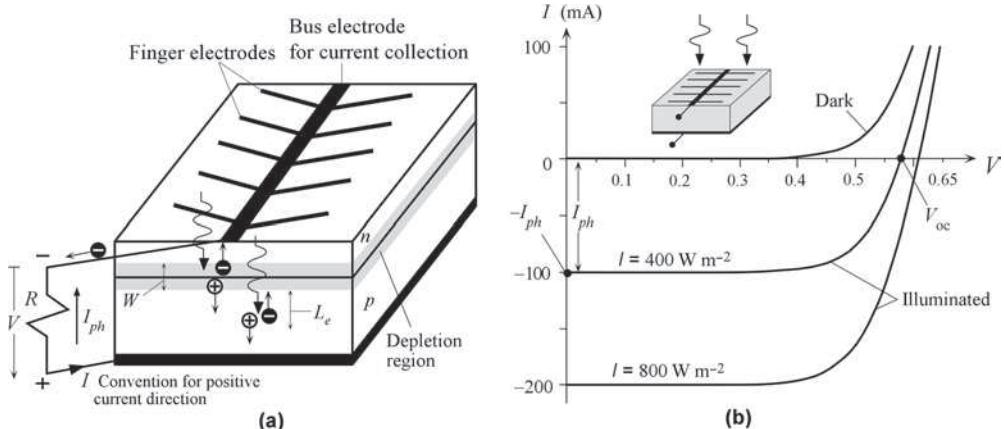


FIGURE 5.40 (a) Typical pn junction solar cell. Finger electrodes on the surface of a solar cell reduce the series resistance. (b) I - V characteristics in the dark and under illumination at intensities corresponding to 400 W m^{-2} and 800 W m^{-2} .

immediately separated by the built-in field, which drifts them apart. The electrons drift toward the n -side and holes toward the p -side, which generates a photocurrent. The excess electrons reach the neutral n -side, then drift around the external circuit, do work, and reach the p -side to recombine with the excess holes on this side. As a result, there will be a continuous external photocurrent during illumination. If the load is a short circuit, then the magnitude of the external current is simply equal to the photocurrent I_{ph} generated by the incident radiation ($I = -I_{ph}$). It is important to realize that without the internal field, it is not possible to drift apart the photogenerated EHPs and hence cause a current flow in the external circuit.

The I - V characteristics of a typical Si solar cell are shown in Figure 5.40 (b). The dark I - V is the usual forward biased pn junction diode equation

pn
junction
diode
current

$$I_{\text{diode}} = I_o [\exp(eV/\eta k_B T) - 1] \quad (5.14.1)$$

where η is the diode ideality factor (not to be confused with the solar cell efficiency) which is between 1 and 2. Under illumination, the I - V dark characteristics are shifted down by an amount that is equal to the photocurrent I_{ph} . The photocurrent I_{ph} is proportional to the photogeneration rate and hence to the incident light intensity I , i.e.,

Photocurrent
and light
intensity

$$I_{ph} = KI \quad (5.14.2)$$

where K is a device specific constant. As can be seen from Figure 5.40 (b), doubling the light intensity has doubled I_{ph} .

The positive I and V convention and the short circuit behavior of an illuminated pn junction are shown in Figure 5.41 (a) and (b). The short circuit current I_{sc} is $-I_{ph}$, and is an important solar cell quantity. If the solar cell is in open circuit, there is no external current ($I = 0$) but there is a voltage, called the **open circuit voltage**, V_{oc} , across the device as in Figure 5.41 (c). This voltage corresponds to the point where the illuminated I - V characteristic cuts the V -axis in Figure 5.40 (b). The photogenerated electrons arriving on the n -side cannot flow in the external circuit and accumulate in the n -side; similarly photogenerated holes accumulate in the p -side. We should clearly see a net positive voltage on the p -side with respect to the n -side. The reality

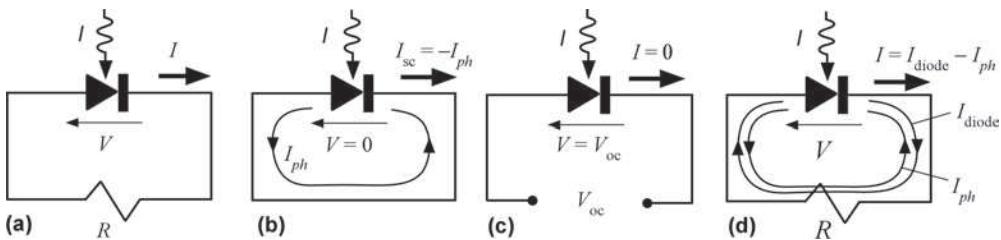


FIGURE 5.41 (a) The solar cell connected to an external load R and the convention for the definitions of positive voltage and positive current. (b) The solar cell in short circuit. The magnitude of the external current I_{sc} in the photocurrent I_{ph} . (c) Under open circuit conditions there is a voltage V_{oc} at the output terminals. (d) The solar cell driving an external load R . There is a voltage V and current I in the circuit.

is a bit more subtle. The extra electrons on the n -side neutralize some of the positive donor ions in the SCL on the n -side. Extra holes neutralize some of negative acceptors in the SCL on the p -side. The internal field becomes decreased, and hence the built-in voltage V_o becomes reduced. It is the reduction in the built-in voltage that appears as V_{oc} at the external terminals.

Suppose that there is an external load R , as shown in Figures 5.40 (a) and 5.41 (d). The external current (no longer simply I_{ph}) will flow through R and generate a voltage V across it. IV is the electric power dissipated in the external load. It should be apparent that the voltage V developed across R now *forward biases* the pn junction, generates a diode current I_{diode} in the normal way, as given by Eq. (5.14.1), and flows in the opposite direction to I_{ph} . The net current, as apparent from Figure 5.41 (d) is

$$I = -I_{ph} + I_{diode} = -I_{ph} + I_o [\exp(eV/\eta k_B T) - 1] \quad (5.14.3) \quad \begin{matrix} \text{Solar cell} \\ I-V \end{matrix}$$

which represents the I - V characteristics of a solar cell. The light intensity increases I_{ph} along Eq. (5.14.2), and hence shifts the normal diode characteristics down as shown in Figure 5.40 (b).



An experimental solar cell aircraft called Helios flying over the coast of Hawaii. (Courtesy of NASA Dryden Research Center.)

B. Operating Current and Voltage and Fill Factor

Equation (5.14.3) gives only the I - V characteristics of the solar cell. When the solar cell is connected to a load as in Figure 5.42 (a), the load has the same voltage as the solar cell and carries the same current. But the current I through R is now in the opposite direction to the convention that current flows from high to low potential. Thus, as shown in Figure 5.42 (a)

$$I = -V/R \quad (5.14.4) \quad \begin{matrix} \text{The load} \\ \text{line} \end{matrix}$$

The actual current I' and voltage V' in the circuit must satisfy both the I - V characteristics of the solar cell, Eq. (5.14.3), and that of the load, Eq. (5.14.4). We can find I' and V' by solving these two equations simultaneously or using a **load line construction**. The I - V characteristics

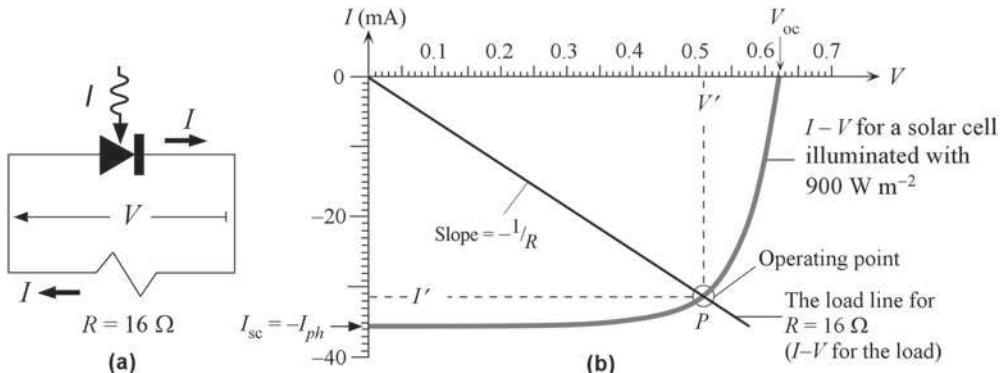


FIGURE 5.42 (a) A solar cell driving a load R and the definitions of positive current I and voltage V . (b) The load line construction for finding the operating point when a load $R_L = 15 \Omega$ is connected across the solar cell.

of the load in Eq. (5.14.4) is a straight line with a negative slope $-1/R$. This is called the **load line** and is shown in Figure 5.42 (b) along with the I - V characteristics of the solar cell under a given intensity of illumination. The load line cuts the solar cell characteristic at P where the load and the solar cell have the same current and voltage, I' and V' . Point P therefore satisfies both Eqs. (5.14.3) and (5.14.4) and thus represents the **operating point of the circuit**.

The **power delivered** to the load is $P_{out} = I'V'$, which is the area of the rectangle bound by I - and V -axes and the dashed lines shown in Figure 5.42 (b). Maximum power is delivered to the load when this rectangular area is maximized (by changing R or the intensity of illumination), when $I' = I_m$ and $V' = V_m$. Since the maximum possible current is I_{ph} and the maximum possible voltage is V_{oc} , $I_{ph}V_{oc}$ represents the desirable goal in power delivery for a given solar cell. It therefore makes sense to compare the maximum power output²⁰, I_mV_m , with $I_{sc}V_{oc}$. The **fill factor** (FF), which is a figure of merit for the solar cell, is defined as

Definition
of fill
factor

$$\text{FF} = \frac{I_mV_m}{I_{sc}V_{oc}} \quad (5.14.5)$$

The FF is a measure of the closeness of the solar cell I - V curve to the rectangular shape (the ideal shape). It is clearly advantageous to have FF as close to unity as possible, but the exponential pn junction properties prevent this. Typically FF values are in the range 70–85% and depend on the device structure. One can vary the operating point P in Figure 5.42 (b) to examine the effects of different load lines on the rectangular area $I'V'$. A load line that cuts the solar cell I - V characteristics around the “knee” point, as in Figure 5.42 (b), corresponds to nearly the maximum power I_m, V_m point.

C. Equivalent Circuit of a Solar Cell

Practical solar cells can deviate substantially from the ideal pn junction solar cell behavior illustrated in Figure 5.40 (b) due to a number of reasons. Consider an illuminated solar cell driving a load resistance R_L and assume that photogeneration takes place in the depletion region. The photogenerated electrons have to traverse a surface semiconductor region to reach the nearest finger electrode as shown in Figure 5.43 (a). All these electron paths in the n -layer surface region of finger electrodes introduce an **effective series resistance** R_s into the photovoltaic circuit. If the finger electrodes are thin, then the resistance of the electrodes themselves will further

²⁰One should not be too concerned with the negative sign for I_mV_m inasmuch as the diode is generating electric power during illumination.

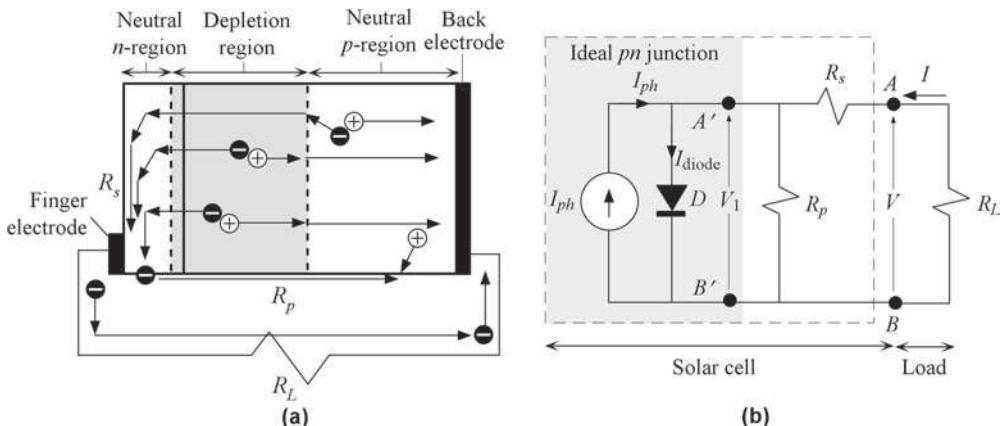


FIGURE 5.43 (a) Series and shunt resistances and various fates of photogenerated EHPs. (b) A simple equivalent circuit diagram for a solar cell. The current I into A , and V across A and B , follow the convention for positive current and voltage. A' and B' represent the ideal diode terminals.

increase R_s . There is also a series resistance due to the neutral p -region but this is generally small compared with the resistance of the electron paths to the finger electrodes.

Figure 5.43 (b) shows the equivalent circuit of a pn junction solar cell. The gray region is the ideal pn junction diode case and has the terminals A' and B' . The photogeneration process is represented by a constant current generator I_{ph} , which generates a current that is proportional to the light intensity. The flow of photogenerated carriers across the junction gives rise to a photovoltaic voltage difference V_1 across the junction, and this voltage leads to the normal diode current $I_{diode} = I_o[\exp(eV_1/\eta kT) - 1]$. This diode current I_{diode} is represented by an ideal pn junction diode D in the circuit as shown in Figure 5.43 (b). As apparent, I_{ph} and I_{diode} are in opposite directions (I_{ph} is “up” and I_{diode} is “down”) so that, in open circuit, the photovoltaic voltage V_1 is such that I_{ph} and I_{diode} have the same magnitude and cancel each other. By convention, positive current I at the output terminal is normally taken to flow into the terminal, and is given by Eq. (5.14.3) (In reality, of course, the solar cell current is negative, which represents a current that is flowing out into the load.) The voltage V between A and B is the actual voltage measured, and available externally across the solar cell terminals. Notice that $V_1 = V - IR_s$ so that we can substitute for V_1 in terms of V in the diode current equation.

The equivalent circuit in Figure 5.43 (b) outside the ideal gray box contains the series resistance R_s in Figure 5.43 (a), which gives rise to a voltage drop, and therefore prevents the ideal photovoltaic voltage from developing at the output between A and B when a current is drawn. A small fraction of the photogenerated carriers can also flow through the crystal surfaces (edges of the device) or through *grain boundaries in polycrystalline devices* instead of flowing through the external load R_L as shown in Figure 5.43 (a). These effects that prevent photogenerated carriers from flowing in the external circuit can be represented by an effective **shunt** or **parallel resistance** R_p that diverts the photocurrent away from the load R_L . Typically R_p is less important than R_s in overall device behavior, unless the device is highly polycrystalline and the current component flowing through grain boundaries is not negligible. The current I flowing into A , and the voltage V across A and B (of A with respect to B) in Figure 5.43 (b) now represent the solar cell characteristics that the user measures.

The series resistance R_s can significantly deteriorate the solar cell performance as illustrated in Figure 5.44, where $R_s = 0$ is the best solar cell case. It is apparent that the available maximum output power decreases with increasing series resistance, which therefore reduces the

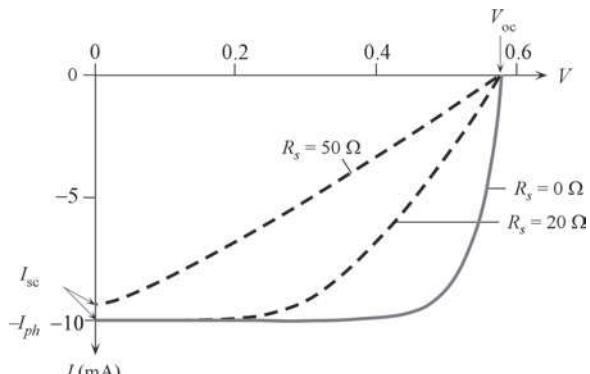


FIGURE 5.44 The effect of series resistance R_s on the I - V characteristics. This example is a Si pn junction solar cell with $\eta = 1.5$ and $I_o = 3 \times 10^{-6}$ mA. The light intensity is such that it generates $I_{ph} = 10$ mA.

cell efficiency. Notice also that, as expected, V_{oc} is unaffected by R_s , though for sufficiently large R_s , the short circuit current can become reduced. Similarly, low shunt resistance values, due to extensive defects in the material, also reduce the efficiency. The difference is that although R_s does not affect the open circuit voltage V_{oc} , low R_p leads to a reduced V_{oc} . (See Question 5.35.)

D. Solar Cell Structures and Efficiencies

Most solar cells use crystalline silicon because silicon-based semiconductor fabrication is now a mature technology that enables cost-effective devices to be manufactured. One of the most important solar cell metrics is the **efficiency**, which is defined as the maximum output power from the device per unit incident radiation power under well-defined conditions. The Sun's spectrum is different at different heights from Earth due to absorption and scattering effects in the atmosphere, which depend on the wavelength. For terrestrial use, one common standard condition for solar cell comparisons is the so-called global **AM1.5 solar spectrum**²¹ that represents the radiation arriving on Earth's surface with an integrated intensity (over all wavelengths) of 1000 W m^{-2} ; and the test temperature is 25°C . Table 5.6 lists the efficiencies and various solar cell parameters such as V_{oc} , J_{sc} (short circuit current density), and the fill factor.

TABLE 5.6 Characteristics of a few selected classes of solar cells, and reported efficiencies under global AM1.5 solar spectrum (1000 W m^{-2}) at 25°C

Semiconductor	V_{oc} (V)	$ J_{sc} $ (mA cm^{-2})	FF (%)	Efficiency (%)	E_g (eV)
Si, single crystal (PERL)	0.707	42.7	82.8	25.0	1.11
Si, polycrystalline	0.664	38.0	80.9	20.4	1.11
Amorphous Si:H (<i>pin</i>)	0.886	16.75	67.0	10.1	~1.7
GaAs, single crystal	1.030	29.8	86.0	26.4	1.42
GaAs, thin film	1.107	29.6	84.1	27.6	1.42
InP, single crystal	0.878	29.5	85.4	22.1	1.35
GaInP/GaAs Tandem	2.488	14.22	85.6	30.3	1.95/1.42
GaInP/GaAs/Ge Tandem	2.622	14.37	85.0	32.0	1.95/1.42/0.66

(Source: Data have been extracted from M. A. Green, K. Emery, Y. Hishikawa, W. Warta, *Progress in Photovoltaics: Research and Applications*, 18, 346, 2010, and 19, 84, 2011. The original tables in the latter have extensive confirmed data on a number of important solar cells and modules with references. In addition, the original tables also list the uncertainties and errors involved in the reported efficiency values.)

²¹ AM1.5 means Air-Mass1.5. In general, the number m in AM m represents $m = \sec \theta$, where θ is the Zenith angle between the shortest path the sun rays can take and the actual path to the solar cell.

Typical Si-based solar cell efficiencies range from about 18% for polycrystalline to 22–25% in high efficiency single crystal devices that have special structures to absorb as many of the incident photons as possible. The reflection of light from the solar cell surface must be minimized to increase the device efficiency. Indeed, best Si *pn* junction solar cell efficiencies are about 24–25% for expensive single crystal PERL—Passivated Emitter Rear Locally diffused—cells.²² The PERL and similar cells have a textured surface that is an array of “inverted pyramids” etched into the surface to capture as much of the incoming light as possible as illustrated in Figure 5.45. Normal reflections from a flat crystal surface lead to a loss of light, whereas reflections inside the pyramid allow a second or even a third chance for absorption. Further, after refraction photons would be entering the semiconductor at oblique angles, which means that they will be absorbed in the useful photogeneration volume, that is, within the electron diffusion length of the depletion layer as shown in Figure 5.45. The largest factors reducing the efficiency of a Si solar cell are the unabsoed photons with $h\nu < E_g$ (long wavelengths) and short wavelength photons that are absorbed near the surface. Both of these factors are improved by the use of multifunction heterostructure.

Research over the years has led to the development of high efficiency heterostructure **tandem** solar cells, which are also called **multiplication solar cells**. Multijunction solar cells essentially use two or more cells in tandem, or in cascade, to increase the absorbed photons from the incident light as illustrated in Figure 5.46 (a). The first cell is made from a wider bandgap material and only absorbs photons with $h\nu > E_{g1}$. The second cell absorbs photons that pass the first cell and have $h\nu > E_{g2}$. The whole structure can be grown by using lattice matched crystalline layers on a suitable substrate, leading to a monolithic tandem cell. The two cells have to be connected, that is, allow the carriers (electrons and holes) to pass. This is done by using a highly doped very thin *pn* junction between the two cells that serves as a tunneling junction; the carriers tunnel through it. All the layers are grown by special techniques on a single substrate. One of the best efficiencies is achieved by using a three junction solar cell, which is illustrated in Figure 5.46 (b). The layers are all grown on a Ge substrate and each cell is an *np* junction. There are two very thin *pn* tunnel junctions that connect the cells in tandem as shown in Figure 5.46 (b), to allow the drifting carriers tunnel through (pass through). The top cell is GaInP₂ with $E_g \approx 1.95$ eV ($\lambda_g = 0.64$ μm), the second is GaAs with $E_g \approx 1.42$ eV ($\lambda_g = 0.87$ μm), and the third is Ge with $E_g \approx 0.66$ eV ($\lambda_g = 1.9$ μm). The three cells have a wide spectral range and are able to capture a very high percentage of the solar radiation. The multijunction solar cell in

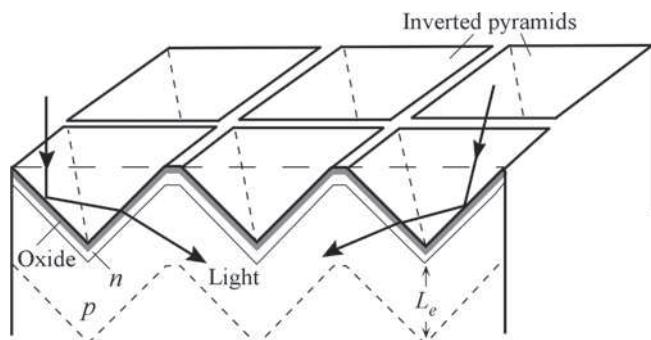


FIGURE 5.45 Inverted pyramid textured surface substantially reduces reflection losses and increases absorption probability in the device.

²²Much of the pioneering work for high efficiency PERL solar cells was done by Martin Green and coworkers at the University of New South Wales, Australia, (J. Zhao, A. Wang, and M. A. Green, *Appl. Phys. Letts.*, 73, 1991, 1998.)

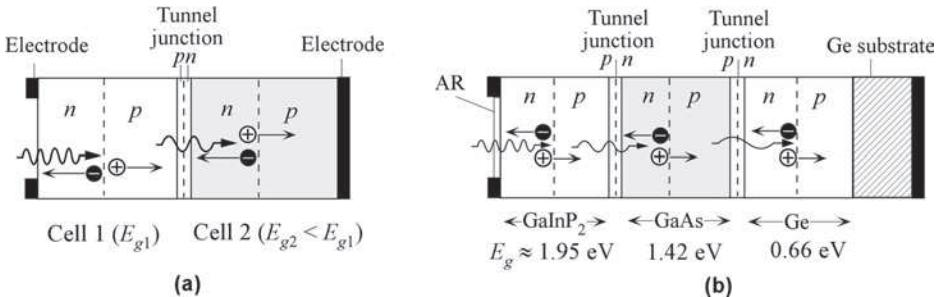


FIGURE 5.46 (a) A heterojunction solar cell that can absorb both high and low energy photons and generate a photocurrent. (b) A tandem solar cell.

Figure 5.46 (b) is commercially available (Spectrolab) with an efficiency of 29.5%. Even higher efficiencies have been reported in research labs using such multijunction heterostructures. If, in addition, light concentrators are also used, the efficiency can be further increased.

Tandem cells are also used in thin film a-Si:H (hydrogenated amorphous Si) *pinn* solar cells to obtain efficiencies up to about 12%. These tandem cells are based on a-Si:H and a-SiGe:H and are easily and inexpensively fabricated in large areas. Amorphous Si:H has an E_g of about 1.8 eV. The alloying of a-Si:H with Ge to produce a-SiGe:H decreases E_g . Further, E_g of a-SiGe:H can be graded by controlling the Ge content.

EXAMPLE 5.14.1 Solar cell driving a load

Consider the solar cell driving a $16\text{-}\Omega$ resistive load as in Figure 5.42 (b). Suppose that the cell has an area of $1\text{ cm} \times 1\text{ cm}$ and is illuminated with light of intensity 900 W m^{-2} as in the figure. What are the current and voltage in the circuit? What is the power delivered to the load? What is the efficiency of the solar cell in this circuit? If you assume it is operating close to the maximum deliverable power, what is the FF?

Solution

The I - V characteristic of the load is the load line as described by Eq. (5.14.4), $I = -V/R$ with $R = 16\text{ }\Omega$. This line is drawn in Figure 5.42 (b) with a slope $1/(16\text{ }\Omega)$. It cuts the I - V characteristics of the solar cell at $I' \approx -31.5\text{ mA}$ and $V' \approx 0.505\text{ V}$, which are the current and voltage in the photovoltaic circuit of Figure 5.42 (b). In fact, from Eq. (5.14.4), V'/I' gives $-16\text{ }\Omega$ as expected. The power delivered to the load is

$$P_{\text{out}} = |I'V'| = (31.5 \times 10^{-3}\text{ A})(0.505\text{ V}) = 0.0159\text{ W} \quad \text{or} \quad 15.9\text{ mW}$$

This is *not* necessarily the maximum power available from the solar cell. The input sunlight power is

$$P_{\text{in}} = (\text{Light intensity})(\text{Surface area}) = (900\text{ W m}^{-2})(0.01\text{ m})^2 = 0.090\text{ W}$$

The efficiency is

$$\text{Efficiency} = 100 \times (P_{\text{out}}/P_{\text{in}}) = 100(15.9\text{ mW}/90\text{ mW}) = 17.7\%$$

This will increase if the load is adjusted to extract the maximum power from the solar cell but the increase will be small as the rectangular area $I'V'$ in Figure 5.42 in it is already close to the maximum. Assuming that $|I'V'|$ is roughly the maximum power available (maximum area for the rectangle $I'V'$), then $I_m \approx I' \approx -31.5\text{ mA}$ and $V_m \approx V' \approx 0.505\text{ V}$. For the solar cell in Figure 5.42 (b), $I_{\text{sc}} = -35.5\text{ mA}$ and $V_{\text{oc}} = 0.62\text{ V}$. Then

$$\text{FF} = I_m V_m / I_{\text{sc}} V_{\text{oc}} \approx (-31.5\text{ mA})(0.505\text{ V}) / (-35.5\text{ mA})(0.62\text{ V}) = 0.72 \quad \text{or} \quad 72\%$$

EXAMPLE 5.14.2 Open circuit voltage and short circuit current

A solar cell under an illumination of 500 W m^{-2} has a short circuit current I_{sc} of -16 mA and an open circuit output voltage V_{oc} of 0.50 V . What are the short circuit current and open circuit voltages when the light intensity is doubled? Assume $\eta = 1$.

Solution

The general I - V characteristics under illumination is given by Eq. (5.14.3). The short circuit current corresponds to the photocurrent so that, from Eq. (5.14.2), at double the intensity, the photocurrent is

$$I_{ph2} = \left(\frac{I_2}{I_1} \right) I_{ph1} = (16 \text{ mA})(1000/500) = 32 \text{ mA}$$

Setting $I = 0$ for open circuit we can obtain the open circuit voltage V_{oc} ,

$$I = -I_{ph} + I_o [\exp(eV_{oc}/\eta k_B T) - 1] = 0$$

Assuming that $V_{oc} \gg \eta k_B T/e$, rearranging the above equation we can find V_{oc}

$$V_{oc} = \frac{\eta k_B T}{e} \ln \left(\frac{I_{ph}}{I_o} \right) \quad (5.14.6)$$

Open
circuit
output
voltage

In Eq. (5.14.6), the photocurrent, I_{ph} , depends on the light intensity I via $I_{ph} = KI$. At a given temperature, then the change in V_{oc} is

$$V_{oc2} - V_{oc1} = \frac{\eta k_B T}{e} \ln \left(\frac{I_{ph2}}{I_{ph1}} \right) = \frac{\eta k_B T}{e} \ln \left(\frac{I_2}{I_1} \right)$$

Assuming $\eta = 1$, the new open circuit voltage is

$$V_{oc2} = V_{oc1} + \frac{\eta k_B T}{e} \ln \left(\frac{I_2}{I_1} \right) = 0.50 \text{ V} + (1)(0.0259 \text{ V}) \ln(2) \approx 0.52 \text{ V}$$

This is a $\sim 4\%$ increase in V_{oc} compared with the 100% increase in illumination and the short circuit current.

Questions and Problems

5.1 Bandgap and photodetection

- (a) Determine the maximum value of the energy gap which a semiconductor, used as a photoconductor, can have if it is to be sensitive to light of wavelength 1550 nm commonly used in optical communication network.
- (b) A photodetector whose area is $4 \times 10^{-2} \text{ cm}^2$ is irradiated with light of wavelength 1300 nm , whose intensity is 3 mW cm^{-2} . Assuming that each photon generates one electron–hole pair, calculate the number of pairs generated per second.
- (c) From the known energy gap of the semiconductor InGaAsP ($E_g = 1.24 \text{ eV}$), calculate the primary wavelength of photons emitted from this crystal as a result of electron–hole recombination. Is this wavelength in the visible range?
- (d) Will a silicon photodetector be sensitive to the radiation from an InGaAsP laser? Why?

- 5.2 Absorption coefficient** In the following questions assume that all the incident radiation onto the semiconductor surface is transmitted into the semiconductor medium; that is, there is a perfect antirelection coating on the surface of the semiconductor.

- (a) If d is the thickness of a photodetector material, I_o is the intensity of the incoming radiation, show that the number of photons absorbed per unit volume of sample is

$$n_{ph} = \frac{I_o [1 - \exp(-\alpha d)]}{d h \nu}$$

- (b) What is the thickness of a Ge and $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ crystal layer that is needed for absorbing 90% of the incident radiation at $1.5 \mu\text{m}$? (Use Figure 5.5)
- (c) Suppose that each absorbed photon liberates one electron (or electron hole pair) in a unity quantum efficiency photodetector and that the photogenerated electrons are immediately collected. Thus, the rate of charge collection is limited by rate of photon generation. What is the external photocurrent density for the photodetectors in (b) if the incident radiation is $100 \mu\text{W mm}^{-2}$?

5.3 Absorption coefficient and device design An optoelectronic device engineer has designed a pn junction photodiode that has a depletion width of $2 \mu\text{m}$. The semiconductor on the light-receiving side is heavily doped (*e.g.*, p^+), the neutral region is very narrow, and can be neglected. If the photogeneration is to occur in the depletion region, what are approximate corresponding wavelengths for the semiconductors in Figure 5.5?

5.4 Absorption coefficient and quantum efficiency Consider the Si pn junction photodiode shown in Figure 5.8. Suppose that the incident optical power is P_o . The incident number of photons per unit time is P_o/hv . Of these, a fraction T is transmitted. The number of photons flowing per second along the device is

$$\frac{dN_{\text{ph}}}{dt} = \frac{TP_o}{hv} \exp(-\alpha x)$$

where x is measured from the illuminated surface of the semiconductor. Show that the photocurrent generated by the photodiode is given by

$$I_{\text{ph}} = \frac{e\eta_i TP_o}{hv} \left\{ \exp[-\alpha(\ell_p - L_e)] - \exp[-\alpha(\ell_p - L_e + W + L_h)] \right\}$$

where η_i is the internal quantum efficiency and the other symbols are defined in Figure 5.8. Consider two photodiodes, one p^+n and the other pin . Their properties are listed in Table 5.7, where the photocurrent has been calculated at 800 nm. Suppose that the incident optical power is $1 \mu\text{W}$, and the antireflection coating on the semiconductor provides $T \approx 1$. Find the photocurrent I_{ph} and hence the responsivity of the two photodiodes at 900 nm and 500 nm. What is your conclusion?

TABLE 5.7 Properties of a p^+n and pin photodiode

	$\ell_p (\mu\text{m})$	$L_e (\mu\text{m})$	$W (\mu\text{m})$	$L_h (\mu\text{m})$	$\alpha (\text{m}^{-1})$ (at 800 nm)	$I_{\text{ph}} (\mu\text{A})$
p^+n	0.5	0.1	1	10	1×10^5	0.45
pin	0.5	0.1	30	0.1	1×10^5	0.59

Note: $\eta_i = 1$, $T = 1$ and $P_o = 1 \mu\text{W}$ were used.

5.5 Ge Photodiode Consider a commercial Ge pn junction photodiode which has the responsivity shown in Figure 5.47. Its photosensitive area is 0.008 mm^2 . It is used under a reverse bias of 10 V when the dark current is $0.3 \mu\text{A}$ and the junction capacitance is 4 pF . The rise time of the photodiode is 0.5 ns .

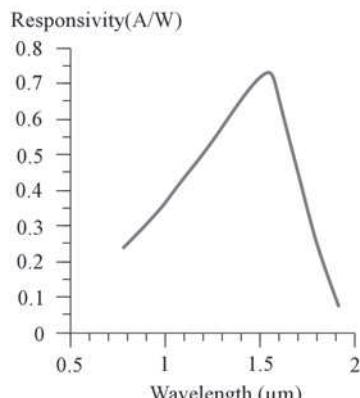


FIGURE 5.47 The responsivity of a commercial Ge pn junction photodiode.

- (a) Calculate its quantum efficiency at 850 nm, 1300 nm, and 1550 nm.
- (b) What is the light intensity of light at $1.55 \mu\text{m}$ that gives a photocurrent equal to the dark current?
- (c) What would be the effect of lowering the temperature on the responsivity curve?
- (d) Given that the dark current is in the range of microamperes, what would be the advantage in reducing the temperature?
- (e) Suppose that the photodiode is used with a $100\text{-}\Omega$ resistance to sample the photocurrent. What limits the speed of response?

5.6 Si pin Photodiodes Consider two commercial Si *pin* photodiodes, type A and type B, both classified as fast *pin* photodiodes. They have the responsivity curves shown in Figure 5.48. Differences in the responsivity are due to the *pin* device structure. The photosensitive area is 0.125 mm^2 (0.4 mm in diameter).

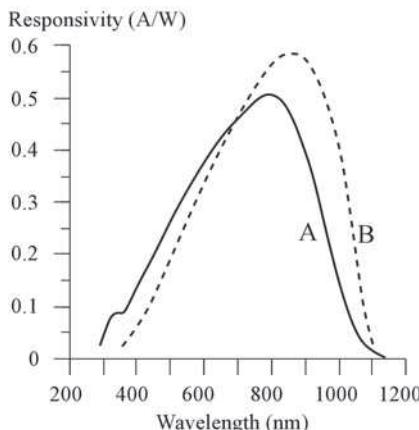


FIGURE 5.48 The responsivity of two commercial Si *pin* photodiodes.

- (a) Calculate the photocurrent from each when they are illuminated with blue light of wavelength 450 nm and light intensity $1 \mu\text{W cm}^{-2}$. What is the QE of each device?
- (b) Calculate the photocurrent from each when they are illuminated with red light of wavelength 700 nm and light intensity $1 \mu\text{W cm}^{-2}$. What is the QE of each device?
- (c) Calculate the photocurrent from each when they are illuminated with infrared light of wavelength 1000 nm and light intensity $1 \mu\text{W cm}^{-2}$. What is the QE of each device?
- (d) What is your conclusion?

5.7 InGaAs pin Photodiodes Consider a commercial InGaAs *pin* photodiode whose responsivity is shown in Figure 5.49. Its dark current is 5 nA.

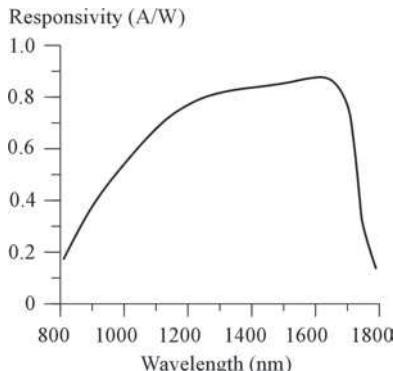


FIGURE 5.49 The responsivity of an InGaAs *pin* photodiode.

- (a) What optical power at a wavelength of $1.55 \mu\text{m}$ would give a photocurrent that is twice the dark current? What is the QE of the photodetector at $1.55 \mu\text{m}$?
- (b) What would be the photocurrent if the incident power in (a) was at $1.3 \mu\text{m}$? What is the QE at $1.3 \mu\text{m}$ operation?

5.8 InAs Photodiode Figure 5.50 shows the responsivity of an InAs *pn* junction photodiode for use in the infrared. It is usually cooled under normal operation.

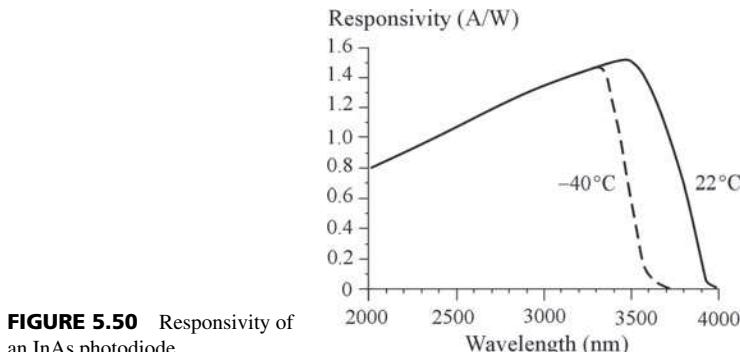


FIGURE 5.50 Responsivity of an InAs photodiode.

- (a) What is the wavelength and the quantum efficiency at the peak responsivity at -40°C ?
- (b) The diode area that is involved in the detection has a diameter of 0.25 mm. What is the intensity of light that corresponds to a photocurrent of 50 nA at peak responsivity?
- (c) Why does the cutoff wavelength become shorter at lower temperatures?

5.9 Photocurrent *pin* photodiode Consider a Si *pin* photodiode with a depletion layer width of $30 \mu\text{m}$ and diameter of $300 \mu\text{m}$. The absorption coefficient at 800 nm is 10^3 cm^{-1} .

Find the quantum efficiency of this photodiode, assuming perfect AR coating. If the load resistance is $10 \text{ k}\Omega$, obtain the RC time constant. What would happen if there is no AR coating so that $T = 0.70$?

5.10 Shockley–Ramo theorem Consider a *pin* photodiode with an *i*-layer width W . A very short pulse of light is absorbed just inside the depletion region on the p^+ -side (Figure 5.9). The photogenerated electrons drift through the *i*-layer with a velocity given in Figure 5.10. What is the current generated by this drift? How long does it last? Suppose that W is $30 \mu\text{m}$, and assume that the quantum efficiency is 0.80. A voltage of 20 V is applied to reverse bias the *pin* detector. The light pulse from a femtosecond laser operating at 515 nm is used for photoexcitation, and the light pulse energy is 37.5 fJ . Assume all of this energy is absorbed very close to the depletion region and p^+ -layer boundary, *i.e.* electrons drift across W and constitute the photocurrent. Calculate the transient photocurrent. How long does it last?

5.11 QE and Responsivity A photodiode has a quantum efficiency of 70% when photons of energy $1.5 \times 10^{-19} \text{ J}$ are incident upon it. What is the operating wavelength of this photodiode? Find out the incident optical power required to obtain a photocurrent of $2.5 \mu\text{A}$ with the above operating wavelength. What would be the responsivity when the QE is 100%?

5.12 Si *pin* photodiode speed Consider Si *pin* photodiodes with a p^+ -layer of thickness $1.0 \mu\text{m}$, and an *i*-Si layer of width $7.5 \mu\text{m}$. It is reverse biased with a voltage of 25 V.

- (a) What is the speed of response due to bulk absorption? What wavelengths would lead to this type of speed of response?
- (b) What is the speed of response due to absorption near the surface? What wavelengths would lead to this type of speed of response?

5.13 Transient photocurrents in a *pin* photodiode Consider a reverse biased Si *pin* photodiode as shown in Figure 5.51. It is appropriately reverse biased so that the field in the depletion region (*i*-Si layer), $E = V_r/W$, is the saturation field. Thus, photogenerated electrons and holes in this layer drift at saturation velocities v_{de} and v_{dh} . Assume that the field E is uniform and that the thickness of the p^+ is negligible. A very short light pulse (infinitesimally short) photogenerates EHPs in the depletion layer as shown in Figure 5.51, which results in an exponentially decaying EHP concentrations across W . Figure 5.51 shows the photogenerated electron concentration at time $t = 0$ and also at a later time t , when the electrons have drifted a distance $\Delta x = v_{de}t$. Those that reach the back electrode B become collected. The electron distribution shifts at a constant velocity until the

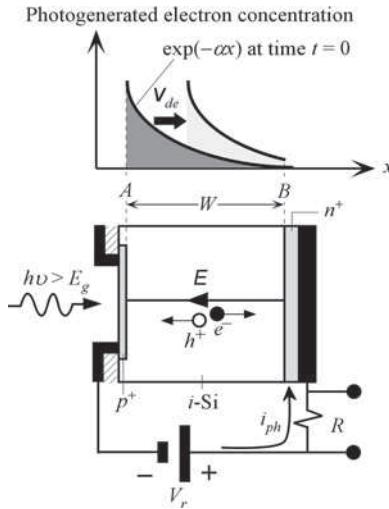


FIGURE 5.51 An infinitesimally short light pulse is absorbed throughout the depletion layer of a reverse-biased ideal *pin* photodetector, and creates an EHP concentration that decays exponentially. The p^+ - and n^+ -layers are assumed to be very thin.

initial electrons at A reach B , which represents the longest transit time $\tau_e = W/v_{de}$. Similar arguments apply to holes, but they drift in the opposite direction and their transit time $\tau_h = W/v_{dh}$ where v_{dh} is their saturation velocity. The photocurrent density at any instant is

$$j_{ph} = j_e(t) + j_h(t) = eN_e v_{de} + eN_h v_{dh}$$

where N_e and N_h are the overall electron and hole concentration in the sample at time t . Assume for convenience that the cross-sectional area $A = 1$ below (derivations are not affected as we are interested in the photocurrent current densities).

- (a) Sketch the hole distribution at time $\tau_h > t > 0$ where $\tau_h = \text{hole drift time} = W/v_{dh}$.
- (b) The electron concentration distribution $n(x)$ at time t corresponds to that at $t = 0$ shifted by $v_{det}t$. Thus the total electrons in W is proportional to integrating this distribution $n(x)$ from A at $x = v_{de}t$ to B at $x = W$. Given $n(x) = n_o \exp(-\alpha x)$ at $t = 0$, where n_o is the electron concentration at $x = 0$ at $t = 0$ we have

$$\text{Total number electrons at time } t = \int_{v_{de}t}^W n_o \exp[-\alpha(x - v_{de}t)] dx$$

and

$$N_e(t) = \frac{\text{Total number electrons at time } t}{\text{Volume}}$$

Then

$$\begin{aligned} N_e(t) &= \frac{1}{W} \int_{v_{de}t}^W n_o \exp[-\alpha(x - v_{de}t)] dx \\ &= \frac{n_o}{W\alpha} \left\{ 1 - \exp \left[-\alpha W \left(1 - \frac{t}{\tau_e} \right) \right] \right\} \end{aligned}$$

$N_e(0)$ is the initial overall electron concentration at time $t = 0$, that is,

$$N_e(0) = \frac{1}{W} \int_0^W n_o \exp(-\alpha x) dx = \frac{n_o}{W\alpha} [1 - \exp(-\alpha W)]$$

We note that n_o depends on the intensity I of the light pulse so that $n_o \propto I$. Show that for holes

$$N_h(t) = \frac{n_o \exp(-\alpha W)}{W\alpha} \left\{ \exp \left[\alpha W \left(1 - \frac{t}{\tau_h} \right) \right] - 1 \right\}$$

- (c) Given $W = 40 \mu\text{m}$, $\alpha = 5 \times 10^4 \text{ m}^{-1}$, $v_{de} = 10^5 \text{ m s}^{-1}$, $v_{dh} = 0.8 \times 10^5 \text{ m s}^{-1}$, $n_o = 10^{13} \text{ cm}^{-3}$, calculate the electron and hole transit time, sketch the photocurrent densities $j_e(t)$ and $j_h(t)$ and hence $j_{ph}(t)$ as a function of time, and calculate the initial photocurrent. What is your conclusion?

- 5.14 Fiber attenuation and InGaAs pin Photodiode** Consider the commercial InGaAs *pin* photodiode whose responsivity is shown in Figure 5.49. This is used in a receiver circuit that needs a minimum of 6 nA photocurrent for a discernible output signal (acceptable signal to noise ratio for the customer). Suppose that the InGaAs *pin* PD is used at 1.3- μm operation with a single mode fiber whose attenuation is 0.4 dB km^{-1} . If the laser diode emitter can launch at most 3 mW of power into the fiber, what is the maximum distance for the communication without a repeater?
- 5.15 Multiplication in APDs** Consider the results of multiplication vs. reverse bias experiments on a commercial InGaAs APD summarized in Table 5.8. Plot $(1 - 1/M)$ vs. V_r and hence find m and V_{br} in Eq. (5.6.2). (This can be done, for example, by using Excel and then fitting a power law trend line; or plotting on a log-log axis in which case the slope would give m .)

TABLE 5.8 Multiplication (M) vs. reverse voltage (V_r) data for an InGaAs commercial APD

V_r	30.6	33.5	36.6	39.5	41.9	43.4	44.3	45.46	45.92	46.21
M	1.78	2.16	2.74	3.74	5.34	7.21	9.0	14.2	19.5	24.8

- 5.16 Effect of temperature on APDs** Experiments on two commercial Si APDs *A* and *B* have examined how the temperature affects the reverse bias V_r needed to keep the multiplication constant. The results are summarized in Table 5.9 in terms of V_r values at different temperatures that would keep $M = 100$ and 200 for *A* and 100 for *B*. Plot V_r vs. T (in K) for *A* and *B*. Do you agree that empirically $V_r \propto T^n$? What is n ?

TABLE 5.9 The reverse bias voltage V_r for given multiplication M at different temperatures for two APDs *A* and *B*

T	-20°C	0°C	20°C	40°C	60°C	Comment
V_r at $M = 100$	125	162.5	200	236	273	APD <i>A</i>
V_r at $M = 100$	135	173.5	211	249	286	APD <i>A</i>
V_r at $M = 100$	147	150	153	155	157	APD <i>B</i>

- 5.17 Multiplication in APDs** The avalanche multiplication gain M is given by

$$M = \frac{1 - k}{\exp[-(1 - k)\alpha_e w] - k}$$

Consider two cases (a) $k = 0$ and (b) $k = 0.1$. Suppose that w is $2 \mu\text{m}$. Increasing the field will increase α_e . Plot M vs. α_e semilogarithmically for the two cases over the range $\alpha_e = 0$ to 2. What is your conclusion? (Note: As α_e increases with the field, k does not strictly remain the same. However, the two cases clearly distinguish two distinctly different types of behavior.)

- 5.18 Si APD** The electron and hole ionization coefficients α_e and α_h in silicon are approximately given in Example 5.6.3. Suppose we would like the avalanche to be achieved at a certain applied field in the avalanche region (corresponding to a particular desirable voltage range). Suppose that we would like the multiplication M to be 100 when $E = 5 \times 10^5 \text{ V cm}^{-1}$. Estimate the width w of the avalanche region.

- 5.19 InP APD design** For InP the impact ionization coefficients are roughly given by $\alpha_e \approx (9.2 \times 10^6) \exp(-3.44 \times 10^6/E)$ and $\alpha_h \approx (4.3 \times 10^6) \exp(-2.72 \times 10^6/E)$, where α_e and α_h are in cm^{-1} , and E is in V cm^{-1} (from K. Takuchi *et al.*, *J. Appl. Phys.*, 59, 476, 1986). Consider the heterojunction InGaAs-InP APD in Figure 5.16. The avalanche is initiated by holes in N -InP. We can define $k = \alpha_e/\alpha_h$ and then the multiplication $M = (1 - k)/\{\exp[-(1 - k)\alpha_h w] - k\}$. Suppose that the width of the N -layer is $1 \mu\text{m}$ and the applied field is $4.6 \times 10^5 \text{ V cm}^{-1}$. What is the multiplication M ? If you increase the field by 2%, what is M ?

5.20 Schottky junction photodiode Consider a *pn* junction and a Schottky junction photodiode. Calculate the photocurrent under an illumination of $2 \mu\text{W}$ at a wavelength of 400 nm for each, and find the responsivity given (a) the *pn* junction has $\ell_p = 0.5 \mu\text{m}$, $L_e = 0.1 \mu\text{m}$, $W = 2 \mu\text{m}$, $L_h = 10 \mu\text{m}$; and (b) the Schottky junction has $W = 3 \mu\text{m}$. Assume that the internal QE is 100%. (Note: You can use the equation in Question 5.3.)

5.21 Photoconductive detector An *n*-type Si photoconductor has a length $L = 100 \mu\text{m}$ and a cross-sectional area $A = 10^{-4} \text{ mm}^2$. The applied bias voltage to the photoconductor is 10 V .

- (a) What are the transit times, t_e and t_h , of an electron and a hole across L ? What is the photoconductive gain?
- (b) It should be apparent that as electrons are much faster than holes, a photogenerated electron leaves the photoconductor very quickly. This leaves behind a drifting hole and therefore a positive charge in the semiconductor. Secondary (*i.e.*, additional electrons) then flow into the photoconductor to maintain neutrality in the sample and the current continues to flow. These events will continue until the hole has disappeared by recombination, which takes on average a time τ . Thus, more charges flow through the contact per unit time than charges actually photogenerated per unit time. What will happen if the contacts are not ohmic, *i.e.*, they are not injecting?
- (c) What can you say about the product $\Delta\sigma$ and the speed of response, which is proportional to $1/\tau$.

5.22 Photoconductive detectors: PbSe for $4 \mu\text{m}$ detection The basic principle in photoconductive detectors is the change in the resistance of the semiconductor upon exposure to light. Figure 5.52 shows how PbS and PbSe photodetectors are often used. The photoconductor (PC) has a dark resistance R_d and is biased by V_B through a load R_L . A chopper (either mechanical or electronic) chops the light at a frequency f_c . The resistance of the PC changes periodically at the chopper frequency. The periodic change in R_d causes a periodic change in the current. This periodic change is the photocurrent $i_{ph}(t)$ signal and is an AC-type signal at the frequency f_c . The photocurrent generates a periodic voltage signal $v_{ph}(t)$ across R_d , which can be coupled through a coupling capacitor C_c into a lock-in amplifier (LIA). This amplifier is synchronized with the chopper and only amplifies the signal if it is in phase with the chopper. Its output is a DC signal that represents the magnitude of v_{ph} that is in phase with the chopped light. Consider a PbSe photoconductor for detecting $4 \mu\text{m}$ radiation. The circuit in Figure 5.52 has been implemented with $V_B = 15 \text{ V}$, $R_d = 1 \text{ M}\Omega$, and $R_L = R_d$. A test experiment has shown that incident optical power of 50 nW on the detector's photosensitive area ($3 \times 3 \text{ mm}^2$) decreases R_d by 6.0Ω . (Put differently, $\delta R_d = -6 \Omega$ for incident radiation of 50 nW .) Find the signal voltage V_{signal} , *i.e.*, the change δV in the voltage across R_d , and the photosensitivity, defined as V_{signal}/P_o .

(b) What is the photosensitivity if $R_L = 2R_d$ and $R_L = R_d/2$? (c) For the case $R_L = R_d$, find the NEP and hence the specific detectivity D^* . Assume that the only noise source is the thermal noise from R_d and R_L . The quoted D^* for this device is $2.5 \times 10^9 \text{ cm Hz}^{1/2} \text{ W}^{-1}$.

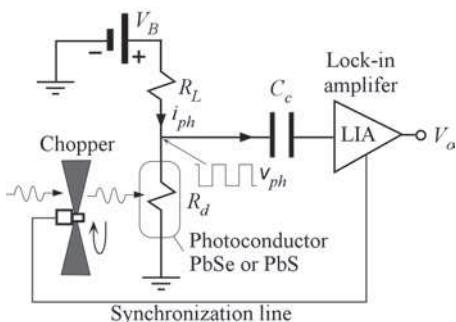


FIGURE 5.52 A photo-detection system based on using a photoconductive detector, a chopper to chop the light, and a lock-in amplifier (LIA).

5.23 SNR of an InGaAs *pin* A particular photodetection application requires an InGaAs photodetector and needs a bandwidth of 2 GHz . The dark current of the InGaAs *pin* detector is 6 nA at 30°C . The minimum signal that is to be measured is 5 nW at 1550 nm where the responsivity is 0.95 A W^{-1} .

- (a) Calculate the SNR in dB at 30°C .
- (b) When the detector is cooled to -30°C , the dark current becomes 0.1 nA , and the responsivity is about the same. What is the new SNR?
- (c) Suppose that we operate the detector over a bandwidth of 20 MHz . What is the new SNR at -30°C ?

5.24 Detector and receiver noise Consider an InGaAs *pin* photodiode used in a receiver circuit as in Figure 5.31 with a load resistor of $30\text{ k}\Omega$. The total capacitance of the detector and the input of the amplifier together is 20 pF . The photodiode has a dark current of 1 nA . The incident radiation is 6 nW at 1550 nm where the responsivity is 0.9 A W^{-1} . Assuming that the amplifier is noiseless, calculate the SNR at 27°C .

5.25 The NEP and Ge and InGaAs photodiodes

- (a) Show that the noise equivalent power of a photodiode is given by

$$\text{NEP} = \frac{P_1}{B^{1/2}} = \frac{hc}{\eta e\lambda} [2e(I_d + I_{ph})]^{1/2}$$

How would you improve the NEP of a photodiode? What is *NEP* for an ideal PD at operating at $\lambda = 1.55\text{ }\mu\text{m}$?

- (b) Given the dark current I_d of a PD, show that for $\text{SNR} = 1$, the photocurrent is

$$I_{ph} = eB \left[1 + \left(1 + \frac{2I_d}{eB} \right)^{1/2} \right]$$

What is the corresponding optical power P_1 ?

- (c) Consider a fast Ge *pn* junction PD which has a photosensitive area of diameter 0.3 mm . It is reverse biased for photodetection and has a dark current of $0.5\text{ }\mu\text{A}$. Its peak responsivity is 0.7 A V^{-1} at $1.55\text{ }\mu\text{m}$ (see Figure 5.47). The bandwidth of the photodetector and the amplifier circuit together is 100 MHz . Calculate its NEP at the peak wavelength and find the minimum optical power and hence minimum light intensity that gives a SNR of 1. How would you improve the minimum detectable optical power?
- (d) Table 5.10 shows the characteristics of typical Ge *pn* junction and InGaAs *pin* photodiodes in terms of responsivity and the current. Fill in the remainder of the columns in the table assuming that there is an ideal, noiseless preamplifier to detect the photocurrent from the photodiode. Assume a working bandwidth, B , of 1 MHz . What is your conclusion?

TABLE 5.10 Ge *pn* junction and InGaAs *pin* PDs. Photosensitive area has a diameter of 1 mm

Photodiode	R at $1.55\text{ }\mu\text{m} (\text{A V}^{-1})$	I_d (nA)	I_{ph} (nA) for	Optical power (nW)			Comment
			SNR = 1 at $B = 1\text{ MHz}$	$B = 1\text{ MHz}$	for SNR = 1 at $B = 1\text{ MHz}$	NEP ($\text{W Hz}^{-1/2}$)	
Ge at 25°C	0.8	400					
Ge at -20°C	0.8	5					Thermoelectric cooling
InGaAs <i>pin</i>	0.95	3					

5.26 Specific detectivity of Si photodiodes Consider the NEP measurements shown in Table 5.11 for a collection of Si photodiodes. (a) For each calculate the specific detectivity. Then find the average and the mean standard deviation. (b) Plot NEP vs. $A^{1/2}$ on log-log plot and find the slope. What does the slope represent?

TABLE 5.11 Photosensitive area and NEP for a collection of Si photodiodes

A (mm^2)	7	13	20	25	66	100	324	784
NEP ($\times 10^{-15}\text{ W Hz}^{-1/2}$)	9.5	13	17	16	25	38	66	86

5.27 Specific detectivity of InGaAs photodiodes Consider the results of NEP measurements shown in Table 5.12 for a few InGaAs *pin* photodiodes with different photosensitive areas. (a) Plot the dark current vs. A on log-log plot and find the slope. What does the slope represent? (b) Plot NEP vs. A on log-log plot and find the slope. What does the slope mean?

TABLE 5.12 Photosensitive area and NEP for a collection of InGaAs photodiodes at -10°C

$A (\text{mm}^2)$	0.785	3.14	7.07	19.6
Dark current (nA)	0.07	0.3	1	2.5
NEP ($\times 10^{-15} \text{ W Hz}^{-1/2}$)	5	10	20	30

- 5.28 Excess avalanche noise in InGaAs communications detector** Consider a particular commercial 1550 nm InGaAs communications detector whose properties are as follows: Diameter of sensitive area is $75\text{ }\mu\text{m}$, the unmultiplied dark current (I_{do}) is 0.7 nA. Its responsivity at 1550 nm (peak responsivity) is 0.75 A V^{-1} . The excess noise factor F has been measured to be 3.5 at $M = 10$. Assume the operating wavelength is 1550 nm. The APD normally biased to operate at a multiplication (gain) of $M = 10$ and has a bandwidth of 2 GHz when connected to a suitable external load (preamplifier). We will ignore the load and the preamplifier connection and simply focus on the APD's performance on its own within a bandwidth of 2 GHz. (a) Find the NEP under $M = 1$ (with $B = 1\text{ Hz}$). (b) What is the NEP when the APD is operating at $M = 10$? (c) What is the incident detectable optical power and hence intensity for operation at an SNR of 10 at $M = 1$. Suppose that this incident power is kept the same but we introduce gain and set $M = 10$. What is the new SNR? What is your conclusion?

- 5.29 Avalanche photodiodes and excess avalanche noise** APDs exhibit excess avalanche noise which contributes to the shot noise of the diode current. The total noise current in the APD is given by

$$i_{n-\text{APD}} = [2e(I_{do} + I_{pho})M^2FB]^{1/2} \quad (\text{P1.1})$$

where F is the excess noise factor which depends in a complicated way not only on M but also on the ionization probabilities of the carriers in the device. It is normally taken simply to be M^x where x is an index that depends on the semiconductor material and device structure.

- (a) Table 5.13 provides measurements of F vs. M on a Ge APD using photogeneration at $1.55\text{ }\mu\text{m}$. Find x in $F = M^x$. How good is the fit?

TABLE 5.13 Data for excess avalanche noise as F vs. M for a Ge APD

M	1	3	5	7	9
F	1.1	2.8	4.4	5.5	7.5

(Source: Data from D. Scansen and S. O. Kasap, *Cnd. J. Phys.* 70, 1070, 1992.)

- (b) The above Ge APD has an unmultiplied dark current of $0.5\text{ }\mu\text{A}$ and an unmultiplied responsivity of 0.8 A V^{-1} at its peak response at $1.55\text{ }\mu\text{m}$ and is biased to operate at $M = 6$ in a receiver circuit with a bandwidth of 500 MHz. What is the minimum photocurrent that will give a SNR = 1? If the photosensitive area is 0.3 mm in diameter what is the corresponding minimum optical power and light intensity?
(c) What should be the photocurrent and incident optical power for SNR = 10? What would be the effect of cooling the APD so that the dark current is reduced by a factor of 100, but M is the same?

5.30 Photodetector materials, devices, and their limitations

- (a) What limits the operation of a photodetector when the absorption depth ($\delta = 1/\alpha$) at short wavelengths becomes so narrow that EHPs are generated almost at the crystal surface?
(b) Quantum efficiency defined in terms of incident photons applies to the whole device and includes the effects arising from reflections from the semiconductor surface. What is the percentage of photons lost due to reflections at a Si crystal surface if the refractive index of Si is 3.5. How can you improve the transmitted number of photons into the semiconductor crystal?
(c) In some applications such as measuring the light intensity in the visible range from a source that also emits extensively in the infrared (such as an incandescent light source), it is necessary to use an infrared heat filter. Why?
(d) Consider a heterojunction APD such as that shown in Figure 5.16. For InP, $E_g = 1.35\text{ eV}$ and for InGaAs $E_g = 0.75\text{ eV}$. Obviously $1.5\text{ }\mu\text{m}$ photons will not be absorbed in InP. What is the effect of mismatch in the refractive index n between the two semiconductors? Is it important if $\Delta n \approx 0.2$, and $n \approx 3.5$?
(e) What determines the speed of operation of the phototransistor in Figure 5.25? Consider how long it takes for the photojected hole into the p -type base to become neutralized by recombination.

5.31 CMOS and CCD image sensors Consider the basic principle of operation of the CMOS and CCD image sensors in Section 5.13. (a) What constitutes the signal in the CMOS and CCD sensors? (b) Using the responsivity or sensitivity definition used for photodetectors, how would you define the responsivity of CMOS and CCD sensors? (c) The **fill factor** defines what fraction of the pixel area is able to capture the light and convert it to a signal. For example, what fraction of the area is allocated to the photodiode in the pixel of a CMOS? Why is FF important? (d) Considering the operation of the CMOS and CCD sensors, which one is faster for a given number of megapixels? (e) Considering the principle of operation, which sensor would have lower noise?

5.32 CCD image sensors

- (a) A particular full frame square CCD image sensor (Kodak KAF-1001) has 1024×1024 pixels, where H and V represent horizontal and vertical directions, respectively. The pixel size is a square with a side $24\text{ }\mu\text{m}$. What is the optical imaging area of the sensor? If the data transfer rate is 2 MHz, what is the total time to transfer one frame of the image?
- (b) In a CCD, charge is transferred from one well to the next, and a quantity of importance is the **charge transfer efficiency** η_t . $1 - \eta_t$ represents the fraction of electrons lost in the transfer process. After n transfers from well to well, the collected charge Q_n with respect to the original signal Q_o is

$$Q_n/Q_o = \eta^n$$

If the CCD in (a) has a $\eta = 0.9997$, calculate the charge collection efficiency of the CCD.

5.33 Solar cell driving a load Consider the I - V characteristics of the solar cell shown in Figure 5.40 (b). Suppose that we connect a resistive load of 5Ω across it. Find the current through and the voltage across the load at 800 W m^{-2} and 400 W m^{-2} .

5.34 Solar cell open circuit voltage and short circuit A solar cell under an illumination of 300 W m^{-2} has a short circuit current I_{sc} of 60 mA and an open circuit output voltage V_{oc} of 0.60 V . What are the short circuit current and open circuit voltages when the light intensity is 600 W m^{-2} ?

5.35 Shunt resistance Consider the equivalent circuit of a solar cell as shown in Figure 5.43 (b).

- (a) Show that

$$I = -I_{ph} + I_{diode} + \frac{V}{R_p} = -I_{ph} + I_o \exp\left(\frac{eV}{\eta k_B T}\right) - I_o + \frac{V}{R_p}$$

- (b) Plot I vs. V for a polycrystalline Si solar cell that has $\eta = 2$ and $I_o = 3 \times 10^{-4}\text{ mA}$, for an illumination such that $I_{ph} = 5\text{ mA}$. Use $R_p = \infty, 1000\Omega$ and then $R_p = 100\Omega$; and plot the curves on the graph for comparison. What is your conclusion?

5.36 Series resistance Consider a solar cell that has $I_o = 30 \times 10^{-6}\text{ mA}$, $\eta = 1.5$, and negligible shunt conductance ($R_p = \infty$). Suppose that the illumination is such that $I_{ph} = 9\text{ mA}$. Plot the individual I - V characteristics for $R_s = 0\Omega, 40\Omega$, and 90Ω .

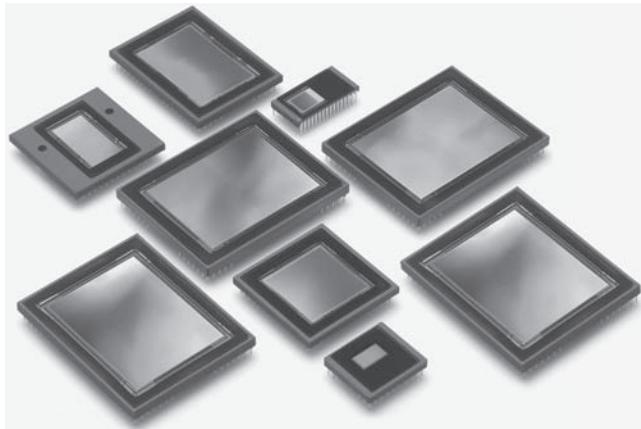
5.37 Solar cell near Eskimo Point

- (a) The intensity of light arriving at a point on earth, where the solar latitude is α , can be approximated by the Meinel and Meinel equation

$$I \approx 1.353(0.7)^{(\text{cosec}\alpha)^{0.678}} \quad (\text{kW m}^{-2})$$

where $\text{cosec } \alpha = 1/\sin \alpha$. The solar latitude α is the angle between the sun rays and the horizon. Around September 23rd and March 22nd, the sun rays arrive parallel to the plane of the equator. Plot intensity I vs. α . What is your conclusion? What is the maximum power available for a photovoltaic device panel of area 1 m^2 if its efficiency of conversion is 20% ?

- (b) Manufacturer's characterization tests on a particular Si *p-n* junction solar cell at 27°C specifies an open circuit output voltage of 0.45 V and a short circuit current of 400 mA when illuminated directly with light of intensity 1 kW m⁻². The fill factor for the solar cell is 0.73. This solar cell is to be used in a portable equipment application near Eskimo Point (Nunavut, Canada) at a latitude (ϕ) of 63°. The latitude ϕ and α are related by $\phi + \alpha = \pi/2$. Calculate the open circuit output voltage and the maximum available power when the solar cell is used at noon on September 23rd, when the temperature is around -10°C. What is the maximum current this solar cell can supply to an electronic equipment? What is your conclusion? (Assume that the ideality factor is 1.)



Various image sensors. (Courtesy of Teledyne-DALSA.)

*As there are two different refractions, I conceived also
that there are two different emanations of the waves of light...*

—Christiaan Huygens



Sailors visiting Iceland during the seventeenth century brought back to Europe calcite crystals (Iceland spar) which had the unusual property of showing double images when objects were viewed through it. The Danish scientist Rasmus Bartholin (Erasmus Bartholinus) described this property in 1669 as the effect of double refraction, and later Christiaan Huygens (1629–1695), a Dutch physicist, explained this double refraction in terms of ordinary and extraordinary waves. Christiaan Huygens made many contributions to optics and wrote prolifically on the subject. (*Morphart Creations Inc./Shutterstock.com*)

6

POLARIZATION AND MODULATION OF LIGHT

6.1 POLARIZATION

A. State of Polarization

A propagating electromagnetic (EM) wave has its electric and magnetic fields at right angles to the direction of propagation. If we place a z -axis along the direction of propagation, then the electric field can be in any direction in the plane perpendicular to the z -axis. The term **polarization** of an EM wave describes the behavior of the electric field vector in the EM wave as it propagates through a medium. If the oscillations of the electric field at a given point at all times are contained within a well-defined line, then the EM wave is said to be **linearly polarized** as shown in Figure 6.1 (a). The field oscillations and the direction of propagation (z) define a plane of polarization (plane of oscillation) so that linear polarization implies a wave that is **plane-polarized**. By contrast, if a beam of light has waves with the E -field in each wave in a random direction perpendicular to z , then this light beam is *unpolarized*. A light beam can be linearly polarized by passing the beam through a polarizer, such as a polaroid sheet, a device that only passes electric field oscillations lying on a well-defined plane at right angles to the direction of propagation.

Suppose that we arbitrarily place x - and y -axes, and describe the electric field in terms of its components E_x and E_y along x and y , as defined by an observer receiving the EM wave; x and y are perpendicular to each other and the direction of propagation¹ as in Figure 6.1 (a). To find the electric field in the wave at any space and time location, we add E_x and E_y *vectorially*. Both E_x and E_y can individually be described by a wave equation, which must have the same angular frequency ω and propagation constant k . However, we must include a phase difference ϕ between the two

$$E_x = E_{x0} \cos(\omega t - kz) \quad (6.1.1)$$

and

$$E_y = E_{y0} \cos(\omega t - kz + \phi) \quad (6.1.2)$$

¹The directions of x - and y -axes are taken as they would be drawn by an observer receiving the light wave. In this case, the observer ignores the rigid Cartesian directions, which would make the x -axis point to the left not right in Figure 6.1 (a). The definitions and discussions that follow are actually unaffected by the choice of x and y in Figure 6.1 (a).

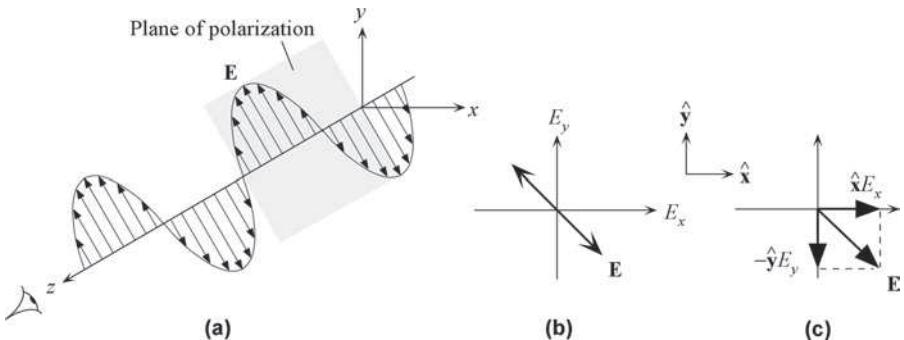


FIGURE 6.1 (a) A linearly polarized wave has its electric field oscillations defined along a line perpendicular to the direction of propagation, z . The field vector E and z define a plane of polarization. (b) The E -field oscillations are contained in the plane of polarization. (c) A linearly polarized light at any instant can be represented by the superposition of two fields, E_x and E_y , with the right magnitude and phase.

where ϕ is the phase difference between E_y and E_x . A phase difference ϕ can arise if one of the components is delayed, and this delay is termed *retardation*.

The linearly polarized wave in Figure 6.1 (a) has the E -field oscillations at -45° to the x -axis as shown in (b). We can generate this field by choosing $E_{xo} = E_{yo}$ and $\phi = \pm 180^\circ (\pm \pi)$ in Eqs. (6.1.1) and (6.1.2). Put differently, E_x and E_y have the same magnitude but they are out of phase by 180° . If \hat{x} and \hat{y} are the unit vectors along x and y , using $\phi = \pi$ in Eq. (6.1.2), the field in the wave is

$$\mathbf{E} = \hat{x}E_x + \hat{y}E_y = \hat{x}E_{xo}\cos(\omega t - kz) - \hat{y}E_{yo}\cos(\omega t - kz)$$

or

$$\mathbf{E} = \mathbf{E}_o\cos(\omega t - kz) \quad (6.1.3)$$

where

$$\mathbf{E}_o = \hat{x}E_{xo} - \hat{y}E_{yo} \quad (6.1.4)$$

Equation (6.1.4) describes the vector amplitude in Eq. (6.1.3), and shows that the vector \mathbf{E}_o is at -45° to the x -axis. Equation (6.1.3) describes the propagation of this \mathbf{E}_o at -45° along the z -direction.

There are many choices for the behavior of the electric field besides the simple linear polarization in Figure 6.1 (a). For example, if the magnitude of the field vector \mathbf{E} remains constant but its tip at a given location on z traces out a circle by rotating in a clockwise sense with time, as observed by the receiver of the wave, then the wave is said to be **right circularly polarized**,² as in Figure 6.2. If the rotation of the tip of \mathbf{E} is counterclockwise, the wave is said to be **left circularly polarized**. From Eqs. (6.1.1) and (6.1.2), it should be apparent that a right circularly polarized wave has $E_{xo} = E_{yo} = A$ (an amplitude), and $\phi = \pi/2$. This means that

$$E_x = A\cos(\omega t - kz) \quad (6.1.5a)$$

²There is a difference in this definition in optics and engineering. The definition here follows that in optics, and it is also prevalent in optoelectronics.

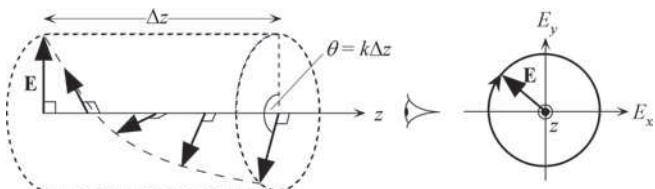


FIGURE 6.2 A right circularly polarized light. The field vector \mathbf{E} is always at right angles to z , rotates clockwise around z with time, and traces out a full circle over one wavelength of distance propagated.

and

$$E_y = -A \sin(\omega t - kz) \quad (6.1.5b)$$

It is relatively straightforward to show Eqs. (6.1.5a) and (6.1.5b) represent a circle, that is,

$$E_x^2 + E_y^2 = A^2 \quad (6.1.6)$$

as shown in Figure 6.2.

The snapshot of the circularly polarized light in Figure 6.2 shows that over a distance Δz , the field \mathbf{E} rotates through an angle $\theta = k\Delta z$. Linear and circular polarization concepts are summarized in Figure 6.3, where, for simplicity, $E_{yo} = 1$ has been taken and the corresponding E_{xo} and ϕ are shown.

An **elliptically polarized**, or elliptic, light has the tip of the \mathbf{E} -vector trace out an ellipse as the wave propagates through a given location in space. As in circular polarization, light can

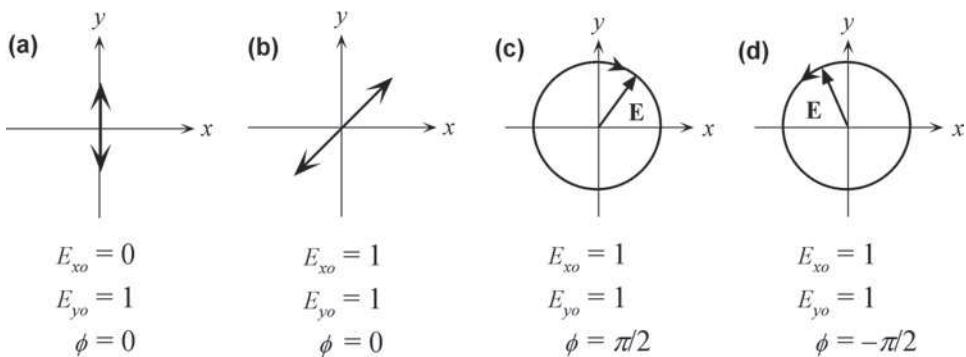


FIGURE 6.3 (a) and (b) are examples of linearly polarized light. (c) and (d) are right circularly and left circularly polarized light, respectively (as seen when the wave directly approaches a viewer).

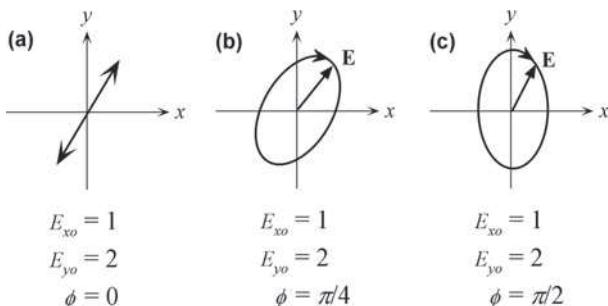


FIGURE 6.4 (a) Linearly polarized light with $E_{yo} = 2E_{xo}$ and $\phi = 0$. (b) When $\phi = \pi/4$ (45°), the light is right elliptically polarized with a tilted major axis. (c) When $\phi = \pi/2$ (90°), the light is right elliptically polarized. If E_{xo} and E_{yo} were equal, this would be right circularly polarized light.

be right and left elliptically polarized depending on clockwise or counterclockwise rotation of the **E**-vector.

Figure 6.4 illustrates how an elliptically polarized light can result for any ϕ not equal to any multiple of π , and when E_{xo} and E_{yo} are not equal in magnitude. Elliptic light can also be obtained when $E_{xo} = E_{yo}$ and the phase difference is $\pm\pi/4$ or $\pm3\pi/4$, etc.

EXAMPLE 6.1.1 Elliptical and circular polarization

Show that if $E_x = A\cos(\omega t - kz)$ and $E_y = B\cos(\omega t - kz + \phi)$, the amplitudes A and B are different and the phase difference ϕ is $\pi/2$, the wave is elliptically polarized.

Solution

From the x and y components we have

$$\cos(\omega t - kz) = E_x/A$$

and

$$\cos(\omega t - kz + \pi/2) = -\sin(\omega t - kz) = E_y/B$$

Using $\sin^2(\omega t - kz) + \cos^2(\omega t - kz) = 1$ we find,

$$\left(\frac{E_x}{A}\right)^2 + \left(\frac{E_y}{B}\right)^2 = 1$$

This is the equation that relates the instantaneous values E_x and E_y of the field along x and y . The equation is a *circle* when $A = B$, as in Figure 6.3 (c) and an *ellipse* when $A \neq B$ as in Figure 6.4 (c).

Further, at $z = 0$, and at $\omega t = 0$, $E = E_x = A$. Later when $\omega t = \pi/2$, $E = E_y = -B$. Thus, the tip of the electric field rotates in a clockwise sense and the wave is right circularly polarized.

B. Malus's Law

There are various optical devices that operate on the polarization state of a wave passing through it and thereby modify the polarization state. A linear polarizer will only allow electric field oscillations along some preferred direction, called the **transmission axis**, to pass through the device, as illustrated in Figure 6.5. A *polaroid sheet* is a good example of a commercially available linear polarizer. Dichroic crystals such as tourmaline crystals are good polarizers because they are optically anisotropic and attenuate EM waves with fields that are *not* oscillating parallel to a particular axis, called the transmission axis. The emerging beam from the polarizer has its field oscillations along the transmission axis and therefore it is *linearly polarized*; see Figure 6.5.

Suppose that the linearly polarized light from the polarizer is now incident on a second identical polarizer. Then, by rotating the transmission axis of this second polarizer we can analyze the polarization state of the incident beam; hence the second polarizer is called an **analyzer**. If the transmission axis of the second polarizer is at an angle θ to the **E**-field of the incident beam (*i.e.*, to the first polarizer) then only the component $E\cos\theta$ of the field will be allowed to pass through the analyzer. The irradiance (intensity) of light passing through the analyzer is proportional to the square of the electric field which means that the detected intensity varies as $(E\cos\theta)^2$. Since all the electric field will pass when $\theta = 0$ (**E** parallel to TA_2),

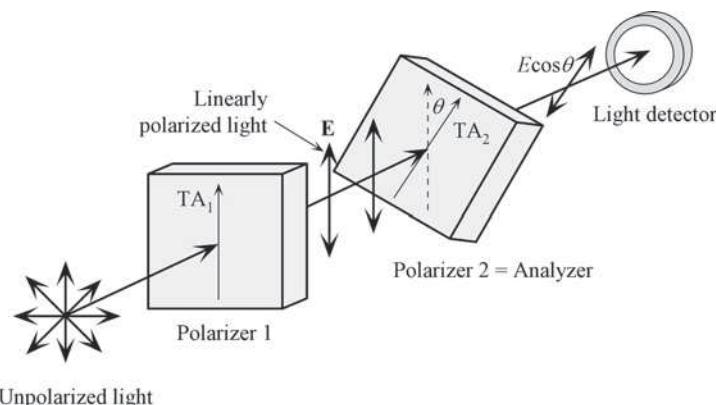


FIGURE 6.5 Randomly polarized light is incident on Polarizer 1 with a transmission axis TA₁. Light emerging from Polarizer 1 is linearly polarized with E along TA₁, and becomes incident on Polarizer 2 (called the analyzer) with a transmission axis TA₂ at an angle θ to TA₁. A detector measures the intensity of the incident light. TA₁ and TA₂ are perpendicular to the light direction.

this is the maximum irradiance condition. The irradiance (or intensity) I at any other angle θ is then given by **Malus's law**³

$$I(\theta) = I(0)\cos^2\theta \quad (6.1.7) \quad \text{Malus's law}$$

Malus's law therefore relates the intensity of a linearly polarized light passing through a polarizer to the angle between the transmission axis and the electric field vector.

6.2 LIGHT PROPAGATION IN AN ANISOTROPIC MEDIUM: BIREFRINGENCE

A. Optical Anisotropy

An important characteristic of crystals is that many of their properties depend on the crystal direction, that is, crystals are generally anisotropic. The dielectric constant ϵ_r depends on electronic polarization, which involves the displacement of electrons with respect to positive atomic nuclei. Electronic polarization depends on the crystal direction inasmuch as it is easier to displace electrons along certain crystal directions. This means that *the refractive index n of a crystal depends on the direction of the electric field in the propagating light beam*. Consequently, the velocity of light in a crystal depends on the direction of propagation and on the state of its polarization, *i.e.*, the direction of the electric field. Most noncrystalline materials, such as glasses and liquids, and all cubic crystals are **optically isotropic**, that is, the refractive index is the same in all directions. For all classes of crystals excluding cubic structures, the refractive index depends on the propagation direction and the state of polarization. The result of optical anisotropy is that, except along certain special directions, any unpolarized light ray entering such a crystal breaks into two different rays with different polarizations and

³ I is used for the intensity in this chapter. Etienn-Louis Malus (1775–1812) was a French physicist and an engineer, who studied the polarization properties of light.

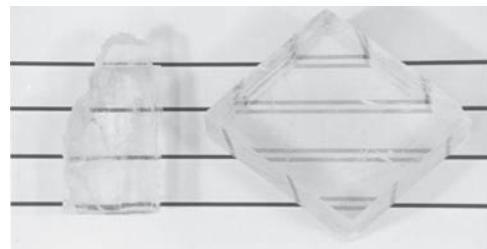


FIGURE 6.6 A line viewed through a cubic sodium chloride (halite) crystal (optically isotropic) and a calcite crystal (optically anisotropic).

phase velocities. When we view an image through a calcite crystal, an optically anisotropic crystal, we see two images, each constituted by light of different polarization passing through the crystal, whereas there is only one image through an optically isotropic crystal as illustrated in Figure 6.6. Optically anisotropic crystals are called **birefringent** because an incident light beam may be *doubly refracted*. There is also a special direction in a birefringent crystal, called the **optic axis**, along which all waves with different polarizations experience the same refractive index or travel with the same phase velocity. A light entering a birefringent crystal along the optic axis does not experience double refraction. Depending on the crystal type, there may be one or two optic axes.

Experiments and theories on most anisotropic crystals, *i.e.*, those with the highest degree of anisotropy, show that we can describe light propagation in terms of *three* refractive indices, called **principal refractive indices**, n_1 , n_2 , and n_3 , along three mutually orthogonal directions in the crystal, say x , y , and z , called *principal axes*. These indices correspond to the polarization state of the wave along these axes.

Crystals that have three distinct principal indices also have *two* optic axes and are called **biaxial crystals**. On the other hand, **uniaxial crystals** have two of their principal indices the same ($n_1 = n_2$) and only have *one* optic axis. Table 6.1 summarizes crystal classifications according to optical anisotropy. Uniaxial crystals, such as quartz, that have $n_3 > n_1$ are called **positive**, and those such as calcite that have $n_3 < n_1$ are called **negative** uniaxial crystals.

TABLE 6.1 Principal refractive indices of some optically isotropic and anisotropic crystals (near 589 nm, yellow Na-D line)

Optically isotropic	$n = n_o$		
Glass (crown)	1.510		
Diamond	2.417		
Fluorite (CaF_2)	1.434		
Uniaxial-Positive			
	n_o	n_e	
Ice	1.309	1.3105	
Quartz	1.5442	1.5533	
Rutile (TiO_2)	2.616	2.903	
Uniaxial-Negative			
	n_o	n_e	
Calcite (CaCO_3)	1.658	1.486	
Tourmaline	1.669	1.638	
Lithium niobate (LiNbO_3)	2.29	2.20	
Biaxial			
	n_1	n_2	n_3
Mica (muscovite)	1.5601	1.5936	1.5977

B. Uniaxial Crystals and Fresnel's Optical Indicatrix

For our discussions of optical anisotropy, we will consider uniaxial crystals such as calcite and quartz. All experiments and theories lead to the following basic principles⁴

Any EM wave entering an anisotropic crystal splits into two orthogonal linearly polarized waves, which travel with different phase velocities, that is, they experience different refractive indices. These two orthogonally polarized waves in uniaxial crystals are called *ordinary* (*o*) and *extraordinary* (*e*) waves. The *o*-wave has the same phase velocity in all directions and behaves like an ordinary wave in which the field is perpendicular to the phase propagation direction. The *e*-wave has a phase velocity that depends on its direction of propagation and its state of polarization; and the electric field in the *e*-wave is not necessarily perpendicular to the phase propagation direction. These two waves propagate with the same velocity only along a special direction called the **optic axis**. The *o*-wave is always perpendicularly polarized to the optic axis and obeys the usual Snell's law.

The two images observed in Figure 6.6 are due to *o*-waves and *e*-waves being refracted differently so that when they emerge from the crystal they have been separated. Each ray constitutes an image but the optical field directions are orthogonal. This is easily demonstrated by using two polaroid analyzers with their transmission axes at right angles as in Figure 6.7. If we were to view an object along the optic axis of the crystal, we would not see two images because the two rays would experience the same refractive index.

As mentioned above, we can represent the optical properties of a crystal in terms of three refractive indices along three orthogonal axes, the principal axes of the crystal, shown as *x*, *y*, and *z* in Figure 6.8 (a). These are special axes along which the polarization vector and the electric field are parallel. (Put differently, the electric displacement⁵ \mathbf{D} and the electric field \mathbf{E} vectors are parallel.) The refractive indices along these *x*-, *y*-, and *z*-axes are the principal indices n_1 , n_2 ,

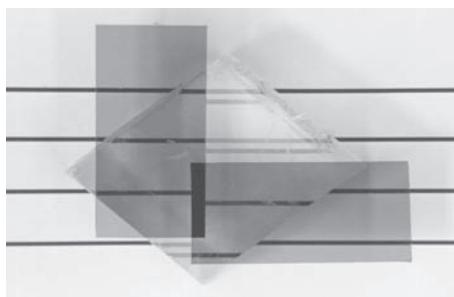


FIGURE 6.7 Two polaroid analyzers are placed with their transmission axes, along the long edges, at right angles to each other. The ordinary ray, undeflected, goes through the left polarizer, whereas the extraordinary wave, deflected, goes through the right polarizer. The two waves therefore have orthogonal polarizations.

⁴These statements can be proved by solving Maxwell's equations in an anisotropic medium. The discussion of light wave propagation in anisotropic media invariably involves dealing with the displacement vector \mathbf{D} , which is not always covered in undergraduate curricula.

⁵Electric displacement \mathbf{D} at any point is defined by $\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}$ where \mathbf{E} is the electric field and \mathbf{P} is the polarization at that point. In many of the discussions in this chapter we will simply ignore the angle between the electric field \mathbf{E} and \mathbf{D} . A readable treatment of EM wave propagation in anisotropic media can be found in G. R. Fowles, *Introduction to Modern Optics*, 2nd Edition (Dover Publications, 1975), Ch. 6.

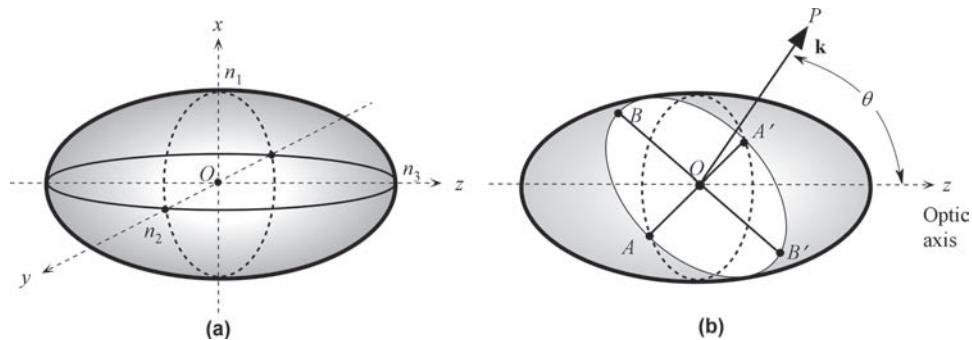


FIGURE 6.8 (a) Fresnel's ellipsoid (for $n_1 = n_2 < n_3$; quartz). (b) An EM wave propagating along OP at an angle θ to the optic axis.

and n_3 , respectively, for electric field oscillations along these directions (not to be confused with the wave propagation direction). For example, for a wave with a polarization parallel to the x -axis, the refractive index is n_1 .

The refractive index associated with a particular EM wave in a crystal can be determined by using Fresnel's *refractive index ellipsoid*, called the **optical indicatrix**,⁶ which is a refractive index surface placed in the center of the principal axes, as shown in Figure 6.8 (a), where the x -, y -, and z -axes have intercepts n_1 , n_2 , and n_3 , respectively. If all three indices were the same, $n_1 = n_2 = n_3 = n_o$, we would have a spherical surface and all electric field polarization directions would experience the same refractive index n_o . Such a spherical surface would represent an optically isotropic crystal. For positive uniaxial crystals such as quartz, $n_1 = n_2 < n_3$, which is the example in Figure 6.8 (a).

Suppose that we wish to find the refractive indices experienced by a wave traveling with an arbitrary wave vector \mathbf{k} , which represents the direction of phase propagation. This phase propagation direction is shown as OP in Figure 6.8 (b) and is at an angle θ to the z -axis. We place a plane perpendicular to OP and passing through the center O of the indicatrix. This plane intersects the ellipsoid surface in a curve $ABA'B'$ which is an *ellipse*. The major (BOB') and minor (AOA') axes of this ellipse determine the field oscillation directions and the refractive indices associated with this wave. Put differently, the original wave is now represented by two orthogonally polarized EM waves.

The line AOA' , the *minor axis*, corresponds to the polarization of the ordinary wave, and its semiaxis OA is the refractive index $n_o = n_2$ of this *o*-wave. The electric displacement and the electric field are in the same direction and parallel to AOA' . If we were to change the direction of OP , we would always find the same minor axis, *i.e.*, n_o is either n_1 or n_2 whatever the orientation of OP (try orientating OP to be along y and along x). Since $n_1 = n_2$ for uniaxial crystals [Figure 6.8 (a)], this means that the *o*-wave always experiences the same refractive index in all directions. The *o*-wave behaves just like an ordinary wave (hence the name).

The line BOB' in Figure 6.8 (b), the *major axis*, corresponds to the electric displacement field (**D**) oscillations in the extraordinary wave and its semiaxis OB is the refractive index $n_e(\theta)$ of this *e*-wave traveling along OP at an angle θ to z . This refractive index is smaller than n_3 but greater than

⁶There are various names in the literature with various subtle nuances: the *Fresnel ellipsoid*, *optical indicatrix*, *index ellipsoid*, *reciprocal ellipsoid*, *Poinsot ellipsoid*, *ellipsoid of wave normals*.

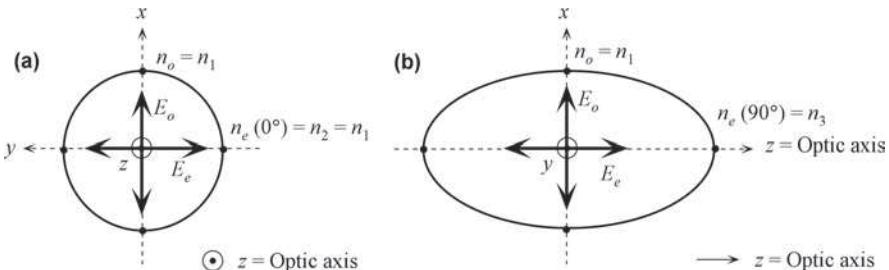


FIGURE 6.9 $E_o = E_o$ -wave and $E_e = E_e$ -wave. (a) Wave propagation along the optic axis. (b) Wave propagation normal to optic axis.

$n_2 (= n_o)$. The e -wave therefore travels more slowly than the o -wave in this particular direction and in this crystal. If we change the direction of OP , we find that the length of the major axis changes with the OP direction. Thus, $n_e(\theta)$ depends on the wave direction, θ . As apparent, $n_e = n_o$ when OP is along the z -axis, that is, when the wave is traveling along z as in Figure 6.9 (a). This direction is the **optic axis** and all waves traveling along the optic axis have the same phase velocity whatever their polarization. When the e -wave is traveling along the y -axis or along the x -axis, $n_e(90^\circ) = n_3 = n_e$ and the e -wave has its slowest phase velocity as shown in Figure 6.9 (b). Along any OP direction that is at an angle θ to the optic axis, the e -wave has a refractive index $n_e(\theta)$ given by

$$\frac{1}{n_e(\theta)^2} = \frac{\cos^2\theta}{n_o^2} + \frac{\sin^2\theta}{n_e^2} \quad (6.2.1)$$

Refractive index of the e -wave

Clearly, for $\theta = 0^\circ$, $n_e(0^\circ) = n_o$ and for $\theta = 90^\circ$, $n_e(90^\circ) = n_e$.

The major axis BOB' in Figure 6.8 (b) determines the e -wave polarization by defining the direction of the displacement vector \mathbf{D} and not \mathbf{E} . Although \mathbf{D} is perpendicular to \mathbf{k} , this is not true for \mathbf{E} . The electric field of the e -wave is orthogonal to that of the o -wave and it is in the plane determined by \mathbf{k} and the optic axis, as discussed below. \mathbf{E} is orthogonal to \mathbf{k} when the e -wave propagates along one of the principal axes.

From the indicatrix, or equivalently from Eq. (6.2.1), we can easily determine the refractive indices of the o - and e -waves in any direction and thereby calculate their wave vectors. We can then construct a **wave vector surface** for all the o and e -waves in the crystal, as shown for a particular cross-section in Figure 6.10 (a). The distance from the origin O to any arbitrary point P on a wave vector surface represents the value of \mathbf{k} along the direction OP . Since the o -wave has the same refractive index in all directions, its wave vector surface will be a sphere of radius $n_o k_{\text{vacuum}}$, where k_{vacuum} is the wave vector in free space. This is shown as a circle in an “xz” cross-section in Figure 6.10 (a). On the other hand, the wave vector of the e -wave depends on the propagation direction and is given by $n_e(\theta) k_{\text{vacuum}}$ so its surface is an ellipse in the “xz” cross-section in Figure 6.10 (a). Two example wave vectors \mathbf{k}_o and \mathbf{k}_e that represent an o -wave and an e -wave propagating along arbitrary directions OP and OQ , respectively, are illustrated in Figure 6.10 (a) (different directions chosen only for clarity).

The electric field \mathbf{E}_o of the o -wave is always orthogonal to its wave vector direction \mathbf{k}_o and also to the optic axis. This fact is shown as dots on the o -wave wave vector surface in Figure 6.10 (a) and is highlighted for an arbitrary \mathbf{k}_o along OP . Since the electric and magnetic fields in the o -wave are normal to \mathbf{k}_o , the o -wave Poynting vector \mathbf{S}_o , i.e., the direction of energy flow, is along \mathbf{k}_o .

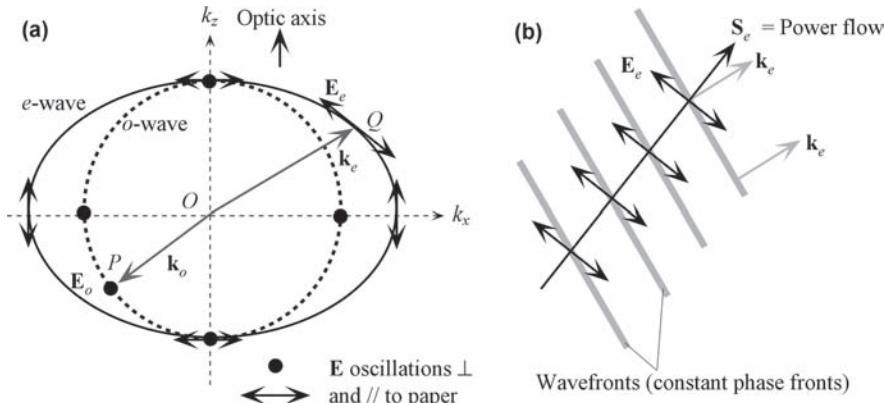


FIGURE 6.10 (a) Wave vector surface cuts in the xz plane for o - and e -waves. (b) An extraordinary wave in an anisotropic crystal with a \mathbf{k}_e at an angle to the optic axis. The electric field is not normal to \mathbf{k}_e . The energy flow (group velocity) is along \mathbf{S}_e , which is different than \mathbf{k}_e .

It may be thought that, as in normal EM wave propagation, *i.e.*, as in the o -wave, the electric field \mathbf{E}_e in the e -wave should be normal to the wave vector \mathbf{k}_e . This is not generally true. The reason is that the polarization of the medium is not parallel to the inducing field in the e -wave and consequently, the overall electric field \mathbf{E}_e in the EM wave is *not* at right angles to the phase propagation direction \mathbf{k}_e as indicated in Figure 6.10 (a). This means that the energy flow (group velocity) and the phase velocity directions are different (a phenomenon called the Poynting vector “walk-off” effect). The energy flow, *i.e.*, the Poynting vector \mathbf{S}_e , direction is taken as the **ray direction** for the e -wave so that the wavefronts advance “sideways” as illustrated in Figure 6.10 (b). The group velocity is in the same direction as energy flow (\mathbf{S}_e).

C. Birefringence of Calcite

Consider a calcite crystal (CaCO_3) which is a negative uniaxial crystal and also well-known for its double refraction. When the surfaces of a calcite crystal have been cleaved, *i.e.*, cut along certain crystal planes, the crystal attains a shape that is called a *cleaved form* and the crystal faces are rhombohedrons (parallelogram with angles 78.08° and 101.92°). A cleaved form of the crystal is called a **calcite rhomb**. A plane of the calcite rhomb that contains the optic axis and is normal to a pair of opposite crystal surfaces is called a **principal section**.

Consider what happens when an unpolarized or natural light enters a calcite crystal at normal incidence and thus also normal to a principal section to this surface, but at an angle to the optic axis as shown in Figure 6.11. The ray breaks into ordinary (o) and extraordinary (e) waves with mutually orthogonal polarizations. The waves propagate in the plane of the principal section, as this plane also contains the incident light. The o -wave has its field oscillations perpendicular to the optic axis. It obeys Snell’s law, which means that it enters the crystal undeflected. Thus the direction of E -field oscillations must come out of the paper so that it is normal to the optic axis and also to the direction of propagation. The field E_\perp in the o -ray is shown as dots, oscillating into and out of the paper.

The e -wave has a polarization orthogonal to the o -wave and in the principal section (which contains the optic axis and \mathbf{k}). The e -wave polarization is in the plane of the paper, indicated as E_{\parallel} , in Figure 6.11. It travels with a different velocity and diverges from the o -wave.

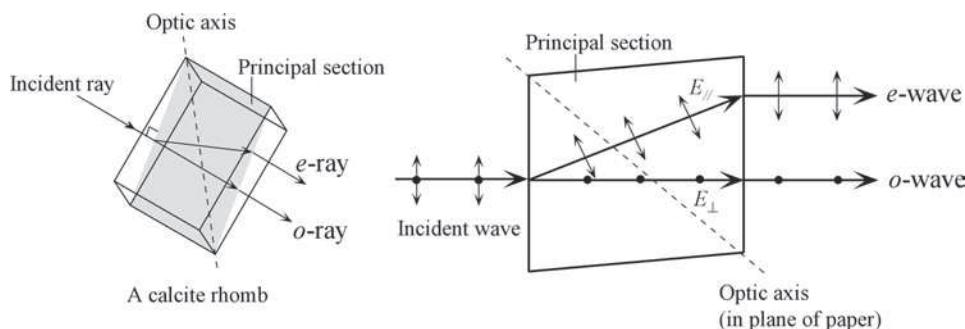


FIGURE 6.11 An EM wave that is off the optic axis of a calcite crystal splits into two waves called ordinary and extraordinary waves. These waves have orthogonal polarizations and travel with different velocities. The *o*-wave has a polarization that is always perpendicular to the optical axis.

Clearly, the *e*-wave does not obey the usual Snell's law inasmuch as the angle of refraction is not zero. We can determine the *e*-ray direction by noting that the *e*-wave propagates sideways as in Figure 6.10 (b) at right angles to E_{\parallel} .

If we were to cut a plate from a calcite crystal so that the optic axis (along z) would be parallel to two opposite faces of the plate as in Figure 6.12 (a), then a ray entering at normal incidence to one of these faces would not diverge into two separate waves. This is the case illustrated in Figure 6.9 (b), that is, propagation along the y -direction, except now $n_e < n_o$. The *o*- and *e*-waves would travel in the same direction but with different speeds. The waves emerge in the same direction as well, which means that we would not see double refraction. This optical arrangement is used in the construction of various optical retarders and polarizers as discussed below. If we were to cut a plate so that the optic axis was perpendicular to the plate faces as in Figure 6.12 (b), then both the *o*- and *e*-waves would be traveling at the same speed [Figure 6.9 (a)] and along the same direction, which means we would not again see any double refraction.

D. Dichroism

In addition to the variation in the refractive index, some anisotropic crystals also exhibit **dichroism**, a phenomenon in which the optical absorption in a substance depends on the direction of propagation and the state of polarization of the light beam. A dichroic crystal is an optically anisotropic crystal in which either the *e*-wave or the *o*-wave is heavily attenuated (absorbed). This means that a

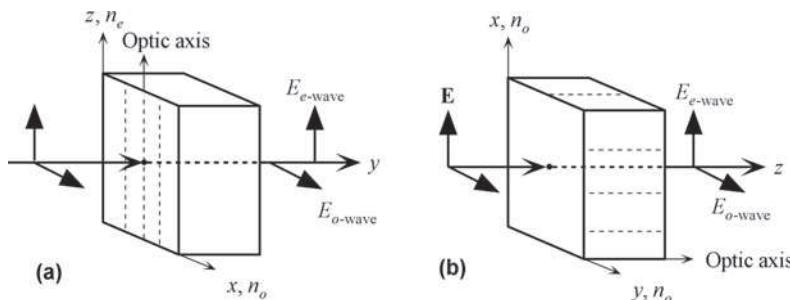


FIGURE 6.12 (a) A birefringent crystal plate with the optic axis parallel to the plate surfaces. (b) A birefringent crystal plate with the optic axis perpendicular to the plate surfaces.

light wave of arbitrary polarization entering a dichroic crystal emerges with a well-defined polarization because the other orthogonal polarization would have been attenuated. Generally, dichroism depends on the wavelength of light. For example, in a tourmaline (aluminum borosilicate) crystal, the *o*-wave is much more heavily absorbed with respect to the *e*-wave.

6.3 BIREFRINGENT OPTICAL DEVICES

A. Retarding Plates

Consider a positive uniaxial crystal plate, such as a quartz ($n_e > n_o$) plate, that has the optic axis (taken along z) parallel to the plate faces as in Figure 6.13. Suppose that a *linearly polarized* wave is incident at normal incidence on a plate face. If the field \mathbf{E} is parallel to the optic axis (shown as E_{\parallel}), then this wave will travel through the crystal as an *e*-wave with a velocity c/n_e slower than the *o*-wave, since $n_e > n_o$. Thus, the optic axis is the “slow axis” for waves polarized parallel to it. If \mathbf{E} is at right angles to the optic axis (shown as E_{\perp}), then this wave will travel with a velocity c/n_o , which will be the fastest velocity in the crystal. Thus, the axis perpendicular to the optic axis (say x) will be the “fast axis” for polarization along this direction. When a light ray enters a crystal at normal incidence to the optic axis and plate surface, as in Figure 6.12 (a), then the *o*- and *e*-waves travel along the same direction as shown in Figure 6.13. We can of course resolve a linear polarization at an angle α to z into E_{\perp} and E_{\parallel} . When the light comes out at the opposite face, these two components would have been phase shifted by ϕ . Depending on the initial angle α of \mathbf{E} and the length of the crystal, which determines the total phase shift ϕ through the plate, the emerging beam can have its initial linear polarization rotated, or changed into an elliptically or circularly polarized light as summarized in Figure 6.14.

If L is the thickness of the plate then the *o*-wave experiences a phase change $k_{o\text{-wave}}L$ through the plate, where $k_{o\text{-wave}}$ is the wave vector of the *o*-wave: $k_{o\text{-wave}} = (2\pi/\lambda)n_o$, where λ is the free-space wavelength. Similarly, the *e*-wave experiences a phase change $(2\pi/\lambda)n_eL$ through the plate. Thus, the phase difference ϕ between the orthogonal components E_{\perp} and E_{\parallel} of the emerging beam is

$$\phi = \frac{2\pi}{\lambda}(n_e - n_o)L \quad (6.3.1)$$

Relative phase through retarder plate

The phase difference ϕ expressed in terms of full wavelengths is called the **retardation** of the plate. For example, a phase difference ϕ of 180° is a half-wavelength retardation.

The polarization of the through beam depends on the crystal type ($n_e - n_o$), and the plate thickness L . We know that, depending on the phase difference ϕ between the orthogonal

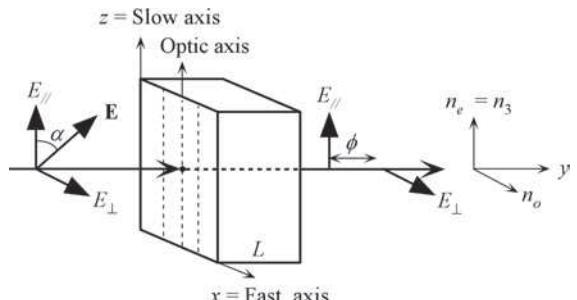


FIGURE 6.13 A retarder plate. The optic axis is parallel to the plate face. The *o*- and *e*-waves travel in the same direction but at different speeds.

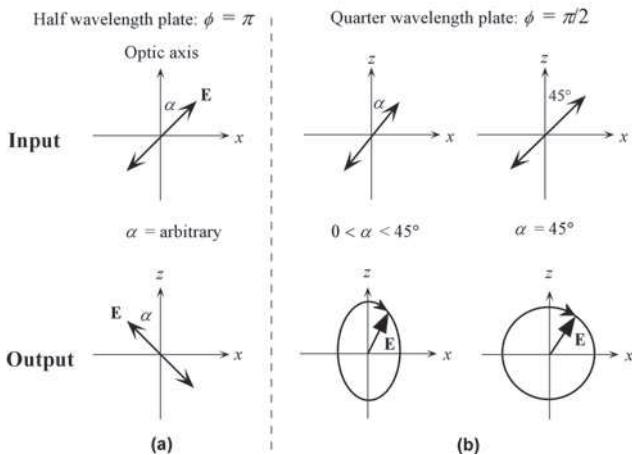


FIGURE 6.14 Input and output polarizations of light through (a) a half-wavelength plate and (b) through a quarter-wavelength plate.

components of the field, the EM wave can be linearly, circularly, or elliptically polarized as in Figure 6.4.

A **half-wave plate retarder** has a thickness L such that the phase difference ϕ is π or 180° , corresponding to a half of wavelength ($\lambda/2$) of retardation. The result is that E_{\parallel} is delayed by 180° with respect to E_{\perp} . If we add the emerging E_{\perp} and E_{\parallel} with this shift ϕ , \mathbf{E} would be at an angle $-\alpha$ to the optic axis and still linearly polarized. \mathbf{E} has been rotated counterclockwise through 2α as shown in Figure 6.14 (a).

A **quarter-wave plate retarder** has a thickness L such that the phase difference ϕ is $\pi/2$ or 90° , corresponding to a quarter of wavelength. If we add the emerging E_{\perp} and E_{\parallel} with this shift ϕ , the emerging light will be elliptically polarized if $0 < \alpha < 45^\circ$ and circularly polarized if $\alpha = 45^\circ$ as illustrated in Figure 6.14 (b).

EXAMPLE 6.3.1 Quartz half-wave plate

What should be the thickness of a half-wave quartz plate for a wavelength $\lambda \approx 590$ nm given the ordinary and extraordinary refractive indices in Table 6.1?

Solution

Half-wavelength retardation is a phase difference of π so that from Eq. (6.3.1)

$$\phi = \frac{2\pi}{\lambda}(n_e - n_o)L = \pi$$

giving,

$$L = \frac{\frac{1}{2}\lambda}{(n_e - n_o)} = \frac{\frac{1}{2}(590 \times 10^{-9} \text{ m})}{(1.5533 - 1.5442)} = 32.4 \mu\text{m}$$

If we were to repeat the calculation for calcite, we would find about $1.7 \mu\text{m}$ thickness, which is not very practical. Typically mica, quartz, or polymeric substances are used as retarder plates because they have a $(n_e - n_o)$ difference that is not too large to result in an impractical thickness.

EXAMPLE 6.3.2 Circular polarization from linear polarization

Consider a linearly polarized light that is incident on a quarter-wavelength plate as in Figure 6.13, such that the polarization is 45° to the slow axis. Show that the output beam is circularly polarized.

Solution

Both the x and z components of the electric field emerging from the retarder plate in Figure 6.13 are propagating along the y -axis with the same $\cos(\omega t - ky)$ harmonic term after emerging from the crystal but with a phase difference ϕ . We are only interested in adding E_x and E_z vectorially at one location so that we can neglect the ky phase term. We can write the field components along fast (x) and slow (z) axes as

$$E_x = E_{\perp} \cos(\omega t) \quad E_z = E_{\parallel} \cos(\omega t - \phi)$$

When the incident polarization is at 45° , then $E_{\parallel} = E_{\perp} = E_o$. Putting $\phi = \pi/2$ for a quarter-wave plate, using $\cos(\omega t - \pi/2) = \sin(\omega t)$, and then $\sin^2(\omega t) + \cos^2(\omega t) = 1$, we find

$$\cos^2(\omega t) + \sin^2(\omega t) = \left(\frac{E_x}{E_o}\right)^2 + \left(\frac{E_z}{E_o}\right)^2 = 1$$

which is the equation of a circle on the E_x and E_z axes [Figure 6.14 (b)] with a radius E_o .

B. Soleil–Babinet Compensator

An optical compensator is a device that allows one to control the retardation (*i.e.*, the phase change) of a wave passing through it. In a wave plate retarder such as the half-wave plate, the relative phase change ϕ between the ordinary and extraordinary waves depends on the plate thickness and cannot be changed. In compensators, ϕ is adjustable. The Soleil–Babinet compensator described below is one such optical device that is widely used for controlling and analyzing the polarization state of light.

Consider the optical structure illustrated in Figure 6.15 which has two quartz wedges touching over their large faces to form a “block” of adjustable height d . Sliding one wedge over the other wedge alters the “thickness” d of this block. The two-wedge block is placed on a parallel plate quartz slab with a fixed thickness D . The slab has its optic axis parallel to its surface face. The optic axes in the wedges are parallel but perpendicular to the optic axis of the slab, as indicated in the figure.

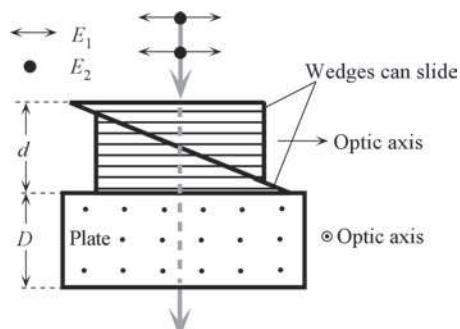


FIGURE 6.15 Soleil–Babinet compensator.



A Soleil–Babinet compensator. (Courtesy of Thorlabs.)



A Wollaston prism. The actual prism is held inside a cylindrical housing. (Courtesy of Thorlabs.)

Suppose that a linearly polarized light is incident on this compensator at normal incidence. We can represent this light by field oscillations parallel and perpendicular to the optic axis of the two-wedge block; these fields are E_1 and E_2 , respectively. The polarization E_1 travels through the wedges (d) experiencing a refractive index n_e (E_1 is along the optic axis) and then travels through the plate (D) experiencing an index n_o (E_1 is perpendicular to the optic axis). Its phase change is

$$\phi_1 = \frac{2\pi}{\lambda} (n_e d + n_o D)$$

But the E_2 -polarization wave first experiences n_o through the wedges (d) and then n_e through the plate (D) so that its phase change is

$$\phi_2 = \frac{2\pi}{\lambda} (n_o d + n_e D)$$

The phase difference $\phi (= \phi_2 - \phi_1)$ between the two polarizations is

$$\phi = \frac{2\pi}{\lambda} (n_e - n_o)(D - d) \quad (6.3.2)$$

Soleil–
Babinet
compensator

It is apparent that, as we can change d continuously by sliding the wedges (by using a micrometer screw), we can continuously alter the phase difference ϕ from 0 to 2π . We can therefore produce a quarter-wave of half-wave plates by simply adjusting this compensator. It should be emphasized that this control occurs over the surface region that corresponds to both the wedges in contact, and in practice this is a narrow region.

C. Birefringent Prisms

Prisms made from birefringent crystals are useful for producing a highly polarized light wave or polarization splitting of light. The **Wollaston prism** is a polarization splitter in which the split beam has orthogonal polarizations. Two calcite (with $n_e < n_o$) right-angle prisms A and B are placed with their diagonal faces touching to form a rectangular block as shown in Figure 6.16. Looking at the cross-section of this block, the optic axis in A is in the plane of the paper, whereas that in B comes out of the paper; the two prisms have their optic axes mutually orthogonal. Further, as shown, the optic axes are parallel to the prism sides.

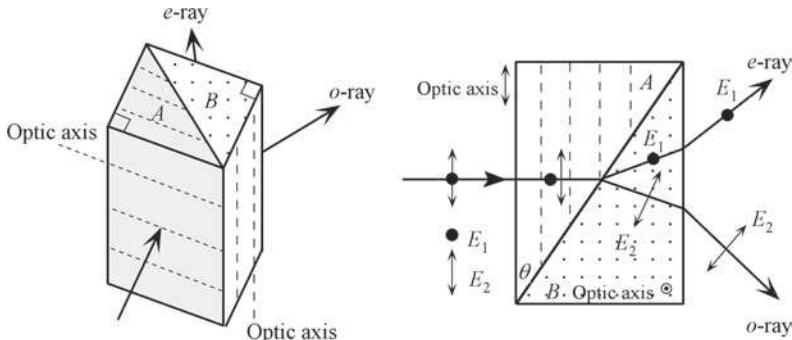


FIGURE 6.16 The Wollaston prism is a beam polarization splitter. E_1 is orthogonal to the plane of the paper and also to the optic axis of the first prism. E_2 is in the plane of the paper and orthogonal to E_1 . Notice that, in prism A, E_1 is the o -wave and E_2 is the e -wave. However, in prism B, E_1 is the e -wave and E_2 is the o -wave.

Consider a light wave of arbitrary polarization at normal incidence to prism A. The light beam entering prism A travels in this prism as two orthogonally polarized waves that have fields E_1 and E_2 as in Figure 6.16. E_1 (normal to the plane of paper) is orthogonal to the optic axis and corresponds to o -waves in A. E_2 (in the plane of the paper) is along the optic axis of A and corresponds to the e -waves. E_1 has a refractive index n_o and E_2 has n_e . However, in prism B, E_1 is the e -wave. This means that in going through the diagonal interface, E_1 experiences a *decrease* from n_o to n_e . On the other hand, the e -wave in A becomes the o -wave in B and experiences an *increase* from n_e to n_o . Notice that E_2 is now orthogonal to the optic axis in B. These refractive index changes are opposite, which means that the two waves are refracted in opposite directions at the interface, as shown in Figure 6.16. The E_1 -wave moves away from the normal to the diagonal face whereas the E_2 -wave moves closer to this normal. The two orthogonal polarizations are therefore angularly separated out by these oppositely sensed refractions. The divergence angle depends on the prism wedge angle θ . Various Wollaston prisms with typical beam splitting angles of 15–45° are commercially available. It is left as an exercise to show that if we rotate the prism about the incident beam by 180°, we would see the two orthogonal polarizations E_1 and E_2 have switched places, and if we use a quartz Wollaston prism, we would find E_1 and E_2 in the figure are again switched.

6.4 OPTICAL ACTIVITY AND CIRCULAR BIREFRINGENCE

When a linearly polarized light wave is passed through a quartz crystal along its optic axis, it is observed that the emerging wave has its **E**-vector (plane of polarization) rotated, which is illustrated in Figure 6.17. This rotation increases continuously with the distance traveled through the crystal (about 21.7° per mm of quartz). The rotation of the plane of polarization by a substance is called **optical activity**. In very simple intuitive terms, optical activity occurs in materials in which the electron motions induced by the external electromagnetic field follow spiraling or helical paths (orbits).⁷ Electrons flowing in helical paths resemble a current flowing in a coil and thus possess a magnetic moment. The optical field in the light therefore induces

⁷The explanation of optical activity involves examining both induced magnetic and electric dipole moments which will not be described here in detail. See Eugene Hecht, *Optics*, 4th Edition (Addison Wesley, Pearson Education, 2002), Section 8.10, for the discussion of optical activity and the derivation of Eq. (6.4.1).

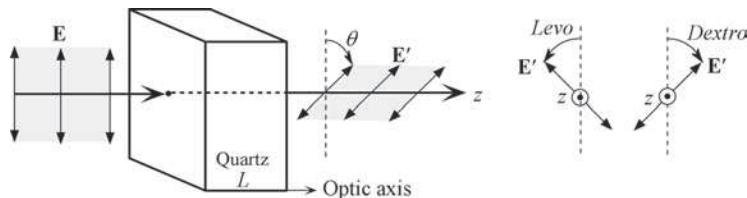


FIGURE 6.17 An optically active material such as quartz rotates the plane of polarization of the incident wave: The optical field \mathbf{E} rotates to \mathbf{E}' . If we reflect the wave back into the material, \mathbf{E}' rotates back to \mathbf{E} .

oscillating magnetic moments which can be either parallel or antiparallel to the induced oscillating electric dipoles. Wavelets emitted from these oscillating induced magnetic and electric dipoles interfere to constitute a forward wave that has its optical field rotated either clockwise or counterclockwise.

If θ is the angle of rotation, then θ is proportional to the distance L propagated in the optically active medium, as illustrated in Figure 6.17. For an observer receiving the wave through quartz, the rotation of the plane of polarization may be *clockwise* (to the right) or *counterclockwise* (to the left), which are respectively called **dextrorotatory** and **levorotatory** forms of optical activity. The structure of quartz is such that atomic arrangements spiral around the optic axis either in clockwise or counterclockwise sense. Quartz thus occurs in two distinct crystalline forms, right-handed and left-handed, which exhibit dextrorotatory and levorotatory types of optical activity, respectively. Although we used quartz as an example, there are many substances that are optically active, including various biological substances and even some liquid solutions (*e.g.*, corn syrup) that contain various organic molecules with a rotatory power.

The **specific rotatory power** is defined as the extent of rotation per unit of distance traveled in the optically active substance, that is, θ/L . Specific rotatory power depends on the wavelength. For example, for quartz this is 49° m^{-1} at 400 nm but 17° m^{-1} at 650 nm .

Optical activity can be understood in terms of left and right circularly polarized waves traveling at different velocities in the crystal, *i.e.*, experiencing different refractive indices. Due to the helical twisting of the molecular or atomic arrangements in the crystal, the velocity of a circularly polarized wave depends on whether the optical field rotates clockwise or counterclockwise. A vertically polarized light with a field \mathbf{E} at the input can be thought of as two right- and left-handed circularly polarized waves, \mathbf{E}_L and \mathbf{E}_R , that are symmetrical with respect to the y -axis, *i.e.*, at any instant $\alpha = \beta$, as shown in Figure 6.18 (α and β are ωt). If they travel at the same velocity through the crystal, then they remain symmetrical with respect to the vertical ($\alpha = \beta$ remains the same) and the resultant is still a vertically polarized light. If, however, these travel at different velocities through a medium, then the output \mathbf{E}'_L and \mathbf{E}'_R are no longer symmetrical with respect to the vertical, $\alpha' \neq \beta'$, and their resultant is a vector \mathbf{E}' at an angle θ to y -axis.

Suppose that n_R and n_L are the refractive indices experienced by the right- and left-handed circularly polarized light, respectively. After traversing the crystal length L , the phase difference between the two optical fields \mathbf{E}'_L and \mathbf{E}'_R at the output leads to a new optical field \mathbf{E}' , which corresponds to \mathbf{E} rotated by θ , given by

$$\theta = \frac{\pi}{\lambda} (n_L - n_R)L \quad (6.4.1) \quad \text{Optical activity}$$

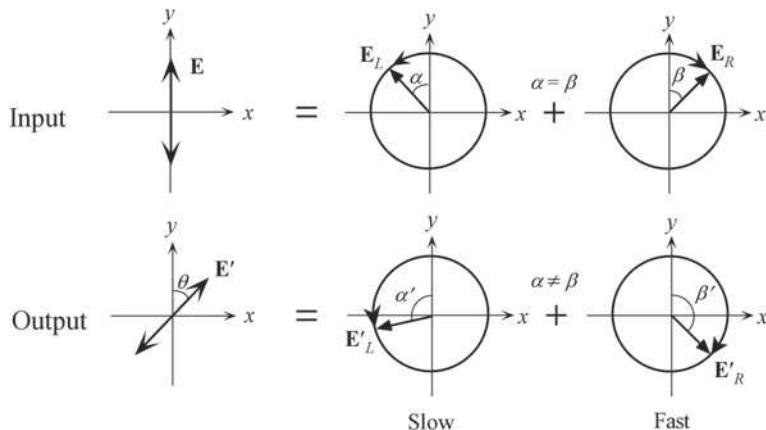


FIGURE 6.18 Vertically polarized wave at the input can be thought of as two right- and left-handed circularly polarized waves that are symmetrical, *i.e.*, at any instant $\alpha = \beta$. If these travel at different velocities through a medium, then at the output they are no longer symmetric with respect to y , $\alpha' \neq \beta'$, and the result is a vector \mathbf{E}' at an angle θ to y .

where λ is the free-space wavelength. For a left-handed quartz crystal, and for 589 nm light propagation along the optic axis, $n_R = 1.54427$ and $n_L = 1.54420$, which means θ is about 21.4° per mm of crystal.

In a **circularly birefringent** medium, the right- and left-handed circularly polarized waves propagate with different velocities and experience different refractive indices, n_R and n_L . Since optically active materials naturally rotate the optical field, it is not unreasonable to expect that a circularly polarized light with its optical field rotating in the same sense as the optical activity will find it easier to travel through the medium. Thus, an optically active medium possesses different refractive indices for right- and left-handed circularly polarized light and exhibits circular birefringence. It should be mentioned that if the direction of the light wave is reversed in Figure 6.17, the ray simply retraces itself and \mathbf{E}' becomes \mathbf{E} .

6.5 LIQUID CRYSTAL DISPLAYS

Many flat-panel televisions and computer displays are liquid crystal displays (LCDs) in which liquid crystals cause changes in the polarization of a passing beam of light. **Liquid crystals** (LCs) are materials that have a rod-like molecular structure as shown in Figure 6.19 (a). These molecules, called **mesogens**, have strong dipoles, which means that the whole structure can be easily polarized. LCs essentially have properties that are between those of a liquid phase and those of a crystalline solid phase; *e.g.*, they can flow like a liquid but, at the same time, possess crystalline domains that lead to anisotropic optical properties. A distinct characteristic of the liquid crystal state is the tendency of the mesogens to point along a common axis called the **director**, a preferred common axis in the liquid crystal that results in an orientationally ordered state, as illustrated in Figure 6.19 (b). This behavior is very different than the way in which molecules behave in a normal liquid phase, where there is no intrinsic order. The orientational order in the liquid crystal state lies between that of a normal crystalline solid, *i.e.*, fully ordered periodic structure, and that of a normal liquid, *i.e.*, nearly fully disordered, and hence is given the name **mesogenic state**.

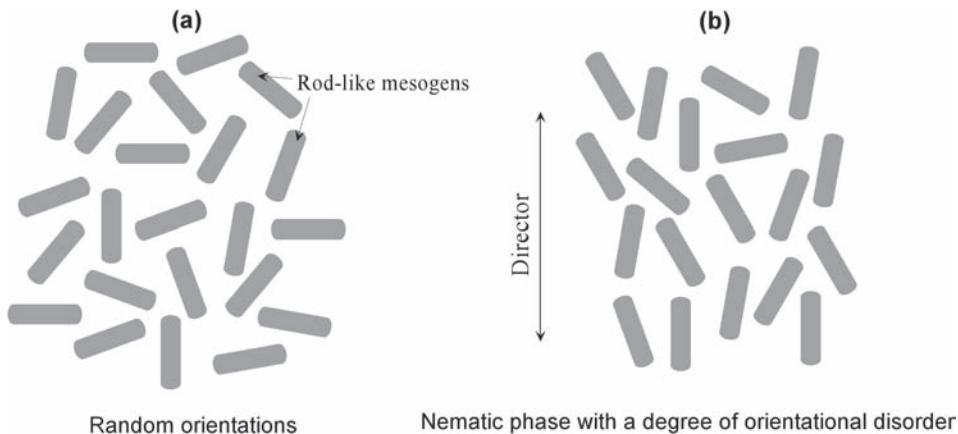


FIGURE 6.19 Schematic illustration of orientational disorder in a liquid with rod-like mesogens. (a) No order, and rods are randomly oriented. (b) There is a tendency for the rods to align with the *director*, the vertical axis in this example.

The number of mesogens pointing along the director, that is, the *degree of anisotropy*, depends on the temperature because thermally induced random motions of the mesogens act against dipole alignment. The degree of alignment will be a maximum at low temperature and decreases with increasing temperature, until at some critical temperature thermal agitation will result in a state of no net ordering. There are a number of possible liquid crystal phases. We will consider the **nematic phase**, which is characterized by mesogens that have no positional order, but tend to point along the same direction, *i.e.*, along the director. Physical properties of these materials tend to depend sensitively on the degree of alignment, and can be highly *anisotropic* for well-aligned materials. A distinct advantage is that the molecular orientation and hence the optical properties can be controlled by an applied field. The molecules have rod-like shapes with lengths typically in the 20–30 nm range as shown in Figure 6.19 (a).

Liquid crystal display is a display that uses a liquid crystal medium whose optical properties can be controlled by an applied field. We can consider it as an LC *light modulator* or a **light valve**. The display has a thin film of *liquid crystal*, *e.g.*, a few microns in thickness, placed between two semitransparent electrically conducting electrodes to form a cell. Most LCDs are based on the **twisted nematic field effect**.⁸ In a **twisted nematic liquid crystal cell**, as shown in Figure 6.20 (a), the two electrodes have surfaces that have been treated, *i.e.*, have an orientational layer, to act as directors for the molecules, and the directors are at right angles to each other. Molecules next to the surfaces are forced to align along these surface directors and hence the molecules homogeneously twist through 90° from one electrode to the other. The **twisted nematic liquid crystal** has its molecules arranged in helical structure, and is able to “twist,” or rotate the optical field in the light that passes through it. Two polarizers at the entrance and the exit ends of the cell have their transmission axes at 90° to each other as shown in Figure 6.20 (a). Polarized light thus enters the cell and has its polarization rotated by 90° as it arrives at the exit polarizer. Since this light has its polarization aligned with the

⁸ Although a number of researchers have reported interesting observations on the optical properties of liquid crystals, the pioneering work on the twisted nematic LCD has been attributed to Martin Schadt and Wolfgang Helfrich (at Hoffmann-LaRoche, Switzerland) in 1970–1971 and James Ferguson (USA) in 1971.

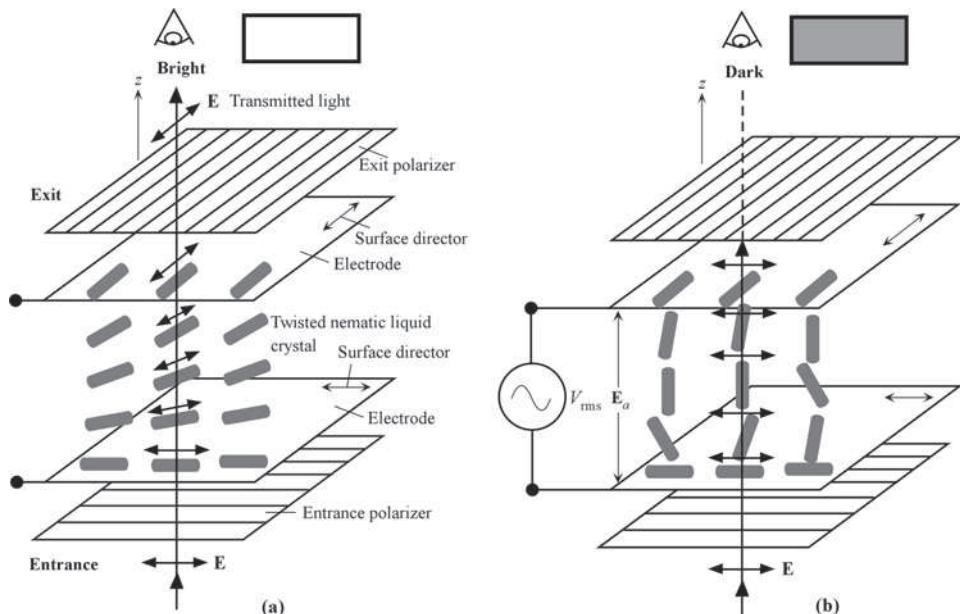


FIGURE 6.20 Transmission-based LCD. (a) In the absence of a field, the liquid crystal has the twisted nematic phase and the light passing through it has its polarization rotated by 90° . The light is transmitted through both polarizers. The viewer sees a bright image. (b) When a voltage, and hence a field \mathbf{E}_a is applied, the molecules in the liquid crystal align with the field and are unable to rotate the polarization of the light passing through it; light therefore cannot pass through the exit polarizer. The light is extinguished, and the viewer sees a dark image.

transmission axis of the exit polarizer, it passes through. Thus, without an applied field, the light is transmitted and the display is bright.

Suppose that an electric field \mathbf{E}_a is applied by supplying an AC voltage (usually a few volts) to the two electrodes comprising the faces of the cell. The applied field now disturbs the alignment of the molecules in the nematic liquid crystal. The field \mathbf{E}_a acts as an externally imposed director and the molecules align with the field, which results in the twisted molecular arrangement being destroyed. Put differently, the helical structure in Figure 6.20 (a) becomes unwound and we end up with the structure shown in Figure 6.20 (b). The polarization in the light entering the cell is unaffected and therefore the light cannot pass through the exit polarizer. The LCD cell therefore appears dark. In fact, the light transmission can be completely extinguished by applying a suitably large field. If a mirror is placed behind the second polarizer, the display can be operated under reflection instead of transmission.

A reverse switching behavior, that is, switching the LCD from dark (without an applied voltage) to bright (with applied voltage), can be achieved by using *parallel* polarizers at the light entrance and exit. By varying the applied voltage between the threshold for reorientation and the saturation field for unwinding the twisted nematic structure, we can obtain gray-scale modulation. The transparent electrodes are typically indium-tin-oxide, and can be patterned by lithographic techniques into various desirable shapes. More than 50% of TV screens use the LCD technology.

The arrangement in Figure 6.20 indicates that an AC voltage must be applied to the LCD. This is indeed the case and LCDs are always operated with an AC voltage; typical operating frequencies

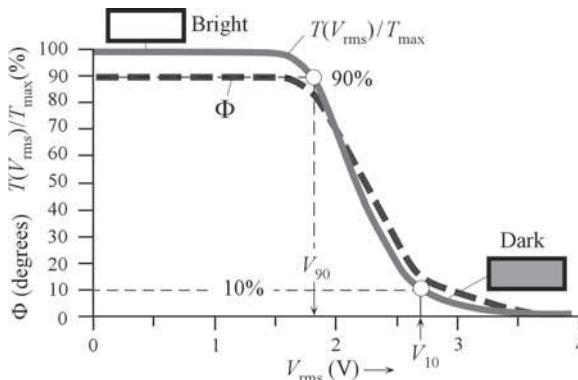


FIGURE 6.21 Plots of the rotation angle Φ of the linearly polarized light vs. the rms voltage V_{rms} across an LCD cell, and the normalized transmittance $T(V_{\text{rms}})/T_{\max}$ (%) vs. V_{rms} for a typical twisted nematic liquid crystal cell.

for LCDs are ~ 1 kHz. The reversal of the field does not change the principle of operation because molecules always try to align parallel to the field, which is along either $+z$ or $-z$. In both cases, optical field is not rotated and the through-light is extinguished at the second polarizer. The extent of transmission through an LCD depends on the rms value of the AC voltage. Manufacturers typically provide the transmittance vs. rms voltage characteristics of their LCDs. Figure 6.21 shows how the rotation angle Φ of the linearly polarized light through the liquid crystal medium depends on the rms voltage V_{rms} across an LCD cell. The normalized transmittance $T' = T(V_{\text{rms}})/T_{\max}$ is also shown as a function of V_{rms} . T_{\max} is the maximum transmittance under bright transmission conditions so that $T(V_{\text{rms}})/T_{\max}$ is 100% with no or very small V_{rms} .

It is apparent from Figure 6.21 that the rms voltage V_{rms} must reach a certain **threshold** value before any effect is seen. A threshold voltage is required to start untwisting the alignment of the mesogens. The rms voltage V_{90} corresponding to 90% normalized transmission T' is usually taken as the **threshold voltage**. The voltage at which T' has dropped to 10% defines the **saturation voltage**, V_{10} . LCD response times for turning on (alignment with the applied field) and off (alignment with the surface directors) depend on the properties of the LC, the thickness of the cell, and temperature. At room temperature these turn on and off times are typically in the millisecond time range, with the turn off time usually being longer than the on time. It is faster to align the molecules with the applied field than the time it takes for them to naturally align with the surface directors when the field is turned off.



The light from an LCD display is linearly polarized. A number of square polarizers have been placed on the screen of this laptop computer at different angles until the light is totally extinguished. There are five polarizers placed on the screen at different angles.

It should be apparent that the whole LCD operation is based on three important effects: the optical activity exhibited by the twisted nematic LC structure in which the “twisted mesogens” rotate the optical field; the ability to rotate or align the mesogens in the LC by a sufficiently large applied (external) field; and the use of two polarizers in converting the rotation of the optical field within the medium to an intensity variation.

6.6 ELECTRO-OPTIC EFFECTS

A. Definitions

Electro-optic effects refer to changes in the refractive index of a material induced by the application of an external electric field, which therefore *modulates* the optical properties; the applied field is not the electric field of any light wave, but a separate external field. We can apply such an external field by placing electrodes on opposite faces of a crystal and connecting these electrodes to a battery. The presence of such a field distorts the electron motions in the atoms or molecules of the substance, or distorts the crystal structure, resulting in changes in the optical properties. For example, an applied external field can cause an optically isotropic cubic crystal such as GaAs to become birefringent. In this case, the field induces principal axes and an optic axis. Typically, changes in the refractive index are small. The frequency of the applied field has to be such that the field appears static over the time scale it takes for the medium to change its properties, *i.e.*, respond, as well as for any light to cross the substance. The electro-optic effects are classified according to **first-** and **second-order effects**.

If we were to take the refractive index n to be a function of the applied electric field E , that is, $n = n(E)$, we can, of course, expand this as a Taylor series in E . The new refractive index n' is

Field-induced refractive index

$$n' = n + a_1 E + a_2 E^2 + \dots \quad (6.6.1)$$

where the coefficients a_1 and a_2 are called the **linear electro-optic effect** and **second-order electro-optic effect** coefficients. Although we would expect even higher terms in the expansion in Eq. (6.6.1), these are generally very small and their effects negligible within highest practical fields. The change in n due to the first E term is called the **Pockels effect**. The change in n due to the second E^2 term is called the **Kerr effect**,⁹ and the coefficient a_2 is generally written as λK where K is called the **Kerr coefficient**. Thus, the two effects are,

Pockels effect

$$\Delta n = a_1 E \quad (6.6.2)$$

and

Kerr effect

$$\Delta n = a_2 E^2 = (\lambda K) E^2 \quad (6.6.3)$$

All materials exhibit the Kerr effect. It may be thought that we will always find some (nonzero) values for a_1 for all materials but this is not true and only certain crystalline materials exhibit the Pockels effect. If we apply a field \mathbf{E} in one direction and then reverse the field and apply $-\mathbf{E}$, then according to Eq. (6.6.2), Δn should change sign. If the refractive index

⁹John Kerr (1824–1907) was a Scottish physicist and faculty member at Free Church Training College for Teachers (1857–1901), Glasgow, where he set up an optics laboratory and demonstrated the Kerr effect (1875).

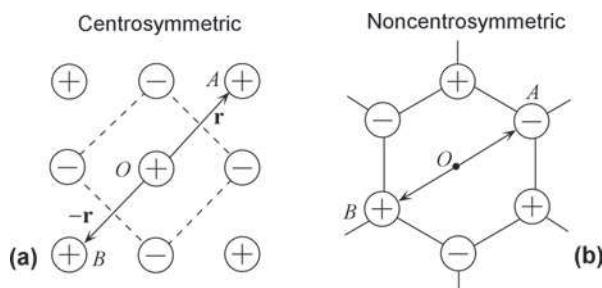


FIGURE 6.22 (a) A centrosymmetric unit cell (such as NaCl) has a center of symmetry at O . (b) An example of a noncentrosymmetric unit cell. In this example, the hexagonal unit cell has no center of symmetry.

increases for \mathbf{E} , it must decrease for $-\mathbf{E}$. Reversing the field should *not* lead to an identical effect (the same Δn). The structure has to respond differently to \mathbf{E} and $-\mathbf{E}$; there must therefore be some *asymmetry* in the structure to distinguish between \mathbf{E} and $-\mathbf{E}$. In a noncrystalline material, Δn for \mathbf{E} would be the same as Δn for $-\mathbf{E}$ as all directions are equivalent in terms of dielectric properties. Thus $a_1 = 0$ for all noncrystalline materials (such as glasses and liquids). Similarly, if the crystal structure has a center of symmetry, as shown in Figure 6.22 (a), then reversing the field direction has an identical effect and a_1 is again zero. A centrosymmetric crystal is such that if we draw a vector \mathbf{r} from the center O to an ion A , and then invert the vector making it $-\mathbf{r}$, we would then find the same type of ion at $-\mathbf{r}$, that is, A and B in Figure 6.22 (a) are identical ions. On the other hand, the hexagonal crystal in Figure 6.22 (b) is **noncentrosymmetric**. If we draw a vector \mathbf{r} from O to A , and then invert this, we would find a different ion B at $-\mathbf{r}$; A and B are different ions. Only noncentrosymmetric crystals exhibit the Pockels effect. For example, an NaCl crystal is centrosymmetric, and therefore exhibits no Pockels effect. On the other hand, a GaAs crystal is noncentrosymmetric and evinces the Pockels effect under an applied field.

B. Pockels Effect

The Pockels effect expressed in Eq. (6.6.2) is an oversimplification because in reality we have to consider the effect of an applied field along a particular crystal direction on the refractive index for light with a given propagation direction and polarization. For example, suppose that x , y , and z are the principal axes of a crystal with refractive indices n_1 , n_2 , and n_3 along these



Friedrich Carl Alwin Pockels (1865–1913), son of Captain Theodore Pockels and Alwine Becker, was born in Vincenza (Italy). He obtained his doctorate from Göttingen University in 1888. From 1900 until 1913 he was a professor of theoretical physics in the Faculty of Sciences and Mathematics at the University of Heidelberg, where he carried out extensive studies on electro-optic properties of crystals—the Pockels effect is the basis of many practical electro-optic modulators. (Courtesy of the Department of Physics and Astronomy, University of Heidelberg, Germany.)

Various KD*P-based Pockels cells used for Q-switching in laser applications. Left, half-wave voltage $V_{\lambda/2}$ of 6.8 kV; right, small, $V_{\lambda/2} = 6.5$ kV; far right large, $V_{\lambda/2} = 6.8$ kV. The cells have AR (antireflection coatings), and can be used from 0.4 to 1.1 μm . Note: KD*P is a KD_2PO_4 crystal. (Courtesy of Eksma Optics, Vilnius, Lithuania.)



directions. For an optically isotropic crystal, these would be the same, whereas for a uniaxial crystal, $n_1 = n_2 \neq n_3$ as illustrated in the xy cross-section in Figure 6.23 (a). Suppose that we suitably apply a voltage across a crystal and thereby apply an external DC field E_a along the z -axis. In Pockels effect, the field will modify the optical indicatrix. The exact effect depends on the crystal structure. For example, a crystal like GaAs, optically isotropic with a spherical indicatrix, becomes birefringent, and a crystal like KDP (KH_2PO_4 —potassium dihydrogen phosphate), *i.e.*, uniaxial, becomes biaxial. In the case of KDP, the field E_a along z rotates the principal axes by 45° about z , and changes the principal indices as indicated in Figure 6.23 (b). The new principal indices are now n'_1 and n'_2 , which means that the cross-section is now an ellipse. Propagation along the z -axis under an applied field in Figure 6.23 (b) now occurs with different refractive indices n'_1 and n'_2 . As apparent in Figure 6.23 (b), the applied field induces new principal axes x' , y' , and z' for this crystal, though in this special case $z' = z$. In the case of LiNbO_3 (lithium niobate), an optoelectronically important uniaxial crystal, a field E_a along the y -direction does not significantly rotate the principal axes but rather changes the principal refractive indices n_1 and n_2 (both equal to n_o) to n'_1 and n'_2 as illustrated in Figure 6.23 (c).

As an example, consider a wave propagating along the z -direction (optic axis) in an LiNbO_3 crystal. This wave will experience the same refractive index ($n_1 = n_2 = n_o$) whatever the polarization, as in Figure 6.23 (a). However, in the presence of an applied field E_a parallel to the principal y -axis as in Figure 6.23 (c), the light propagates as two orthogonally polarized waves (parallel to x and y) experiencing different refractive indices n'_1 and n'_2 . The applied field thus *induces a birefringence* for light traveling along the z -axis. The field-induced rotation of the principal axes in this case, though present, is small and can be neglected. Before the field E_a

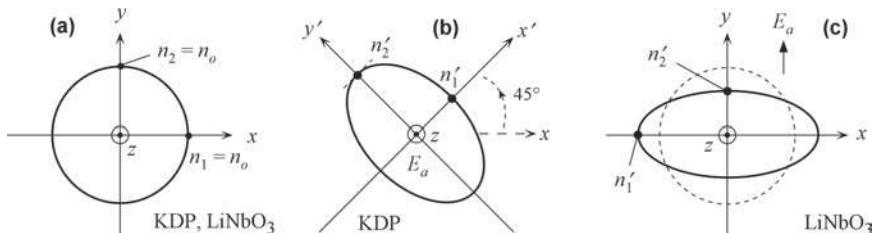


FIGURE 6.23 (a) Cross-section of the optical indicatrix with no applied field, $n_1 = n_2 = n_o$. (b) The applied external field E_a along z modifies the optical indicatrix. In a KDP crystal, it rotates the principal axes by 45° to x' and y' , and n_1 and n_2 change to n'_1 and n'_2 . (c) Applied field along y in LiNbO_3 modifies the indicatrix and changes n_1 and n_2 to n'_1 and n'_2 .

is applied, the refractive indices n_1 and n_2 are both equal to n_o . The Pockels effect then gives the new refractive indices n'_1 and n'_2 in the presence of E_a as

$$n'_1 \approx n_1 + \frac{1}{2}n_1^3 r_{22} E_a \quad \text{and} \quad n'_2 \approx n_2 - \frac{1}{2}n_2^3 r_{22} E_a \quad (6.6.4) \quad \begin{matrix} \text{Pockels} \\ \text{effect} \end{matrix}$$

where r_{22} is a constant, called a **Pockels coefficient**, that depends on the crystal structure and the material. The reason for the seemingly unusual subscript notation is that there are more than one constant and these are elements of a tensor that represents the optical response of the crystal to an applied field along a particular direction with respect to the principal axes (the exact theory is more mathematical than intuitive). We therefore have to use the correct Pockels coefficients for the refractive index changes for a given crystal and a given field direction.¹⁰ If the field were along z , the Pockels coefficient in Eq. (6.6.4) would be r_{13} .

It is clear that the control of the refractive index by an external applied field (and hence a voltage) is a distinct advantage that enables the phase change through a Pockels crystal to be controlled or modulated; such a **phase modulator** is called a **Pockels cell**. In the **longitudinal Pockels cell phase modulator**, the applied field is in the direction of light propagation, whereas in the **transverse phase modulator**, the applied field is transverse to the direction of light propagation. For light propagation along z , the longitudinal and transverse effects are illustrated in Figures 6.23 (b) and (c), respectively.

Consider the transverse phase modulator in Figure 6.24. In this example, the applied electric field, $E_a = V/d$, is applied parallel to the y -direction, normal to the direction of light propagation along z . Suppose that the incident beam is linearly polarized (shown as \mathbf{E}), say at 45° to the y -axis. We can represent the incident light in terms of polarizations \mathbf{E}_x and \mathbf{E}_y along the x - and y -axes. These components, \mathbf{E}_x and \mathbf{E}_y , experience refractive indices n'_1 and n'_2 , respectively. Thus when \mathbf{E}_x traverses the distance L , its phase changes by ϕ_1 ,

$$\phi_1 = \frac{2\pi n'_1}{\lambda} L = \frac{2\pi L}{\lambda} \left(n_o + \frac{1}{2} n_o^3 r_{22} \frac{V}{d} \right)$$

When the component \mathbf{E}_y traverses the distance L , its phase changes by ϕ_2 , given by a similar expression except that r_{22} changes sign. Thus, the phase change $\Delta\phi$ between the two field components is

$$\Delta\phi = \phi_1 - \phi_2 = \frac{2\pi}{\lambda} n_o^3 r_{22} \frac{L}{d} V \quad (6.6.5) \quad \begin{matrix} \text{Transverse} \\ \text{Pockels} \\ \text{effect} \end{matrix}$$

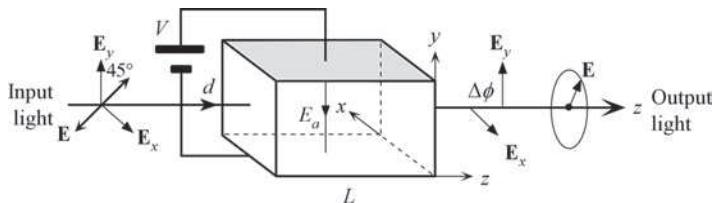


FIGURE 6.24 Transverse Pockels cell phase modulator. A linearly polarized input light into an electro-optic crystal emerges as a circularly polarized light.

¹⁰The reader should not be too concerned with the subscripts but simply interpret them as identifying the right Pockels coefficient value for the particular electro-optic problem at hand.



Electro-optic phase modulator using LiNbO_3 .

The socket is for the RF modulation input.

(Courtesy of Thorlabs.)

The applied voltage thus inserts an adjustable phase difference $\Delta\phi$ between the two field components. The polarization state of output wave can therefore be controlled by the applied voltage and the Pockels cell is a **polarization modulator**. We can change the medium from a quarter-wave to a half-wave plate by simply adjusting V . The voltage $V = V_{\lambda/2}$, the **half-wave voltage**, corresponds to $\Delta\phi = \pi$ and generates a half-wave plate. The advantage of the transverse Pockels effect is that we can independently reduce d , and thereby increase the field, and increase the crystal length L , to build up more phase change; $\Delta\phi$ is proportional to L/d . This is not the case in the longitudinal Pockels effect. If L and d were the same, typically $V_{\lambda/2}$ would be a few kilovolts but tailoring d/L to be much smaller than unity would bring $V_{\lambda/2}$ down to desirable practical values.

From the polarization modulator in Figure 6.24, we can build an **intensity modulator** by inserting a polarizer P and an analyzer A before and after the phase modulator as in Figure 6.25 (a) such that they are cross-polarized, *i.e.*, P and A have their transmission axes at 90° to each other. The transmission axis of P is at 45° to the y -axis (hence A also has its transmission axis at 45° to y) so that the light entering the crystal has equal \mathbf{E}_x and \mathbf{E}_y components. In the absence of an applied voltage, the two components travel with the same refractive index, and polarization output from the crystal is the same as its input. There is no light detected at the detector as A and P are at right angles ($\theta = 90^\circ$ in Malus's law).

An applied voltage inserts a phase difference $\Delta\phi$ between the two electric field components. The light leaving the crystal now has an elliptical polarization and hence a field component along

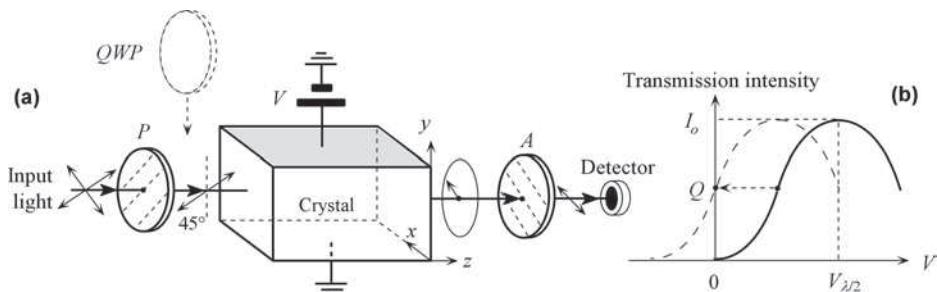


FIGURE 6.25 (a) A transverse Pockels cell intensity modulator. The polarizer P and analyzer A have their transmission axis at right angles and P polarizes at an angle 45° to y -axis. (b) Transmission intensity vs. applied voltage characteristics. If a quarter-wave plate (QWP) is inserted after P , the characteristic is shifted to the dashed curve.

the transmission axis of A . A portion of this light will therefore pass through A to the detector. The transmitted intensity now depends on the applied voltage V . The field components at the analyzer will be out of phase by an amount $\Delta\phi$. We have to find the total field \mathbf{E} and the component of this field along the transmission axis of A . Suppose that E_o is the amplitude of the wave incident on the crystal face. The amplitudes along x - and y -axes will be $E_o/\sqrt{2}$ each (notice that the x -component of \mathbf{E} is along the $-x$ -direction). The total field \mathbf{E} at the analyzer is

$$\mathbf{E} = -\hat{\mathbf{x}} \frac{E_o}{\sqrt{2}} \cos(\omega t) + \hat{\mathbf{y}} \frac{E_o}{\sqrt{2}} \cos(\omega t + \Delta\phi)$$

A factor $\cos(45^\circ)$ of each component passes through A . We can resolve \mathbf{E}_x and \mathbf{E}_y along A 's transmission axis, then add these components, and finally use a trigonometric identity to obtain the field emerging from A . The final result is

$$E = E_o \sin\left(\frac{1}{2}\Delta\phi\right) \sin\left(\omega t + \frac{1}{2}\Delta\phi\right)$$

The intensity I of the detected beam is the square of the term multiplying the sinusoidal time variation, *i.e.*,

$$I = I_o \sin^2\left(\frac{1}{2}\Delta\phi\right) \quad (6.6.6)$$

which can be written as

$$I = I_o \sin^2\left(\frac{\pi}{2} \cdot \frac{V}{V_{\lambda/2}}\right) \quad (6.6.7) \quad \begin{matrix} & \\ & \text{Pockels} \\ & \text{intensity} \\ & \text{modulator} \end{matrix}$$

where I_o is the light intensity under full transmission and $V_{\lambda/2}$ is the voltage that gives a phase change $\Delta\phi$ of π in Eq. (6.6.5).

An applied voltage of $V_{\lambda/2}$ is needed to allow full transmission. The transmission characteristics of this modulator is shown in Figure 6.25 (b). In digital electronics, we would switch a light pulse on and off so that the nonlinear dependence of transmission intensity on V in Eq. (6.6.7) would not be a problem. However, if we wish to obtain a linear modulation between the intensity I and V we need to bias this structure about the apparent “linear region” of the curve at half-height. This is done by inserting a quarter-wave plate (QWP) after the polarizer P , as in Figure 6.25 (a), which provides a circularly polarized light as input. The insertion of QWP means that $\Delta\phi$ is already shifted by $\pi/4$ before any applied voltage. The applied voltage then, depending on the sign, increases or decreases $\Delta\phi$. The new transmission characteristic is shown as a dashed curve in Figure 6.25 (b). We have, effectively, optically biased the modulator at point Q on the new characteristics. (Will it matter, if instead, we were to insert the quarter-wave plate before the analyzer¹¹?)

¹¹The answer is no, since it only serves to introduce a phase change of $\pi/4$ between P and A .

EXAMPLE 6.6.1 Pockels Cell Modulator

What should be the aspect ratio d/L for the transverse lithium niobate (LiNbO_3) phase modulator in Figure 6.24, that will operate at a free-space wavelength of $1.3 \mu\text{m}$ and will provide a phase shift $\Delta\phi$ of π (half wavelength) between the two field components propagating through the crystal for an applied voltage of 12 V ? At $\lambda = 1.3 \mu\text{m}$, LiNbO_3 has $n_o \approx 2.21$, $r_{22} \approx 5 \times 10^{-12} \text{ m V}^{-1}$.

Solution

From Eq. (6.6.5), putting $\Delta\phi = \pi$ for the phase difference between the field components \mathbf{E}_x and \mathbf{E}_y in Figure 6.24 gives

$$\Delta\phi = \frac{2\pi}{\lambda} n_o^3 r_{22} \frac{L}{d} V_{\lambda/2} = \pi$$

or

$$\frac{d}{L} = \frac{1}{\Delta\phi} \cdot \frac{2\pi}{\lambda} n_o^3 r_{22} V_{\lambda/2} \approx \frac{1}{\pi} \cdot \frac{2\pi}{(1.3 \times 10^{-6})} (2.21)^3 (5 \times 10^{-12})(12)$$

giving

$$d/L \approx 1 \times 10^{-3}$$

This particular transverse phase modulator has the field applied along the y -direction and light traveling along the z -direction as in Figure 6.24. If we were to use the transverse arrangement in which the field is applied along the z -axis, and the light travels along the y -axis, the relevant Pockels coefficients would be greater, and the corresponding aspect ratio d/L would be $\sim 10^{-2}$. We cannot arbitrarily set d/L to any ratio we like for the simple reason that when d becomes too small, the light will suffer diffraction effects that will prevent it from passing through the device. d/L ratios $10^{-3}\text{--}10^{-2}$ in practice can be implemented by fabricating an integrated optical device.

C. Kerr Effect

Suppose that we apply a strong electric field to an otherwise optically isotropic material such as glass (or liquid). The change in the refractive index will be due to the Kerr effect, a second-order effect. We can arbitrarily set the z -axis of a Cartesian coordinate system along the applied field as illustrated in Figure 6.26 (a). The applied field distorts the electron motions (orbits)

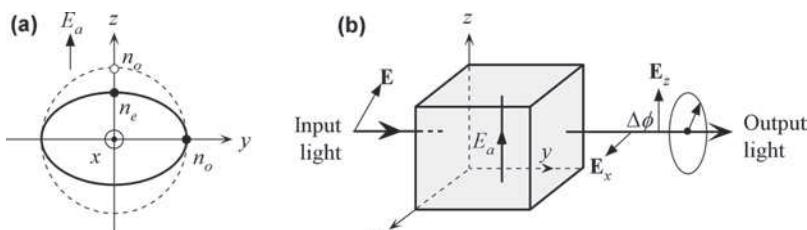


FIGURE 6.26 (a) An applied electric field, via the Kerr effect, induces birefringence in an otherwise optically isotropic material. (b) A Kerr cell phase modulator.

in the constituent atoms and molecules, including those valence electrons in covalent bonds, in such a way that it becomes more difficult for the electric field in the light wave to displace electrons parallel to the applied field direction. Thus, a light wave with a polarization parallel to the z -axis will experience a smaller refractive index, reduced from its original value n_o to n_e . Light waves with polarizations orthogonal to the z -axis will experience the same refractive index n_o . The applied field thus induces birefringence with an optic axis parallel to the applied field direction as shown in Figure 6.26 (a). The material becomes birefringent for waves traveling off the z -axis.

The polarization modulator and intensity modulator concepts based on the Pockels cell can be extended to the Kerr effect as illustrated in Figure 6.26 (b), which shows a Kerr cell phase modulator. In this case, the applied field again induces birefringence. The phase modulator in Figure 6.26 (b) uses the Kerr effect, whereupon the applied field E_a along z induces a refractive index n_e parallel to the z -axes, whereas that along the x -axis will still be n_o . The light components \mathbf{E}_x and \mathbf{E}_z then travel along the material with different velocities and emerge with a phase difference $\Delta\phi$ resulting in an elliptically polarized light. However, the Kerr effect is small as it is a second-order effect, and therefore only accessible for modulation use at high fields. The advantage, however, is that all materials, including glasses and liquids, exhibit the Kerr effect, and the response time in solids is very short, much less than nanoseconds leading to high modulation frequencies; greater than GHz.

If E_a is the applied field, then the change in the refractive index for polarization parallel to the applied field, as in Figure 6.26 (a), can be shown to be given by

$$\Delta n = \lambda K E_a^2 \quad (6.6.8) \text{ Kerr effect}$$

where K is the Kerr coefficient. We can now use Eq. (6.6.8) to find the induced phase difference $\Delta\phi$ and hence relate it to the applied voltage.

Table 6.2 summarizes the Pockels and Kerr coefficients for various materials. The Kerr effect also occurs in anisotropic crystals but the effect there cannot be simply characterized by a single coefficient K .

TABLE 6.2 Pockels (r) and Kerr (K) coefficients in a few selected materials.
Values in parentheses for r are at very high applied field frequencies

Material	Crystal	Indices	Pockels Coefficients ($\times 10^{-12} \text{ m V}^{-1}$)		$K(\text{m V}^{-2})$	Comment
			r_{13}	r_{33}		
LiNbO ₃	Uniaxial	$n_o = 2.286$ $n_e = 2.200$	$r_{13} = 9.6(8.6); r_{33} = 30.9(30.8)$ $r_{22} = 6.8(3.4); r_{51} = 32.6(28)$			$\lambda \approx 633 \text{ nm}$
KDP (KH ₂ PO ₄)	Uniaxial	$n_o = 1.512$ $n_e = 1.470$	$r_{41} = 8.8; r_{63} = 10.3$			$\lambda \approx 546 \text{ nm}$
KD*P (KD ₂ PO ₄)	Uniaxial	$n_o = 1.508$ $n_e = 1.468$	$r_{41} = 8.8; r_{63} = 26.8$			$\lambda \approx 546 \text{ nm}$
GaAs	Isotropic	$n_o = 3.43$	$r_{41} = 1.43$			$\lambda \approx 1.15 \mu\text{m}$
Glass	Isotropic	$n_o \approx 1.5$	0		$(1 - 20) \times 10^{-15}$	Typical values
Nitrobenzene	Isotropic	$n_o \approx 1.5$	0		3×10^{-12}	

(Source: Data selectively extracted from A. Yariv and P. Yeh, *Photonics*, 6th Edition (Oxford University Press, 2007), Ch. 9, and other sources.)

EXAMPLE 6.6.2 Kerr Effect Modulator

Suppose that we have a glass rectangular block of thickness (d) 100 μm and length (L) 20 mm and we wish to use the Kerr effect to implement a phase modulator in a fashion illustrated in Figure 6.26 (b). Note that d is the sample dimension along z and the applied field, and L is the sample length along y , the direction of light propagation. The input light has been polarized parallel to the applied field E_a direction, along the z -axis. Assume that the glass medium has a Kerr coefficient, $K = 3 \times 10^{-15} \text{ m V}^{-2}$. What is the applied voltage that induces a phase change of π (half-wavelength)?

Solution

The phase change $\Delta\phi$ for the optical field E_z is

$$\Delta\phi = \frac{2\pi\Delta n}{\lambda}L = \frac{2\pi(\lambda KE_a^2)}{\lambda}L = \frac{2\pi LKV^2}{d^2}$$

For $\Delta\phi = \pi$, $V = V_{\lambda/2}$,

$$V_{\lambda/2} = \frac{d}{\sqrt{2LK}} = \frac{(100 \times 10^{-6})}{\sqrt{2(20 \times 10^{-3})(3 \times 10^{-15})}} = 9.1 \text{ kV}$$

Although the Kerr effect is fast, it comes at a costly price. Notice that K depends on the wavelength and so does $V_{\lambda/2}$.

6.7 INTEGRATED OPTICAL MODULATORS

A. Phase and Polarization Modulation

Integrated optics refers to the integration of various optical devices and components on a single common substrate, for example lithium niobate, just as in integrated electronics all the necessary devices for a given function are integrated in the same semiconductor crystal substrate (chip). There is a distinct advantage to implementing various optically communicated devices, for example laser diodes, waveguides, splitters, modulators, and photodetectors on the same substrate, as it leads to miniaturization and also to an overall enhancement in performance and usability (typically).

One of the simplest examples is the polarization modulator shown in Figure 6.27 where an embedded waveguide has been fabricated by implanting an LiNbO_3 substrate with Ti atoms, which increase the refractive index. Two coplanar strip electrodes run along the waveguide and enable the application of a transverse field E_a to the light propagation direction z . The external modulating voltage $V(t)$ is applied between the coplanar drive electrodes and, by virtue of the Pockels effect, induces a change Δn in the refractive index and hence a voltage dependent phase shift through the device. We can represent light propagation along the guide in terms of two orthogonal modes, E_x along x and E_y along y . These two modes experience symmetrically opposite phase changes.¹² The phase shift $\Delta\phi$ between the E_x and E_y polarized waves would normally be given by Eq. (6.6.5). However, in this case the applied (or induced) field is not uniform between the electrodes and, further, not all applied field lines are inside the waveguide.

¹²These are called transverse electric (TE) and transverse magnetic (TM) modes.

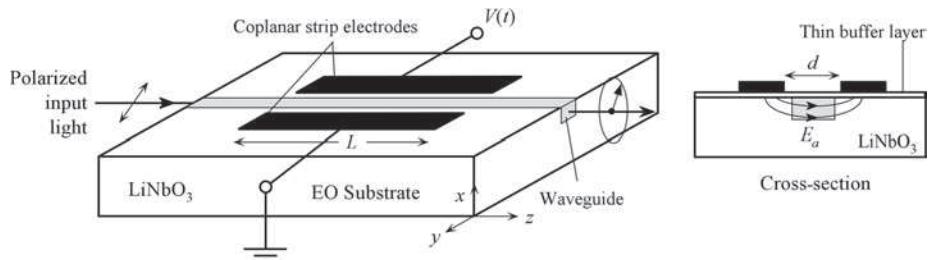


FIGURE 6.27 Integrated transverse Pockels cell phase modulator in which a waveguide is diffused into an electro-optic (EO) substrate. Coplanar strip electrodes apply a transverse field E_a through the waveguide. The substrate is an x -cut LiNbO_3 and typically there is a thin dielectric buffer layer (e.g., ~ 200 nm thick SiO_2) between the surface electrodes and the substrate to separate the electrodes away from the waveguide.

The useful electro-optic effect takes place over the spatial overlap region between the applied field and the optical fields. This **spatial overlap efficiency** is represented by a coefficient Γ , and the phase shift $\Delta\phi$ is written as

$$\Delta\phi = \Gamma \frac{2\pi}{\lambda} n_o^3 r_{22} \frac{L}{d} V \quad (6.7.1) \quad \text{Phase modulator}$$

where typically $\Gamma \approx 0.5\text{--}0.7$ for various integrated polarization modulators of this type. Since the phase shift depends on the product of V and L , a comparative device parameter would be the $V \times L$ product for a phase shift of π (half-wavelength), i.e., $V_{\lambda/2}L$. At $\lambda = 1.5 \mu\text{m}$ for an x -cut LiNbO_3 modulator as in Figure 6.27, with $d \approx 10 \mu\text{m}$, $V_{\lambda/2}L \approx 35 \text{ V cm}$. For example, a modulator with $L = 2 \text{ cm}$ has a half-wave voltage $V_{\lambda/2} = 17.5 \text{ V}$. By comparison, for a z -cut LiNbO_3 plate, that is for light propagation along the y -direction and E_a along z , the relevant Pockels coefficients (r_{13} and r_{33}) are much greater than r_{22} which leads to $V_{\lambda/2}L \approx 5 \text{ V cm}$.

B. Mach-Zehnder Modulator

Induced phase shift by applied voltage can be converted to an amplitude variation by using an interferometer, a device that interferes two waves of the same frequency but different phase. Consider the structure shown in Figure 6.28 which has implanted single-mode waveguides in a LiNbO_3 (or other electro-optic) substrate in the geometry shown. The waveguide at the input branches out at C to two arms A and B , and these arms are later combined at D to constitute the output.¹³ The splitting operation at C and combining the waves at D involve simple Y-junction waveguides. In the ideal case, the power is equally split at C so that the field is scaled by a factor $(2)^{1/2}$ going into each arm. The structure acts as an interferometer because the two waves traveling through the arms A and B interfere at the output port D , and the output amplitude depends on the phase difference (optical path difference) between the A and B -branches. Two back-to-back identical phase modulators enable the phase changes in A and B to be modulated. Notice that the applied field in branch A is in the opposite direction to that in branch B . The refractive index changes are therefore opposite, which means the phase changes in arms A and B are also opposite.

¹³For those who read Section 1.13 under Additional Topics, this is essentially the Mach-Zehnder interferometer explained in that section. The interferometer was named after Ludwig Mach and Ludwig Zehnder. (Ludwig Mach was the son of Ernst Mach.)

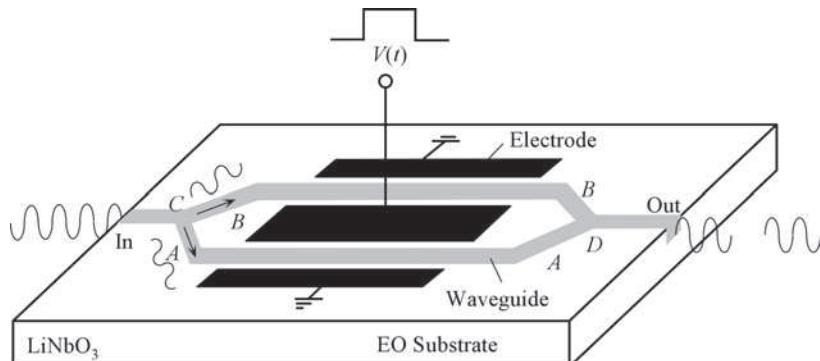


FIGURE 6.28 An integrated Mach–Zehnder optical intensity modulator. The input light is split into two coherent waves A and B, which are phase shifted by the applied voltage, and then the two are combined again at the output.

For example, if the applied voltage induces a phase change of $\pi/2$ in arm A, this will be $-\pi/2$ in arm B so that A and B would be out of phase by π . These two waves will then interfere destructively and cancel each other at D. The output intensity would be zero. Since the applied voltage controls the phase difference between the two interfering waves A and B at the output, this voltage also controls the output light intensity, though the relationship is not linear.

It is apparent that the relative phase difference between the two waves A and B is therefore doubled with respect to a phase change ϕ in a single arm. We can predict the output intensity by



An LiNbO_3 -based phase modulator for use from the visible spectrum to telecom wavelengths, with modulation speeds up to 5 GHz. This particular model has $V_{\lambda/2} = 10 \text{ V}$ at 1550 nm. (© JENOPTIK Optical System GmbH.)



An LiNbO_3 -based Mach–Zehnder amplitude modulator for use from the visible spectrum to telecom wavelengths, with modulation frequencies up to 5 GHz. This particular model has $V_{\lambda/2} = 5 \text{ V}$ at 1550 nm. (© JENOPTIK Optical System GmbH.)

adding waves A and B at D . If A is the amplitude of wave A and B (assumed equal power splitting at C), the optical field at the output is

$$E_{\text{out}} \propto A \cos(\omega t + \phi) + A \cos(\omega t - \phi) = 2A \cos \phi \cos(\omega t)$$

The output power P_{out} is proportional to E_{out}^2 , which is maximum when $\phi = 0$. Thus, at any phase difference ϕ , we have

$$\frac{P_{\text{out}}(\phi)}{P_{\text{out}}(0)} = \cos^2 \phi \quad (6.7.2)$$

Mach–Zehnder
amplitude
modulator

Although the derivation is oversimplified,¹⁴ it nonetheless represents approximately the right relationship between the power transfer and the induced phase ϕ change per modulating arm. The power transfer is zero when $\phi = \pi/2$ as expected. In practice, the Y-junction losses and uneven splitting result in less than ideal performance; A and B do not totally cancel out when $\phi = \pi/2$. Manufacturers often quote the extinction ratio, that is, the ratio between maximum and minimum power that can be transferred through the modulator by the application of a voltage $V(t)$ in Figure 6.28. Another quantity of interest that is quoted is the half-voltage $V_{\lambda/2}$. In the case of an amplitude modulator, this half-wave voltage is the voltage needed to induce a phase difference of π between the two arms A and B and hence extinguish the output. Thus, the half-wave voltage $V_{\lambda/2}$ is the voltage for switching the modulator from the on to the off state.

The output impedance of the driving circuit for the Mach–Zehnder modulator in Figure 6.28 has to be properly matched to the input impedance of the modulator to obtain the best frequency response. Commercially available high-end Mach–Zehnder modulators can handle digital modulation rates up to 40 Gb s⁻¹ and beyond. The half-wave voltages are typically less than 10 V, with extinction ratios in the 30–40 dB range.

C. Coupled Waveguide Modulators

When two parallel optical waveguides A and B are sufficiently close to each other, the electric fields associated with the propagation modes in A and B can overlap, as illustrated in Figure 6.29 (a). This implies that light can be coupled from one guide to another in a reminiscent way to frustrated total internal reflection (see Chapter 1). We can use qualitative arguments to understand the nature of light coupling between these two guides. Suppose that we launch a light wave into the guide A operating in single mode. Since the separation d of the two guides is small, some of the electric field in the evanescent wave of this mode extends into guide B , and therefore some electromagnetic energy will be transferred from guide A to B . This energy transfer will depend on the efficiency of coupling between the two guides and the nature of the modes in A and B , which in turn depend on the geometries and refractive indices of guides and the substrate (acting as a cladding).

As light in guide A travels along z , it leaks into B and, if the mode in B has the right phase, the transferred light waves build up along z as a propagating mode in B , as indicated in Figure 6.29 (b). By the same token, the light now traveling in B along z can be transferred back into A if the mode in A has the right phase. The efficient transfer of energy back and forth

¹⁴ See R. Syms and J. Cozens, *Optical Guided Waves and Devices* (McGraw-Hill, 1992), Ch. 9.

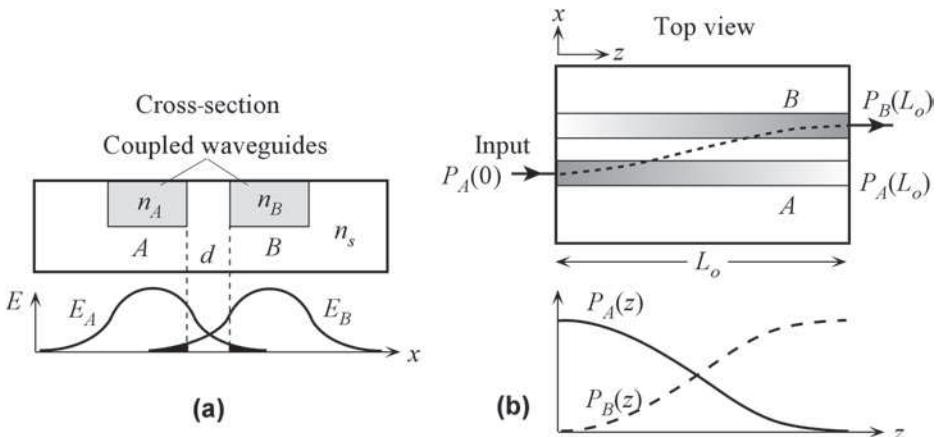


FIGURE 6.29 (a) Cross-section of two closely spaced waveguides A and B (separated by d) embedded in a substrate. The evanescent field from A extends into B and vice versa. Note: n_A and $n_B > n_s$ ($=$ substrate index). (b) Top view of the two guides A and B that are coupled along the z -direction. Light is fed into A at $z = 0$, and it is gradually transferred to B along z . At $z = L_o$, all the light has been transferred to B . Beyond this point, light begins to be transferred back to A in the same way.

between the two guides A and B requires that the two modes be in phase to allow the transferred amplitude to build up along z . If the two modes are out of phase, the waves transferred into a guide do not reinforce each other and the coupling efficiency is poor. Suppose that β_A and β_B are the propagation constants of the fundamental modes in A and B , then there is a phase mismatch per unit length along z , that is, $\Delta\beta = \beta_A - \beta_B$. The efficiency of energy transfer between the two guides depends on this phase mismatch. If the phase mismatch $\Delta\beta = 0$, then full transfer of power from A to B will require a certain coupling distance L_o , called the **transfer distance**, as shown in Figure 6.29 (b). This transfer distance depends on the efficiency of coupling C between the two guides A and B , which in turn depends on the refractive indices and geometries of the two guides. C depends on the extent of overlap of the mode fields E_A and E_B in Figure 6.29 (a). The transmission length L_o is inversely proportional to C (in fact, the theory shows that $L_o = \pi/C$).

In the presence of no mismatch, $\Delta\beta = 0$, full transmission would occur over the distance L_o . However, if there is a mismatch $\Delta\beta$, then the transferred power ratio over the distance L_o becomes a function of $\Delta\beta$. Thus, if $P_A(z)$ and $P_B(z)$ represent the light power in the guides A and B at z , then

$$\frac{P_B(L_o)}{P_A(0)} = f(\Delta\beta) \quad (6.7.3)$$

where f is a function designation. $P_B(L_o)/P_A(0)$ vs. phase mismatch $\Delta\beta$ is shown in Figure 6.30, which has its maximum when $\Delta\beta = 0$ (no mismatch) and then decays to zero at $\Delta\beta = \pi\sqrt{3}/L_o$. If we could induce a **phase mismatch** of $\Delta\beta = \pi\sqrt{3}/L_o$ by applying an electric field, which modulates the refractive indices of the guides, we could then prevent the transmission of light power from A to B . The light in A is then *not* transferred to B at the applied voltage that induces the required phase change.

Figure 6.31 shows an integrated directional coupler where two implanted symmetrical guides A and B are coupled over a transmission length L_o and also have electrodes placed on

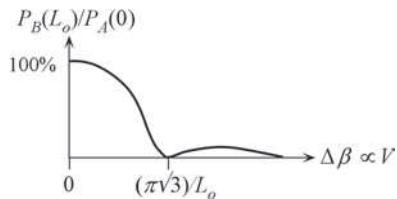


FIGURE 6.30 Transmission power ratio from guide A to guide B over the transmission length L_o as a function of mismatch $\Delta\beta$.

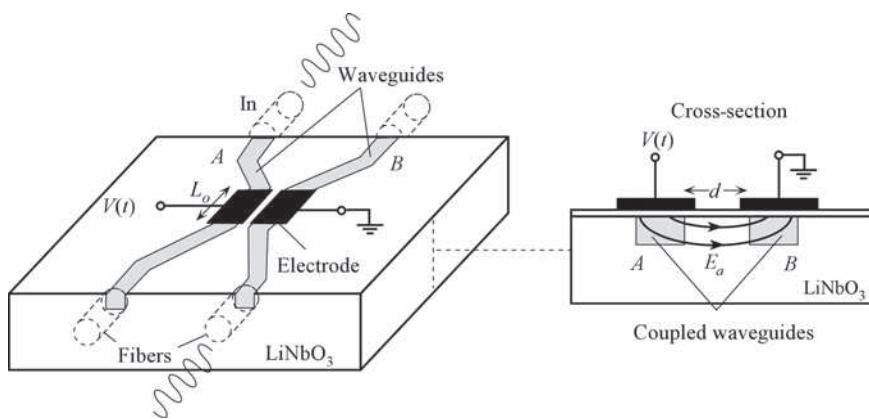


FIGURE 6.31 An integrated directional coupler. The applied field E_a alters the refractive indices of the two guides (A and B) and therefore changes the strength of coupling.

them. In the absence of an applied field, $\Delta\beta = 0$ (no mismatch) and there is a full transmission from guide A to B. If we apply a voltage between the electrodes, the two guides experience an applied field E_a in opposite directions, and hence experience opposite changes in their refractive indices. Suppose that $n (= n_A = n_B)$ is the refractive index of each guide, and Δn is the induced index change in each guide by the Pockels effect. The induced index difference Δn_{AB} between the guides is $2\Delta n$.

We can find Δn from the Pockels effect. The electric field, as a first approximation, can be written as $E_a \approx V/d$. However, we need to scale this field down because we need the effective field E_{eff} inside the crystal in the region of overlap with the optical mode that is propagating in guides A and B. This effective field E_{eff} is the field that modifies the refractive index, and can be written as $\Gamma V/d$, where Γ , called the **overlap factor**, is a factor that scales V/d to the right effective field for use in the Pockels effect; and represents the extent of overlap between the applied field and the optical mode that is propagating.¹⁵ Typical values for Γ are 0.5–0.8. Thus, the mismatch $\Delta\beta$ is

$$\Delta\beta = \Delta n_{AB} \left(\frac{2\pi}{\lambda} \right) \approx 2 \left(\frac{1}{2} n^3 r \Gamma \frac{V}{d} \right) \left(\frac{2\pi}{\lambda} \right) \quad (6.7.4)$$

Phase mismatch and voltage

¹⁵The calculation of Γ is actually quite difficult because the field is not uniform and we need to know what fraction of V/d appears in the optical mode region. Further, we need to average Pockels effect over this mode volume. In the present case, it will suffice to simply use a typical value for Γ .

where r is the appropriate Pockels coefficient. Setting $\Delta\beta = \pi\sqrt{3}/L_o$ for the prevention of transfer, the corresponding switching voltage V_o is

*Directional
coupler
switching
voltage*

$$V_o \approx \frac{\sqrt{3}\lambda d}{2\Gamma n^3 r L_o} \quad (6.7.5)$$

Since L_o depends inversely on the coupling efficiency C , for a given wavelength, V_o depends on the refractive indices, Pockels coefficient, and the geometry of the guides. Note that we simply used a magic factor Γ to represent the effective field that can be used in the Pockels effect to generate the right Δn that gives the right phase mismatch. The correct approach is to use coupled-wave theory, which is beyond the scope of this book. Equations (6.7.4) and (6.7.5) above are perfectly satisfactory for approximate calculations and clearly show the controlling factors.

EXAMPLE 6.7.1 Modulated Directional Coupler

Suppose that two optical guides embedded in a substrate such as LiNbO₃ are coupled as in Figure 6.31 to form a directional coupler, and the transmission length $L_o = 10$ mm. The coupling separation d is 10 μm, $\Gamma \approx 0.7$, the operating wavelength is 1.3 μm where Pockels coefficient $r \approx 10 \times 10^{-12}$ m V⁻¹, and $n \approx 2.20$. What is the switching voltage for this directional coupler?

Solution

Using Eq. (6.7.5), we can calculate the switching voltage

$$V_o \approx \frac{\sqrt{3}\lambda d}{2\Gamma n^3 r L_o} = \frac{\sqrt{3}(1.3 \times 10^{-6})(10 \times 10^{-6})}{2(0.7)(2.20)^3(10 \times 10^{-12})(10 \times 10^{-3})} = 15.1 \text{ V}$$

6.8 ACOUSTO-OPTIC MODULATOR

A. Photoelastic Effect and Principles

Experiments show that when we induce a strain (S) in a crystal, its refractive index n changes. This change in n with S is called the **photoelastic effect**. The strain changes the density of the crystal and distorts the bonds (and hence the electron orbits) which lead to a change in the refractive index n . If we were to examine the change in $1/n^2$ instead of n , then we would find that this is proportional to the induced strain S , and the proportionality constant is the **photoelastic coefficient** p , *i.e.*,

*Photoelastic
effect*

$$\Delta\left(\frac{1}{n^2}\right) = pS \quad (6.8.1)$$

The relationship is not as simple as stated in Eq. (6.8.1) since we must consider the effect of a strain S along one direction in the crystal on the induced change in n for a particular light propagation direction and some specific polarization. Equation (6.8.1) in reality is a tensor relationship.¹⁶

¹⁶As in the case of Pockels effect, we will only use the appropriate photoelastic coefficient, which would be given for the particular problem at hand.

The photoelastic effect in Eq. (6.8.1) provides a convenient means to implement **acousto-optic (AO) modulators** to modulate light. The basic principle is illustrated in Figure 6.32. We can easily generate acoustic or ultrasonic waves in a suitable AO crystal by attaching an ultrasonic transducer, that is, a piezoelectric transducer to one of the crystal faces. The ultrasonic transducer has a piezoelectric crystal with two electrodes, which are driven by an RF source. The piezoelectric crystal vibrates and generates acoustic waves, which are coupled into the AO crystal. The acoustic waves then propagate along the crystal, as visualized in Figure 6.32. These acoustic waves in the crystal propagate by rarefactions and compressions of the crystal, which lead to a periodic variation in the strain and hence a periodic variation in the refractive index in synchronization with the acoustic wave amplitude. Put differently, the periodic variation in the strain S leads to a periodic variation in n , owing to the photoelastic effect in Eq. (6.8.1).

The periodic variation in the refractive index works just like a diffraction grating even though it is generated by an acoustic wave, which is moving. If light is incident on this periodic variation of refractive index, it becomes diffracted. There is a transmitted, or through beam and one or many diffracted beams, depending on the mode of operation of the AO modulator. There are essentially two modes of operation. In the **Raman–Nath regime**¹⁷ of operation, as shown in Figure 6.33 (a), the width L of the diffraction grating is so small that the interaction of the incident light and the grating occurs almost along a line in the x -direction. The grating is a **thin grating**, and we can view this as the line grating we examined in Chapter 1. [See Figure 1.39 (a).] The thin grating diffracts the incident light into many diffracted beams as shown in Figure 6.33 (a).

On the other hand, in the **Bragg regime**, shown in Figure 6.33 (b), the width of the grating L is sufficiently large that we need to consider the interference of waves emanating from points along the width L or along z . The grating behaves like a **thick grating**, and the diffraction is not as straightforward as in the case of a single-line diffraction grating. For a given incident light wavelength λ and acoustic frequency f , or acoustic wavelength Λ in the crystal, there is a through (transmitted) beam and only one diffracted beam, corresponding to a first-order diffraction, if the angle of incidence is correct as shown in Figures 6.32 and 6.33 (b). Most practical

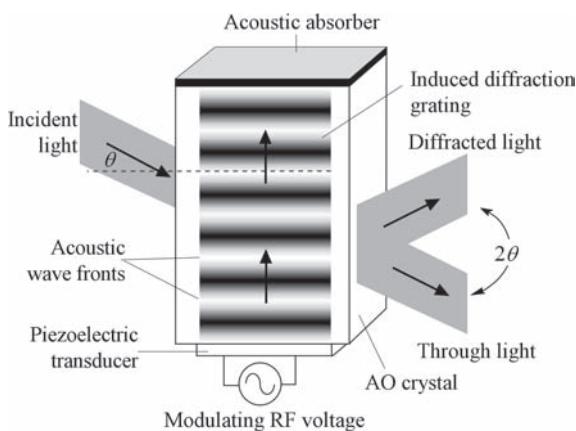


FIGURE 6.32 A schematic illustration of the principle of the acousto-optic modulator.

¹⁷ Also known as **Debye–Sears regime**.

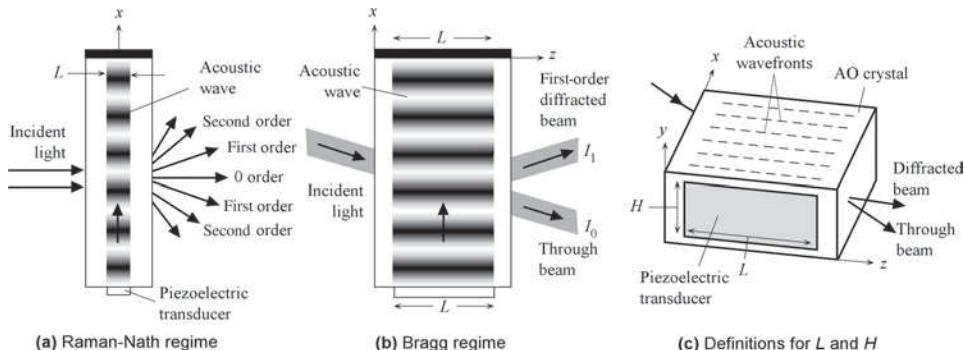


FIGURE 6.33 Illustration of (a) Raman–Nath and (b) Bragg regimes of operation for an acousto-optic modulator. In the Raman regime, the diffraction occurs as if it were occurring from a line grating. In the Bragg regime, there is a through beam and only one diffracted beam. (c) Definitions of L and H based on the transducer and the AO modulator geometry used.

modulators operate in the Bragg regime. In both Figure 6.33 (a) and (b), L essentially represents the **interaction length** of the incident light with the acoustic grating, and is often called the **equivalent acoustic beam length**. The condition for the Bragg regime is that the interaction length L should satisfy

Bragg
regime
condition

$$L \gg \frac{\Lambda^2}{\lambda} \quad (6.8.2)$$

in which Λ is the acoustic wavelength given by v_a/f , where v_a is the acoustic velocity and f is the acoustic frequency.

If $L \ll \Lambda^2/\lambda$, then we have the Raman–Nath regime. Suppose we take the light wavelength λ to be 10^{-6} m, and generate 100 MHz acoustic waves in a TeO_2 crystal (a common AO modulator crystal) in which $v_a = 4.2 \times 10^3$ m s $^{-1}$. The Λ^2/λ ratio in Eq. (6.8.2) is 1.8 mm. In most practical devices, L would be greater than this value and the AO modulator would be operating in the Bragg regime. In most cases, high operating acoustic frequencies (leading to shorter Λ) lead to the Bragg regime and low frequencies (longer Λ) to the Raman–Nath regime. The ultrasonic transducer attached to the AO crystal would have certain dimensions, which are represented as $L \times H$ as shown in Figure 6.33 (c), where L is the length and H is the height of the transducer along x and y , respectively. The acoustic waves generated by this transducer would have a cross-section $L \times H$. We assume that this acoustic wave does not significantly diverge as it travels along x . In addition, we need the height H of the beam to be wider than the width (or size) of the incident beam along y , a condition that is easily satisfied in practice.

B. Acousto-Optic Modulators

Consider the Bragg regime of operation. We can simplistically view the crystal region in which an acoustic wave is propagating as alternations in the refractive index from minimum n_{\min} to maximum n_{\max} as illustrated in Figure 6.34. We effectively have planes of layers with alternating indices n_{\min} and n_{\max} . The incident light interacts with these alternating planes of high and low refractive index. The acoustic wavelength Λ represents the separation of index boundaries. We can treat the incident optical beam as a stream of parallel coherent waves, two of which, A and B , are shown in Figure 6.34. These waves will be partially reflected at O and O_1 by the

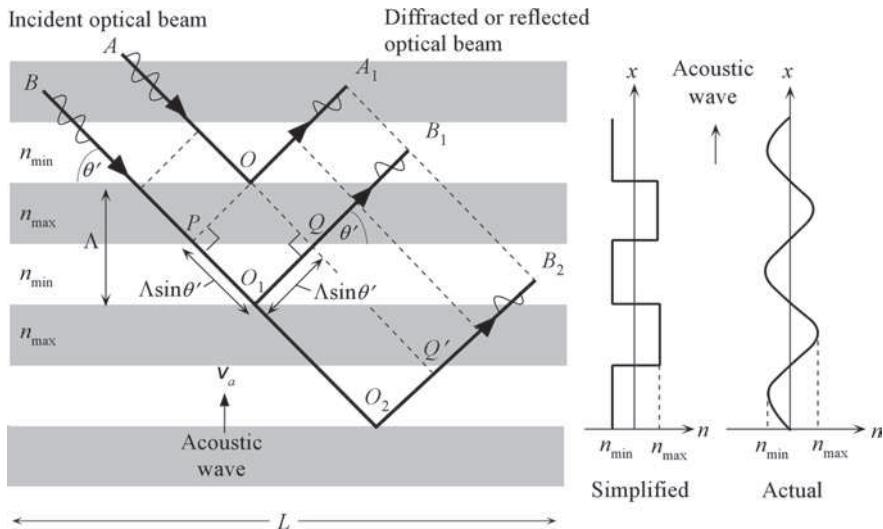


FIGURE 6.34 Consider two coherent optical waves A and B being reflected from two adjacent identical acoustic wavefronts to become A_1 and B_1 . These reflected waves can only constitute the diffracted beam if they are in phase. The angle θ' is exaggerated (typically, this is a few degrees).

index changes to become A_1 and B_1 . Further, as in ordinary reflection, we can take the angles of reflection and incidence to be the same. If the reflected waves A_1 and B_1 are in phase, then they would reinforce each other and thereby constitute a *diffracted beam*. Put differently, if there is a diffracted beam, then any two points on A_1 and B_1 , such as O and Q , must be on the same wavefront. Note also that the diffracted beam is essentially a *reflected beam*, though we will continue to use the term *diffracted beam*.

The points P and O on the incident beam (AB) are on the same wavefront. If there is a diffracted beam, then O and Q must also be on the same wavefront after reflection. The angle θ' represents the angle of the incident light with the acoustic wavefronts inside the crystal. The optical path difference from P to Q is PO_1Q , or $2\Delta \sin\theta'$. This must be a whole optical wavelength in the medium for the diffracted beam to exist, i.e., $2\Delta \sin\theta'$ must be λ/n , where λ is the free-space wavelength and n is the refractive index of the medium (normally taken as the unperturbed index). This condition is satisfied, and a diffracted beam is generated, only when the angle θ' satisfies

$$2\Delta \sin\theta' = \lambda/n; \quad \theta' = \theta'_B \quad (6.8.3)$$

The angle θ' that satisfies Eq. (6.8.3) is called the **Bragg angle** θ'_B . As mentioned below, *there is only one angle and only one diffracted beam*, which is called the **first-order diffracted beam**. Note that λ is much smaller than Λ , so that $\sin\theta'$ must be small, and we can use $\sin\theta' \approx \theta'$ in Eq. (6.8.3). (Clearly, the diffraction shown in Figure 6.34 is highly exaggerated.)

Since the interaction length L of the grating is large, we also need to consider the partial reflections from different planes. For example, the wave B will also experience partial reflections from similar boundaries, one of which is O_2 in Figure 6.34. We then need to examine the condition that will allow Q' on B_2 to be in phase with Q on B_1 if they are to be on the same wavefront. The examination of partial reflections from deeper planes in Figure 6.34 again leads to Eq. (6.8.3). The statement above that there is only one diffracted beam is not obvious in the

Bragg
condition
inside the
crystal

current discussion since any integer multiples of λ/n would also satisfy Eq. (6.8.2), and seemingly generate more diffraction angles. The existence of a single diffracted beam, which is a first-order beam, comes out if we carry out a proper treatment of the problem in which n varies sinusoidally.¹⁸ In deriving Eq. (6.8.3) we assumed that the incident optical wave and the acoustic wave are both plane waves as shown in Figure 6.34. In practice, both the optical and the acoustic beams have finite sizes and their wavefronts are not perfect plane waves. Further, the interaction distance L is not infinite. Consequently, the peak diffraction still occurs at $\theta = \theta_B$, but there are angles over a small spread around θ_B that can still yield some diffraction.

There is one further complication. In deriving Eq. (6.8.3) we assumed that the waves A and B were already inside the AO crystal. However, the light incident on the AO crystal is refracted into and then refracted out from the crystal. We are interested in the external angle of incidence and reflection θ shown in Figure 6.32. At the surface of the crystal, Snell's law for refraction relates the external angle of incidence θ and θ' as $\sin\theta = n\sin\theta'$ so that Eq. (6.8.3) becomes

Bragg condition with external angle

$$2\Lambda\sin\theta = \lambda; \quad \theta = \theta_B \quad (6.8.4)$$

The frequency of the diffracted light is not the same as that of the incident light. Suppose that ω is the angular frequency of the incident optical wave. The optical wave reflections occur from a moving diffraction pattern which moves with an acoustic velocity v_a as shown in Figure 6.34. As a result of the *Doppler effect*, the diffracted beam has either a slightly higher or slightly lower frequency depending on the direction of the traveling acoustic wave. If Ω is the frequency of the acoustic wave, then the diffracted beam has a Doppler-shifted frequency given by

Doppler shift

$$\omega' = \omega \pm \Omega \quad (6.8.5)$$

When the acoustic wave is traveling toward the incoming optical beam as in Figure 6.34, the diffracted optical beam frequency is up-shifted, *i.e.*, $\omega' = \omega + \Omega$. If the acoustic wave is traveling away from the incident optical beam, then the diffracted frequency is down-shifted, $\omega' = \omega - \Omega$. It is apparent that we can tune the frequency (wavelength) of the diffracted light beam by changing the frequency of the acoustic wave. (The diffraction angle is then also changed.)

Although the Bragg condition specifies the incidence angle θ_B for beam deflection, it does not say anything about the diffracted optical intensity. As shown in Figure 6.34, there are many partial reflections from numerous planes of refractive index change in the interaction region. The proper analysis is quite complicated, but the final result is straightforward. The intensity of the diffracted (reflected) first-order beam I_1 depends on the induced photoelastic change Δn , which depends on the amplitude and hence intensity of the acoustic waves. In the Bragg regime, the ratio of first-order diffracted (reflected) intensity I_1 to intensity I_i of incident beam can be viewed as the diffraction (deflection) efficiency η_{DE} of the AO modulator. η_{DE} increases with the power P_a in the acoustic wave and the interaction length L , and depends on the material properties of the AO crystal,¹⁹ *i.e.*,

Diffracted intensity

$$\eta_{DE} = \frac{I_1}{I_i} = \sin^2 \left[\frac{\pi}{\lambda} \left(\frac{L}{2H} M_2 P_a \right)^{1/2} \right] \quad (6.8.6)$$

¹⁸ See, for example, B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, 2nd Edition (Wiley, 2007), Ch. 19.

¹⁹ See, for example, I.-C. Chang, Ch. 6, "Acousto-Optic Modulators," in *The Handbook of Optics*, Vol. V, Ed. M. Bass *et al.*, (McGraw-Hill, 2010); or D. A. Pinnow, *IEEE J. Quantum Electron.*, QE6, 223, 1970.

where P_a is the power in the acoustic wave, L is the acoustic beam length, and H the acoustic beam height, shown in Figure 6.33 (c), and M_2 is a material dependent parameter that gauges the usefulness of the medium for AO modulation. M_2 is called the **AO figure of merit** for the medium, and is defined by

$$M_2 = \frac{n^6 p^2}{\rho v_a^3} \quad (6.8.7) \quad \text{AO figure of merit}$$

where ρ is the density, and p is the appropriate photoelastic coefficient. The n^6 in the numerator implies that the changes in n have the largest effect on the AO modulation efficiency. Table 6.3 summarizes the properties of a few selected useful AO materials. Notice how M_2 can vary by a factor of about 100 from the lowest to the highest.

It should be apparent from Eq. (6.8.6) that the input RF power in the acoustic wave modulates the diffracted beam intensity, I_1 . The through beam is also modulated since power is transferred from the zero-order to the first-order diffracted beam by the diffraction process. Thus, modulating the RF power applied to the piezoelectric transducer, we can modulate power in the acoustic wave and hence modulate both the through and the reflected beams. Figures 6.35 (a) and (b) illustrate how AO modulators can be used in analog and digital modulation. Note how both the diffracted (I_1) and the through beam (I_0) are modulated by the voltage applied to the transducer.

The modulation speed and hence the bandwidth of an AO are determined by the transit time of the acoustic waves across the light beam waist. If the optical beam has a waist (diameter) d in the modulator, then rise time (in digital modulation) is inversely proportional to d/v_a . Reducing d to increase the speed, however, results in the reduction of diffraction efficiency so there is a compromise between the speed of response and diffraction efficiency.

Acousto-optic modulators are used in a wide range of applications such as light deflectors, scanners, modulators, digital optical switches, frequency shifters, Q -switches, fiber attenuators, and various other applications. Although we attached a piezoelectric transducer to a crystal to generate acoustic waves, this is actually not necessary. If we use a piezoelectric crystal such as LiNbO₃, we can place periodic interdigital electrodes on the crystal surface at one end of

TABLE 6.3 Properties and figures of merit M_2 for various acousto-optic materials

Material	LiNbO ₃	TeO ₂	Ge	GaAs	GaP	PbMoO ₄	Fused silica	Ge ₃₃ Se ₅₅ As ₁₂ glass
Useful λ (μm)	0.6–4.5	0.4–5	2–20	1–11	0.6–10	0.4–1.2	0.2–4.5	1.0–14
ρ (g cm ⁻³)	4.64	6.0	5.33	5.34	4.13	6.95	2.2	4.4
n (at μm)	2.2 (0.633)	2.26 (0.633)	4 (10.6)	3.37 (1.15)	3.31 (1.15)	2.4 (0.633)	1.46 (0.633)	2.7
Maximum p (at 0.63 μm)	0.18 (p_{31})	0.34 (p_{13})	-0.07 ^a (p_{44})	-0.17 ^b (p_{11})	-0.151 (p_{11})	0.3 (p_{33})	0.27 (p_{12})	0.21 ^c (p_{11}, p_{12})
v_a (km s ⁻¹)	6.6	4.2	5.5	5.3	6.3	3.7	6	2.5
M_2 (s ³ kg ⁻¹) × 10 ⁻¹⁵	7	35	181	104	45	36	1.5	248

(Source: Extracted from I.-C. Chang, Ch. 6, "Acousto-Optic Modulators" in *The Handbook of Optics*, Vol. V, Ed. M. Bass *et al.*, McGraw-Hill, 2010.)

Notes: n is the refractive index, v_a is the acoustic velocity, and p is the maximum photoelastic coefficient; the exact p_{ij} is shown in parentheses. M_2 values must be multiplied by 10⁻¹⁵. ^a $\lambda = 2.0\text{--}2.2\text{ }\mu\text{m}$; ^b1.15 μm; ^c1.06 μm.

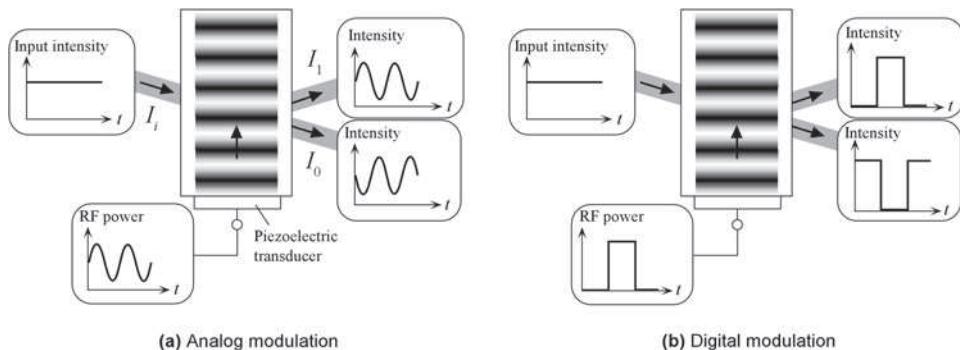


FIGURE 6.35 (a) Analog modulation of an AO modulator. I_i is the input intensity, I_0 is the zero-order diffraction, *i.e.*, the transmitted light, and I_1 is the first-order diffracted (reflected) light. (b) Digital modulation of an AO modulator.

the crystal, and then apply the RF voltage to these electrodes as shown in Figure 6.36. Notice that the voltage polarity changes between two neighboring electrodes. The period of the inter-digital electrodes (between two closest same-polarity electrodes) is chosen to be the same as the acoustic wave to be generated. The alternating voltage on the digital electrodes generates an alternating field inside the LiNbO_3 crystal near the surface. Due to the piezoelectric effect, the oscillating RF field generates oscillating strains, that is, lattice vibrations. These lattice vibrations propagate along the surface of the LiNbO_3 crystal, as shown in Figure 6.36, and are called **surface acoustic waves** (SAWs). They travel along the surface of the LiNbO_3 within a depth of about Λ , the acoustic wavelength, and give rise to the desired induced periodic variation in n through the photoelastic effect.

We need to increase the interaction of the light wave with the acoustic grating to obtain efficient diffraction. The surface region of LiNbO_3 is doped (*e.g.*, with Ti) to increase its refractive index and hence create a thin planar dielectric waveguide at the surface of the LiNbO_3 crystal. The light is fed into this thin waveguide where it is guided. This thin waveguide contains the SAW and hence the diffraction grating, from which the light becomes diffracted as illustrated in Figure 6.36.

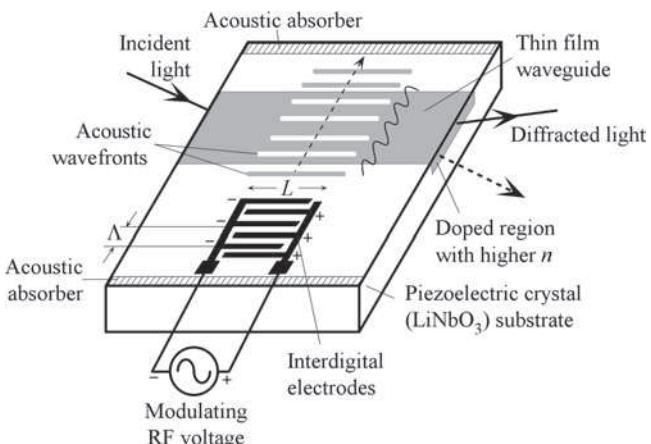


FIGURE 6.36 A simplified and schematic illustration of a surface acoustic wave (SAW)-based waveguide AO modulator. The polarity of the electrodes shown is at one instant, since the applied voltage is from an AC (RF) source.

EXAMPLE 6.8.1 AO Modulator

Suppose that we generate 150 MHz acoustic waves in a TeO₂ crystal. The RF transducer has a length (L) of 10 mm and a height (H) of 5 mm. Consider modulating a red-laser beam from a He-Ne laser, $\lambda = 632.8$ nm. Using the values in Table 6.3 for TeO₂, calculate the acoustic wavelength and hence the Bragg deflection angle. What is the Doppler shift in the wavelength? What is the relative intensity in the first-order deflected beam if the RF acoustic power is 1.0 W?

Solution

If f is the frequency of the acoustic waves, and v_a is the acoustic velocity, then the acoustic wavelength Λ is

$$\Lambda = \frac{v_a}{f} = \frac{(4.2 \times 10^3 \text{ m s}^{-1})}{(150 \times 10^6 \text{ s}^{-1})} = 2.8 \times 10^{-5} \text{ m} = 28 \mu\text{m}$$

Note that $\Lambda^2/\lambda = (2.8 \times 10^{-5} \text{ m})^2/(0.6328 \times 10^{-6} \text{ m}) = 1.2 \text{ mm}$. Since $L (= 10 \text{ mm})$ is much longer than 1.2 mm, we can assume we are in the Bragg regime.

The external Bragg angle θ is given by

$$\sin\theta = \frac{\lambda}{2\Lambda} = \frac{(632.8 \times 10^{-9} \text{ m})}{2(2.8 \times 10^{-5} \text{ m})} = 0.0113$$

so that $\theta = 0.65^\circ$ or a deflection angle 2θ of 1.3° . Note that we could have easily used $\sin\theta \approx \theta$.

The Doppler shift in frequency from Eq. (6.8.5) is simply the acoustic frequency 150 MHz.

The diffraction efficiency into the first order can be found from Eq. (6.8.6)

$$\eta_{DE} = \sin^2 \left[\frac{\pi}{(632.8 \times 10^{-9})} \left(\frac{10 \times 10^{-3}}{2(5 \times 10^{-3})} (35 \times 10^{-15})(1) \right)^{1/2} \right] = 0.64 \quad \text{or} \quad 64\%$$

where we have used $M_2 = 35 \times 10^{-15} \text{ s}^3 \text{ kg}^{-1}$ from Table 6.3.

6.9 FARADAY ROTATION AND OPTICAL ISOLATORS

When an optically inactive material such as glass is placed in a strong magnetic field and then a plane polarized light is sent *along* the direction of the magnetic field, it is found that the emerging light's plane of polarization has been rotated. This magneto-optic phenomenon is called the **Faraday effect**, also known as the **Faraday rotation**, as originally observed by Michael Faraday (1845). The magnetic field can be applied, for example, by inserting the material into the core of a magnetic coil—a solenoid. Strong permanent magnets are also used. The induced specific rotatory power, rotation per unit length, θ/L , has been found to be proportional to the magnitude of applied magnetic field, B . The amount of rotation θ is given by

$$\theta = VBL \quad (6.9.1) \quad \begin{matrix} & \text{Faraday} \\ & \text{effect} \end{matrix}$$

where B is the magnetic field (flux density), L is the length of the medium, and V is the so-called **Verdet constant** or **coefficient**.²⁰ It depends on the material and the wavelength.

²⁰The term *Verdet constant* is more common.

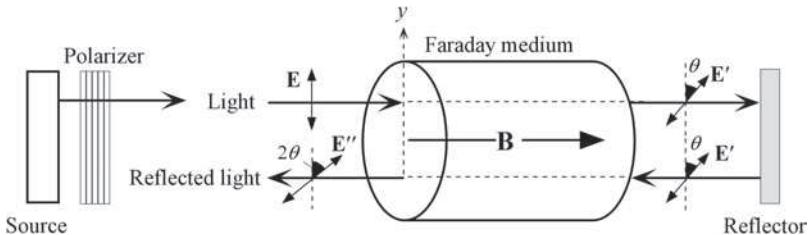


FIGURE 6.37 The sense of rotation of the optical field \mathbf{E} depends only on the direction of the magnetic field for a given medium (given Verdet constant). If light is reflected back into the Faraday medium, the field rotates a further θ in the same sense to come out as \mathbf{E}'' with a 2θ rotation with respect to \mathbf{E} .

It seems to appear that an “optical activity” has been induced by the application of a strong magnetic field to an otherwise optically inactive material. There is, however, an important distinction between the natural optical activity in Figure 6.17 and the Faraday effect. The sense of rotation θ in the Faraday effect, for a given material (Verdet constant), depends only on the direction of the magnetic field \mathbf{B} . If V is positive, for light propagating parallel to \mathbf{B} , the optical field \mathbf{E} rotates in the same sense as an advancing right-handed screw pointing in the direction of \mathbf{B} , as in Figure 6.37. The direction of light propagation, as in Figure 6.37, does not change the absolute sense of rotation of θ . If we reflect the wave to pass through the medium again, the rotation increases to 2θ .

The Verdet constant depends not only on the wavelength λ but also on the material dispersion, that is, n vs. λ , through

$$\text{Verdet constant} \quad V = -\frac{(e/m_e)}{2c} \lambda \frac{dn}{d\lambda} \quad (6.9.2)$$

We know that near a polarization resonance, n changes rapidly with λ , and the change, $dn/d\lambda$, becomes small as we move away from this resonance peak, e.g., as λ increases. Thus, the Verdet constant generally decreases with wavelength.

An **optical isolator** allows light to pass in one direction and not in the opposite direction. For example, the light source can be isolated from various reflections by placing a polarizer and a Faraday rotator that rotates the field by 45° as in Figure 6.37. The reflected light will have a $2\theta = 90^\circ$ rotation and will not pass through the polarizer back to the source. The magnetic field is typically applied by enclosing the Faraday medium in a rare-earth magnet ring.

EXAMPLE 6.9.1 Faraday rotation

Suppose we pass a polarized beam at 633 nm (from a He-Ne laser) through a 5 cm long SF57 dense flint glass rod. If the magnetic field along the rod is 0.7 T, what is the rotation of the optical field?

Solution

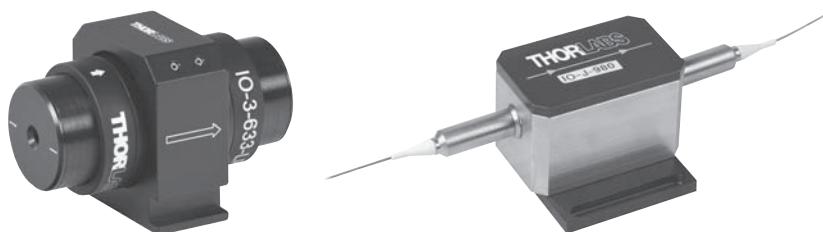
From Table 6.4, at 633 nm, $V = 20 \text{ rad T}^{-1} \text{ m}^{-1}$ so that the rotation is

$$\theta = VBL = (20 \text{ rad B}^{-1} \text{ m}^{-1})(0.7 \text{ T})(0.05 \text{ m}) = 0.70 \text{ rad} \quad \text{or} \quad 40^\circ$$

TABLE 6.4 Verdet constants of some selected materials at various wavelengths

Material and the wavelength	Quartz 589 nm	Tb ₃ Ga ₅ O ₁₂ 633 nm	ZnS 589 nm	ZnTe 633 nm	NaCl 589 nm	Crown glasses 633 nm	Dense flint glass (SF57) 633 nm
$V(\text{rad m}^{-1} \text{T}^{-1})$	4.0	-134	65.8	188	10	4–6	20

(Source: Data from various sources, including, M.J. Weber, *The Handbook of Optical Materials*, CRC Press, 2003.)



Optical isolators. Left: Free-space optical isolator for use at 633 nm up to 3 W. Right: Fiber-coupled optical isolator for use at 980 nm up to 2 W of optical power. (Courtesy of Thorlabs.)

6.10 NONLINEAR OPTICS AND SECOND HARMONIC GENERATION

The application of an electric field E to a dielectric material causes the constituent atoms and molecules to become polarized. The medium responds to the field E by developing a polarization P which represents the net induced dipole moment per unit volume. In a linear dielectric medium, the induced polarization P is proportional to the electric field E at that point and the two are related by $P = \epsilon_0\chi E$, where χ is the electric susceptibility. However, this linearity breaks down at high fields and the P vs. E behavior deviates from the linear relationship as shown in Figure 6.38 (a). P becomes a function of E which means that we can expand it in terms of increasing powers of E . It is customary to represent the induced polarization as

$$P = \epsilon_0\chi_1 E + \epsilon_0\chi_2 E^2 + \epsilon_0\chi_3 E^3 \quad (6.10.1) \quad \text{Induced polarization}$$

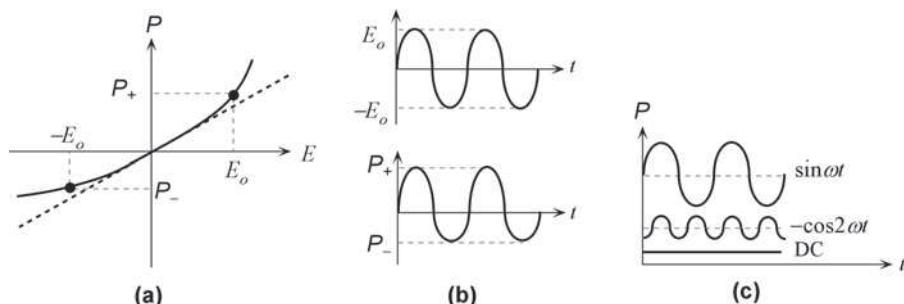


FIGURE 6.38 (a) Induced polarization vs. optical field for a nonlinear medium. (b) Sinusoidal optical field oscillations between $\pm E_o$ result in polarization oscillations between P_+ and P_- . (c) The polarization oscillation can be represented by sinusoidal oscillations at angular frequencies ω (fundamental), 2ω (second harmonic), and a small DC component.

where χ_1 , χ_2 , and χ_3 are the **linear**, **second-order**, and **third-order susceptibilities**, respectively. The coefficients decrease rapidly for higher terms and are not shown in Eq. (6.10.1). The importance of the second and third terms, *i.e.*, nonlinear effects, depends on the field strength E . Nonlinear effects begin to become observable when fields are very large, *e.g.*, $\sim 10^7 \text{ V m}^{-1}$. Such high fields require light intensities ($\sim 1000 \text{ kW cm}^{-2}$) that invariably require lasers. All materials, whether crystalline or noncrystalline, possess a finite χ_3 coefficient. However, only certain classes of crystals have a finite χ_2 , and the reason is the same as for the observation of the Pockels effect.²¹ Only those crystals, such as quartz, that have no center of symmetry have nonzero χ_2 coefficient; these crystals are also piezoelectric. One of the most important consequences of the nonlinear effect is the **second harmonic generation (SHG)**,²² when an intense light beam of frequency²³ ω passing through an appropriate crystal (*e.g.*, quartz) generates a light beam of double the frequency, 2ω . SHG is based on a finite χ_2 coefficient in which the effect of χ_3 is negligible.

Consider a beam of monochromatic light, with a well-defined angular frequency ω , passing through a medium. The optical field E at any point in the medium will polarize the medium at the same point in synchronization with the optical field oscillations. An oscillating dipole moment is well known as an electromagnetic emission source (just like an antenna). These secondary electromagnetic emissions from the dipoles in the medium interfere and constitute the actual wave traveling through the medium (Huygen's construction from secondary waves). Suppose that the optical field is oscillating sinusoidally between $\pm E_o$ as shown in Figure 6.38 (b). In the linear regime, *i.e.*, under small fields, P oscillations will also be sinusoidal with a frequency ω .

If the field strength is sufficiently large, the induced polarization will not be linear as indicated in Figure 6.38 (a), and will not oscillate in a simple sinusoidal fashion as shown in Figure 6.38 (b). The polarization now oscillates between P_+ and P_- and is not symmetrical. The oscillations of the dipole moment P now emit waves not only at the frequency ω but also at 2ω . In addition there is a DC component (light is said to be rectified, *i.e.*, it gives rise to a small permanent polarization). The *fundamental* ω and the *second harmonic*, 2ω , components, along with the DC, are shown in Figure 6.38 (c).

If we write the optical field as $E = E_o \sin(\omega t)$ and substitute this into Eq. (6.10.1), after some trigonometric manipulation,²⁴ and neglecting χ_3 terms, we would find the induced P as

$$P = \varepsilon_o \chi_1 E_o \sin(\omega t) - \frac{1}{2} \varepsilon_o \chi_2 E_o \cos(2\omega t) + \frac{1}{2} \varepsilon_o \chi_2 E_o \quad (6.10.2)$$

The first term is the fundamental, the second is the second harmonic, and the third is the DC term (permanent polarization in the medium) as summarized in Figure 6.38 (c). The second harmonic (2ω) oscillation of local dipole moments generates secondary second harmonic (2ω) waves in the crystal. It may be thought that these secondary waves will interfere constructively and result in a second harmonic beam just as the fundamental (ω) secondary waves interfere and give rise to the propagating light beam. However, the crystal will normally possess

²¹ See Section 6.6 B on the Pockels Effect.

²² Second harmonic generation was first experimentally demonstrated by Peter A. Franken and coworkers at the University of Michigan (1961), where they focused a 3 kW pulse of red light (694.3 nm) from a ruby laser onto a quartz crystal and observed some ultraviolet light (347.15 nm) coming out. The conversion efficiency was 1×10^{-8} [Eugene Hecht, *Optics*, 2nd Edition (Addison Wesley, 1987), p. 612].

²³ In this section, for brevity, the adjective *angular* is dropped from the frequency, though implied.

²⁴ That is, $\sin^2(\omega t) = \frac{1}{2}[1 - \cos(2\omega t)]$.

different refractive indices $n(\omega)$ and $n(2\omega)$ for frequencies ω and 2ω , which means that the ω - and 2ω -waves propagate with different phase velocities v_1 and v_2 , respectively. As the ω -wave propagates in the crystal, it generates secondary 2ω -waves along its path, indicated as S_1, S_2, S_3, \dots in Figure 6.39. When wave S_2 is generated, S_1 must arrive there in phase, which means S_2 must travel with the same velocity as the fundamental wave, and so on as in Figure 6.39. It is apparent that only if these 2ω -waves are in phase, that is, they propagate with the same velocity as the ω -wave, they can interfere constructively and constitute a second harmonic beam. Otherwise, the S_1, S_2 , and S_3 will eventually fall out of phase and destroy each other, and there will be either no or very little second harmonic beam. The condition that the second harmonic waves must travel with the same phase velocity as the fundamental wave to constitute a second harmonic beam is called **phase matching** and requires $n(\omega) = n(2\omega)$. For most crystals, this is not possible as n is dispersive; that is, it depends on the wavelength.

SHG efficiency depends on the extent of phase matching, $n(\omega) = n(2\omega)$. One method is to use a birefringent crystal as these have two refractive indices: ordinary index n_o and extraordinary index n_e . Suppose that along a certain crystal direction at an angle θ to the optic axis, $n_e(2\omega)$ at the second harmonic is the same as $n_o(\omega)$ at the fundamental frequency: $n_e(2\omega) = n_o(\omega)$. This is called **index matching** and the angle θ is the **phase matching angle**. Thus, the fundamental would propagate as an ordinary wave and the second harmonic as an extraordinary wave and both would be in phase. This would maximize the conversion efficiency, though this would still be limited by the magnitude of the second term with respect to the first in Eq. (6.10.1). To separate the second harmonic beam from the fundamental beam, something like a diffraction grating, a prism or an optical filter will have to be used at the output as illustrated in Figure 6.40 (a). The phase matching angle θ depends on the wavelength (or ω) and is sensitive to temperature.

Most green lasers, including many green laser pointers, use SHG in a similar way to that illustrated in Figure 6.40 (a). A diode-pumped Nd³⁺-ion solid-state laser (e.g., Nd³⁺:YVO₄ or Nd³⁺:YAG) produces a CW lasing emission at 1064 nm. This is a diode-pumped solid-state laser, and needs a laser diode emitting at 808 nm for pumping as shown in Figure 6.40 (b). The 1064 nm laser radiation from the Nd³⁺:YVO₄ is fed into the KTP crystal placed next to it, and is often optically cemented to it. The 1064 nm laser intensity is sufficiently high to allow the SHG in KDP to produce a green radiation output beam. Most importantly, the YVO₄ and KTP crystals are both placed *within* the optical cavity of the laser that is producing the ω -radiation (the 1064 nm radiation), as shown in Figure 6.40 (b). Within the optical cavity, the ω -radiation has high intensity, which enhances SHG, a distinct advantage. The cemented YVO₄ and KTP crystals are typically a few millimeters in total thickness and come with the end-mirrors coated.

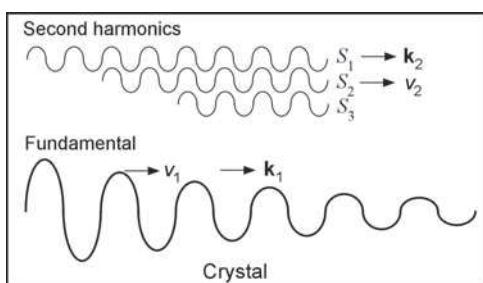


FIGURE 6.39 As the fundamental wave propagates, it periodically generates second harmonic waves (S_1, S_2, S_3, \dots) and if these are in phase then the amplitude of the second harmonic light builds up.

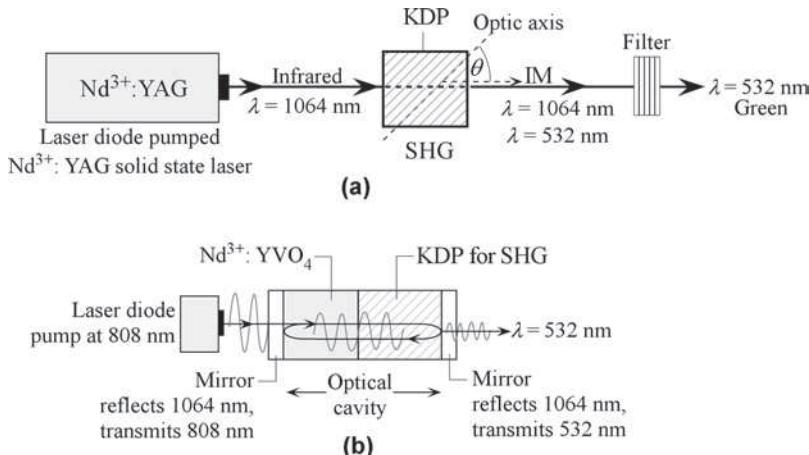


FIGURE 6.40 (a) A simplified schematic illustration of optical frequency doubling using a KDP (potassium dihydrogen phosphate) crystal. IM is the index-matched direction at an angle θ (about 35°) to the optic axis. Along the IM direction, $n_e(2\omega) = n_o(\omega)$. The focusing of the laser beam onto the KDP crystal and the collimation of the light emerging from the crystal are not shown. (b) Typical SHG-based green laser principle. The KTP crystal is next to the $\text{Nd}^{3+}\text{:YVO}_4$ crystal and both inside the laser optical cavity. The end mirrors reflect 1064 nm radiation and hence allow the 1064 nm radiation to build-up in the cavity.

It is instructive to consider the SHG process in terms of photon interactions inside the crystal, as shown in Figure 6.41. Two fundamental mode photons in the incident beam interact with the dipole moments in the crystal to produce a single second harmonic photon. From modern physics we know that the photon momentum is $\hbar\mathbf{k}$ and its energy is $\hbar\omega$. Suppose that subscripts 1 and 2 refer to fundamental and second harmonic photons. In general terms, we can write the following two equations. Conservation of momentum requires that

$$\hbar\mathbf{k}_1 + \hbar\mathbf{k}_1 = \hbar\mathbf{k}_2 \quad (6.10.3)$$

The conservation of energy requires that

$$\hbar\omega_1 + \hbar\omega_1 = \hbar\omega_2 \quad (6.10.4)$$

We tacitly assumed that the interaction does not result in phonon (lattice vibration) generation or absorption. We can satisfy Eq. (6.10.4) by taking $\omega_2 = 2\omega_1$, so that the frequency of the second harmonic is indeed twice the fundamental. To satisfy Eq. (6.10.3) we need $\mathbf{k}_2 = 2\mathbf{k}_1$. The phase velocity v_2 of the second harmonic waves is

$$v_2 = \frac{\omega_2}{k_2} = \frac{2\omega_1}{2k_1} = \frac{\omega_1}{k_1} = v_1$$

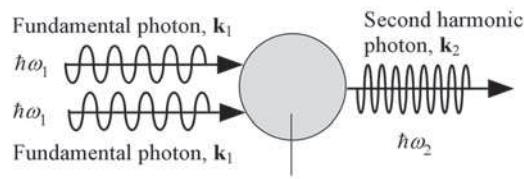


FIGURE 6.41 Photonic interpretation of second harmonic generation.

Dipole moment and photon interaction region

Thus, the fundamental and second harmonic photons are required to have the same phase velocity, which is tantamount to the phase matching criterion developed above in terms of pure waves. If \mathbf{k}_2 is not exactly $2\mathbf{k}_1$, *i.e.*, $\Delta k = k_2 - 2k_1$ is not zero (there is a mismatch), then SHG is only effective over a limited length l_c , which can be shown to be given by $l_c = \pi/\Delta k$. This length l_c is essentially the *coherence length* of the second harmonic; depending on the index difference, this may be quite short, *e.g.*, $l_c \approx 1 - 100 \mu\text{m}$. If the crystal size is longer than this, the second harmonics will interfere randomly with each other and the SHG efficiency will be very poor, if not zero. Phase matching is therefore an essential requirement for SHG. The conversion efficiency depends on the intensity of exciting laser beam, the χ_2 coefficient of the material, and the extent of phase matching and can be substantial (as high as 70–80%) if well-engineered, for example, by placing the converting crystal into the cavity of the laser itself.

Additional Topics

6.11 JONES VECTORS

We can represent the state of polarization of a light wave by using a matrix, called a **Jones matrix**²⁵ (or vector). Various operations on the polarization state of light then correspond to multiplying this matrix with another matrix that represents the optical operation. Consider a light wave traveling along z with field components E_x and E_y along x and y . These components are orthogonal and, in general, would be of different magnitudes and have a phase difference ϕ between them. If we use the exponential notation then

$$E_x = E_{xo} \exp[j(\omega t - kz + \phi_x)] \quad \text{and} \quad E_y = E_{yo} \exp[j(\omega t - kz + \phi_y)]$$

Jones matrix is a column matrix whose elements are E_x and E_y without the common $\exp[j(\omega t - kz)]$ factor

$$\mathbf{E} = \begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} E_{xo} \exp(j\phi_x) \\ E_{yo} \exp(j\phi_y) \end{bmatrix} \quad (6.11.1)$$

Usually Eq. (6.11.1) is normalized by dividing by the total amplitude $E_o = (E_{xo}^2 + E_{yo}^2)^{1/2}$. We can factor out $\exp(j\phi_x)$ to further simplify the matrix to obtain the **Jones matrix**

$$\mathbf{J} = \frac{1}{E_o} \begin{bmatrix} E_{xo} \\ E_{yo} \exp(j\phi) \end{bmatrix} \quad (6.11.2)$$

where $\phi = \phi_y - \phi_x$. For example, a light wave that is linearly polarized with the field at 45° to the x -axis can be conveniently written as

$$\mathbf{J} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

²⁵R. Clark Jones introduced these matrices circa 1941.

TABLE 6.5 Examples of Jones vectors

Jones vector \mathbf{J}_{in}	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}$	$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ j \end{bmatrix}$	$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -j \end{bmatrix}$
Polarization	Linear; horizontal \mathbf{E}	Linear; \mathbf{E} at 45° to x -axis	Linear; \mathbf{E} at θ to x -axis	Right circularly polarized	Left circularly polarized
Transmission matrix \mathbf{T}	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} e^{j\phi} & 0 \\ 0 & e^{j\phi} \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & j \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & e^{-j\Gamma} \end{bmatrix}$
Optical operation	Linear polarizer; horizontal transmission axis	Isotropic phase changer	Quarter-wave plate	Half-wave plate	Wave retarder; fast axis along x

Table 6.5 shows Jones vectors for various polarizations. Note that right and left circularly polarized waves have matrices with complex numbers. In this case $\pm j$ for the bottom matrix element, by virtue of Eq. (6.11.2), implies that E_y is $\pm 90^\circ$ out of phase with E_x , which results in a circularly polarized light.

If we pass a wave of given Jones vector \mathbf{J}_{in} through an optical device, then this operation would be multiplying \mathbf{J}_{in} by the **transmission matrix \mathbf{T}** of the device. If \mathbf{J}_{out} is the Jones vector for the output light through the device, then

$$\mathbf{J}_{out} = \mathbf{T} \mathbf{J}_{in} \quad (6.11.3)$$

For example, the transmission matrix for **quarter-wave operation** would be

$$\mathbf{T} = \begin{bmatrix} 1 & 0 \\ 0 & j \end{bmatrix}$$

The output will be

$$\mathbf{J}_{out} = \mathbf{T} \mathbf{J}_{in} = \begin{bmatrix} 1 & 0 \\ 0 & j \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ j \end{bmatrix}$$

which represents a wave that has E_y and E_x out of phase by $\pi/2$, and hence corresponds to a right circularly polarized light as given in Table 6.5. We can represent a chain of optical manipulations by multiplying Jones matrices and eventually find the output light. Various optical operations are also shown in Table 6.5.

Questions and Problems

6.1 Polarization Suppose that we write the E_y and E_x components of a light wave generally as

$$E_x = E_{xo} \cos(\omega t - kz) \quad \text{and} \quad E_y = E_{yo} \cos(\omega t - kz + \phi)$$

Show that at any instant E_x and E_y satisfy the ellipse equation on the E_y vs. E_x coordinate system:

$$\left(\frac{E_x}{E_{xo}}\right)^2 + \left(\frac{E_y}{E_{yo}}\right)^2 - 2\left(\frac{E_x}{E_{xo}}\right)\left(\frac{E_y}{E_{yo}}\right) \cos\phi = \sin^2\phi$$

Sketch schematically what this ellipse looks like assuming $E_{xo} = 2E_{yo}$. When would this ellipse form (a) an ellipse with its major axis on the x -axis, (b) a linearly polarized light at 45° , (c) right and left circularly polarized light?

- 6.2 Linear and circular polarization** Show that a linearly polarized light wave can be represented by two circularly polarized light waves with opposite rotations. Consider the simplest case of a wave linearly polarized along the y -axis. What is your conclusion?

- 6.3 Wire-grid polarizer** Figure 6.42 shows a wire-grid polarizer which consists of closely spaced parallel thin conducting wires. The light beam passing through the wire grid is observed to be linearly polarized at right angles to the wires. Can you explain the operation of this polarizer?

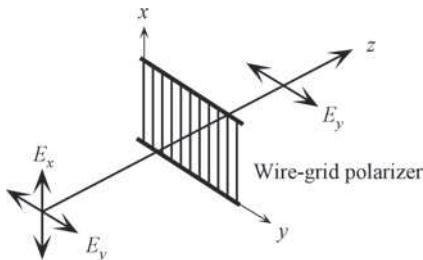


FIGURE 6.42 The wire grid acts as a polarizer.

- 6.4 Anisotropic scattering and polarization** It is well known that an oscillating electric dipole emits electromagnetic radiation. Figure 6.43 (a) shows a snapshot of the electric field pattern around an oscillating electric dipole moment $p(t)$ parallel to the y -axis. There is no field radiation along the dipole axis y . The irradiance I of the radiation along a direction at an angle θ perpendicular to the dipole axis is proportional to $\cos^2\theta$. Sketch the relative radiation intensity pattern around the oscillating dipole. Suppose that the electric field in an incoming electromagnetic wave induces dipole oscillations in a molecule of the medium. Explain how scattering of an incident unpolarized electromagnetic wave by this molecule leads to waves with different polarizations along the x - and y -axes in Figure 6.43 (b).

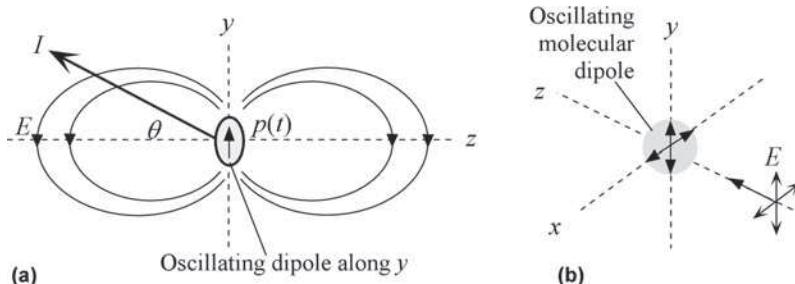


FIGURE 6.43 (a) A snapshot of the field pattern around an oscillating dipole moment in the y -direction. Maximum electromagnetic radiation is perpendicular to the dipole axis and there is no radiation along the dipole axis. (b) Scattering of electromagnetic waves from induced molecular dipole oscillations is anisotropic.

- 6.5 Malus's law** Find the angle between the transmission axes of two polarizers for the transmitted light intensity through both polarizers to be 60%, and 30%.

- 6.6 Modulation by Malus's law** Suppose that a linearly polarized light is passed through a polarizer placed with its transmission axis at an angle $\pi/4$ (or 45°) to the incoming optical field. Suppose now that we *rotationally modulate* the transmission axis of the polarizer by small amounts ϕ about $\pi/4$ (by using a suitable piezoelectric actuator). Show that the changes in the transmission intensity is

$$\Delta I \propto -\phi + \frac{2}{3}\phi^3 - \dots$$

where ϕ is in radians. What is the extent of change (in degrees) in ϕ so that the second term is only 1% of the first term? What is your conclusion?

- 6.7 Birefringence** Consider a negative uniaxial crystal such as calcite ($n_o < n_e$) plate that has the optic axis (taken along z) parallel to the plate face. Suppose that a linearly polarized wave is incident at normal incidence on a plate face. If the optical field is at an angle 45° to the optic axis, sketch the rays through the calcite plate.
- 6.8 Wave plates** A quarter-wave plate is rotated between two crossed polaroids. If an unpolarized beam is incident on the first polaroid, describe the variation of intensity of the emergent beam as the quarter-wave plate is rotated. What would happen if we have a half-wave plate in place of the quarter-wave plate?
- 6.9 Soleil Compensator** Consider a Soleil compensator as shown in Figure 6.15 that uses a quartz crystal. Given a light wave with a wavelength $\lambda \approx 633 \text{ nm}$, a lower plate thickness of 6 mm, calculate the range of d values in Figure 6.15 that provide a retardation from 0 to $\pi/2$ (quarter-wavelength).
- 6.10 Quartz Wollaston prism** Draw a quartz Wollaston prism and clearly show and identify the directions of orthogonally polarized waves traveling through the prisms. How would you test the polarization states of the emerging rays? Consider two identical Wollaston prisms, one from calcite and the other from quartz. Which will have a greater beam-splitting ability? Explain.
- 6.11 Glan–Foucault prism** Figure 6.44 shows the cross-section of a Glan–Foucault prism which is made of two right angle calcite prisms with a prism angle of 38.5° . Both have their optic axes parallel to each other and to the block faces as in the figure. Explain the operation of the prisms and show that the o -wave does indeed experience total internal reflection.

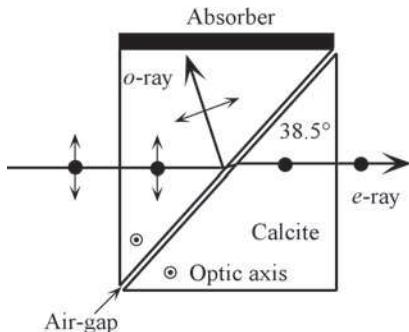


FIGURE 6.44 The Glan–Foucault prism provides linearly polarized light.

6.12 Optical activity

- (a) Consider an optically active medium. The experimenter A (Alan) sends a vertically polarized light into this medium as shown in Figure 6.17. The light that emerges from the back of the crystal is received by an experimenter B (Barbara). B observes that the optical field \mathbf{E} has been rotated to \mathbf{E}' counterclockwise. She reflects the wave back into the medium so that A can receive it. Describe the observations of A and B . What is your conclusion?
- (b) Figure 6.45 shows a simplified version of the Fresnel prism that converts an incoming unpolarized light into two divergent beams that have opposite circular polarizations (R = right, L = left; divergence is exaggerated).

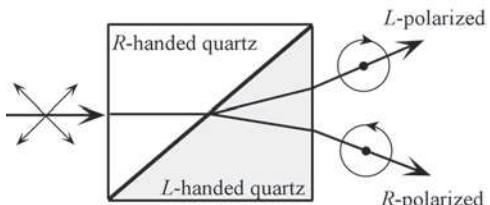


FIGURE 6.45 The Fresnel prism for separating unpolarized light into two divergent beams with opposite circular polarizations (R = right, L = left; divergence is exaggerated).

6.13 Liquid crystal display

- Consider the LCD display shown in Figure 6.20.
- (a) The LCD display shown in Figure 6.20 would be one of the pixels on an LCD display. Based on Chapter 5, how would you use an active matrix array technology to drive the pixels on an LCD screen?
- (b) Sketch how you would construct an LCD cell that is normally black and becomes bright upon the application of an AC voltage.

- 6.14 Liquid crystal display** The LCD characterization measurements of M. Schadt and W. Helfrich reported in their paper in *Applied Physics Letters* in 1971 (18, 127, 1971) are summarized in Table 6.6. Φ is the rotation of the linearly polarized light through the liquid crystal and V_{rms} is the rms voltage of the 1 kHz AC voltage used to drive the LCD. The normalized transmittance $T' = T(V_{\text{rms}})/T_{\text{max}}$ is the measured transmittance divided by the maximum transmittance. Their LCD cell was 10 μm in thickness, whereas most LCDs in modern displays have a thickness of about 5 μm or less. (a) Plot Φ vs. V_{rms} . (b) Plot T' vs. V_{rms} . (c) What are the threshold and saturation voltages?

**TABLE 6.6 Data obtained by M. Schadt and W. Helfrich on a twisted nematic liquid crystal cell.
The normalized transmission T' is the measured transmittance divided by the maximum transmittance, i.e., $T(V_{\text{rms}})/T_{\text{max}}$**

V_{rms} (V)	0.17	0.98	3.00	3.57	3.97	4.49	4.90	5.30	5.99	6.99
Φ	89.4°	89.2°	89.6°	84.5°	71.7°	52.3°	34.6°	16.9°	8.84°	2.75°
V_{rms} (V)	0.20	1.50	3.00	3.39	3.71	4.01	4.50	5.00	5.46	6.00
T' (%)	3.41	3.41	3.41	6.83	15.1	30.1	60.0	80.1	91.8	96.2

(Source: Data extracted from Figure 2 in the original paper.)

- 6.15 LiNbO₃ phase modulator** What should be the aspect ratio d/L for the transverse LiNbO₃ phase modulator in Figure 6.24 that will operate at a free-space wavelength of 1.3 μm and will provide a phase shift $\Delta\phi$ of π (half wavelength) between the two orthogonal field components propagating through the crystal for an applied voltage of 12 V?

We cannot arbitrarily set d/L to any ratio we like for the simple reason that when d becomes too small, the light will suffer diffraction effects that will prevent it from passing through the device. Consideration of diffraction effects leads to²⁶

$$d \approx 2 \left(\frac{\lambda L}{n_o \pi} \right)^{1/2}$$

Taking the crystal length $L \approx 20$ mm, calculate d and hence the new aspect ratio.

- 6.16 Transverse Pockels cell with LiNbO₃** Suppose that instead of the configuration in Figure 6.25, the field is applied along the z -axis of the crystal and the light propagates along the y -axis. The z -axis is the polarization of the ordinary wave and x -axis that of the extraordinary wave. Light propagates through as o - and e -waves. Given that $E_a = V/d$, where d is the crystal length along z (crystal thickness), the indices are

$$n'_o \approx n_o + \frac{1}{2} n_o^3 r_{13} E_a \quad \text{and} \quad n'_e \approx n_e + \frac{1}{2} n_e^3 r_{33} E_a$$

Show that the phase difference between the o - and e -waves emerging from the crystal is

$$\Delta\phi = \phi_e - \phi_o = \frac{2\pi L}{\lambda} (n_e - n_o) + \frac{2\pi L}{\lambda} \left[\frac{1}{2} (n_e^3 r_{33} - n_o^3 r_{13}) \right] \frac{V}{d}$$

where L is the crystal length along the y -axis.

Explain the first and second terms. How would you use two such Pockels cells to cancel the first terms in the total phase shift for the two cells?

If the light beam entering the crystal is linearly polarized in the z -direction, show that

$$\Delta\phi = \frac{2\pi n_e L}{\lambda} + \frac{2\pi L}{\lambda} \left[\frac{(n_e^3 r_{33})}{2} \right] \frac{V}{d}$$

Consider a nearly monochromatic light beam of the free-space wavelength $\lambda = 633$ nm and polarization along z -axis. Calculate the voltage V_π needed to change the output phase $\Delta\phi$ by π given a LiNbO₃ crystal with $d/L = 0.01$ (see Table 6.2).

²⁶ A. K. Ghatak and K. Thyagarajan, *Optical Electronics* (Cambridge University Press, 1989), pp. 477–479.

6.17 Longitudinal Pockels cell

- (a) Sketch schematically the structure of a longitudinal Pockels cell in which the applied field is along the direction of light propagation, both parallel to the z -axis (optic axis). Suggest schemes that would allow light to enter the crystal along the applied field direction.
- (b) Suppose that an LiNbO₃ crystal is used. LiNbO₃ is uniaxial and $n_1 = n_2 = n_o$ (polarizations parallel to x and y) and $n_3 = n_e$ (polarization parallel to z). Neglecting the rotation of the axes (same principal axes in the presence of an applied field), the new ordinary refractive index is

$$n'_o \approx n_o + \frac{1}{2} n_o^3 r_{13} E_a$$

Calculate the half-wave voltage required to induce a retardation of π between the emerging and incident waves if the free-space wavelength is 1 μm . What are their polarizations? (Note: For LiNbO₃ at 633 nm, $n_o \approx 2.28$ and $r_{13} \approx 9 \times 10^{-12} \text{ m V}^{-1}$.)

- (c) Suppose that a KDP crystal is used. KDP is uniaxial and $n_1 = n_2 = n_o$ (polarizations parallel to x and y) and $n_3 = n_e$ (polarization parallel to z). The principal axes x and y are rotated by 45° to become x' and y' as in Figure 6.23 and

$$n'_1 \approx n_o - \frac{1}{2} n_o^3 r_{63} E_a \quad n'_2 \approx n_o + \frac{1}{2} n_o^3 r_{63} E_a \quad \text{and} \quad n'_3 = n_3 = n_e$$

Calculate the half-wave voltage required to induce a retardation of π between the emerging components of the electric field for free-space wavelength of 633 nm if, for KDP at this wavelength, $n_o \approx 1.51$ and $r_{63} \approx 10.3 \times 10^{-12} \text{ m V}^{-1}$.

- 6.18 Response time and bandwidth of an EO phase modulator** Consider the electro-optic modulator in Figure 6.46 (a). Suppose that we suddenly apply a step-voltage from a supply with an output resistance $R_s = 50 \Omega$ (or this may be the impedance of a coaxial cable bringing in the pulse). The EO crystal between the electrodes is a dielectric with a relative permittivity ϵ_r . The time required to charge (and discharge) the capacitance between the electrodes is determined by the time constant of the electric circuit, $\tau = R_s C_{\text{EO}}$, where C_{EO} is the capacitance of the EO crystal. On the other hand, light has to propagate through the length L of the EO crystal. This transit time τ_{light} of light is determined by the length L and the refractive index n . Thus,

$$\tau = \frac{R_s L W \epsilon_o \epsilon_r}{D} \quad \text{and} \quad \tau_{\text{light}} = \frac{L}{c/n}$$

- (a) Calculate these characteristic times for a GaAs crystal assuming an operation at $\lambda = 1.3 \mu\text{m}$, length $L = 20 \text{ mm}$, width $W = 2 \text{ mm}$, and thickness $D = 2 \text{ mm}$ given that $\epsilon_r \approx 12$ and $n \approx 3.6$.
- (b) Consider Figure 6.46 (b) in which a step voltage is applied to an integrated modulator. The effective capacitance between the electrodes is $C_{\text{eff}} = g L W \epsilon_o \epsilon_r / d$, where g is a numerical geometric factor that accounts for the fact that field lines have to bend from one electrode to the other. Consider a LiNbO₃ crystal based modulator. Take length $L \approx 10 \text{ mm}$, width $W \approx 10 \mu\text{m}$, electrode separation $d \approx 10 \mu\text{m}$, $g \approx 0.6$, $n \approx 2.2$, $\epsilon_r \approx 5$ (high frequency), and calculate τ and τ_{light} .

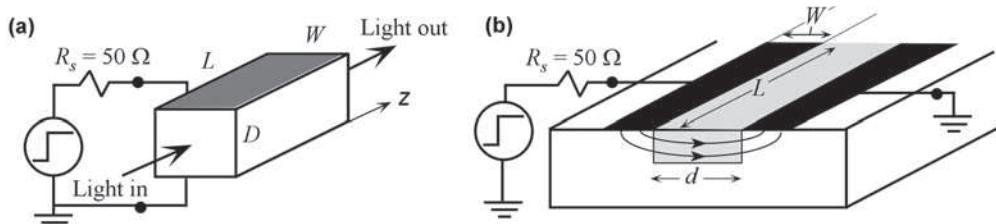


FIGURE 6.46 (a) A step voltage is suddenly applied to an EO modulator. L = length, W = width, D = thickness or depth. (b) A step voltage is applied to an integrated phase modulator. L = electrode length, W = width of electrode, d = separation of electrodes.

- 6.19 Acousto-optic figure of merit.** Using the values for n , p , ρ , and v_a in Table 6.3, calculate the AO figure of merit for LiNbO₃ and GaP.
- 6.20 Acousto-Optic Modulator** A TeO₂ crystal is used to generate 220 MHz acoustic waves. The RF transducer has a length (L) of 8 mm and a height (H) of 8 mm. Calculate the acoustic wavelength, and hence the Bragg deflection angle. What is the Doppler shift in the wavelength? What is the relative intensity in the first order reflected beam for RF acoustic powers 1 W, and 7 W?
- 6.21 Acousto-Optic Modulator Comparison** An AO modulator operating at 1.55 μm can switch the through beam to the deflected beam with 90% efficiency, that is, I_1/I_i is 90% when we apply an RF voltage to the piezoelectric transducer. Assume that L/H is roughly 1. Calculate the required acoustic power for an AO modulator using TeO₂ and a chalcogenide glass based AO modulator using the values in Table 6.3.
- 6.22 LiNbO₃ SAW AO modulator** A SAW AO modulator is shown in Figure 6.36 and operates at a wavelength of 1.5 μm . The interdigital electrodes are driven from a 500 MHz RF source. What should be the periodicity of the interdigital electrodes? What is the deflection angle? If the length L of the transducer is 5 mm, and the power in the acoustic wave is 200 mW, estimate the diffraction efficiency.
- 6.23 Bragg acousto-optic modulator and diffraction** Consider the acousto-optic modulator in Figure 6.32. We can represent the incident and diffracted optical waves in terms of their wave vectors \mathbf{k} and \mathbf{k}' . The incident and the diffracted photons will have energies $\hbar\omega$ and $\hbar\omega'$ and momenta $\hbar\mathbf{k}$ and $\hbar\mathbf{k}'$. An acoustic wave consists of **lattice vibrations** (vibrations of the crystal atoms) and these vibrations are quantized just like electromagnetic waves are quantized in terms of photons. A quantum of lattice vibration is called a **phonon**. A traveling lattice wave is essentially a strain wave and can be represented by $S = S_o \cos(\Omega t - Kx)$ where S is the instantaneous strain at x , Ω is the angular acoustic frequency, K is the wave vector, $K = 2\pi/\Lambda$, S_o is the amplitude of the strain wave, and Λ is the wavelength of the lattice vibration. A phonon has an energy $\hbar\Omega$ and a momentum $\hbar K$. When an incoming photon is diffracted, it does so by interacting with a phonon; it can absorb or generate a phonon. We can treat the interactions as we do between any two particles; they must obey the conservation of momentum and energy rules:

$$\begin{aligned}\hbar\mathbf{k}' &= \hbar\mathbf{k} \pm \hbar\mathbf{K} \\ \hbar\omega' &= \hbar\omega \pm \hbar\Omega\end{aligned}$$

The positive sign case is illustrated in Figure 6.47 which involves absorbing a phonon. Since the acoustic frequencies are orders of magnitude smaller than optical frequencies ($\Omega \ll \omega$), we can assume that, in magnitude, $k' \approx k$. Hence using the above rules, which correspond to Figure 6.47, derive the Bragg diffraction condition.

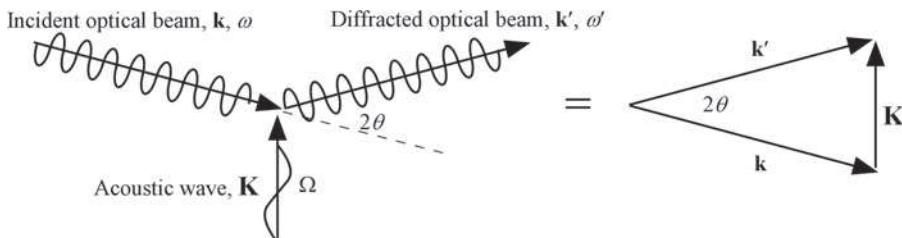


FIGURE 6.47 Wave vectors for the incident and diffracted optical waves and the acoustic wave.

- 6.24 Faraday effect** Given that glass and ZnTe have Verdet constants of about 5 and 190 rad T⁻¹m⁻¹, respectively, for sodium light having wavelength 593 nm, calculate the necessary magnetic field for a rotation of 2° over a length 20 mm. What is the rotation per unit magnetic field for a medium of length 15 cm?
- 6.25 Faraday effect** Consider a ZnTe crystal which rotates the optical field of the 633 nm polarized laser beam from a He-Ne laser. If the crystal length is 2 cm, and the magnetic field is 0.9 T, what is the rotation of the optical field?
- 6.26 Verdet constant** Silica-germania (SiO₂-13.5%GeO₂) glass has a refractive index of 1.4698 at 800 nm. The group index N_g at this wavelength is 1.4890. What is its Verdet constant?
- 6.27 Magneto-optic modulator and isolator**
- Sketch schematically how you would construct a light intensity modulator using the Faraday effect. What would be its advantages and disadvantages?
 - Sketch schematically how you would construct an optical isolator, using a Faraday rotator and two polarizers, that allows light to travel in one direction but not in the opposite direction.

- 6.28 SHG** The mismatch between k_2 for the second harmonic wave vector and k_1 for the fundamental wave is defined by $\Delta k = k_2 - 2k_1$. Perfect match means $k_2 = 2k_1$ and $\Delta k = 0$. When $\Delta k \neq 0$, then the coherence length l_c is given by $l_c = \pi/(\Delta k)$. Show that

$$l_c = \frac{\lambda}{4(n_2 - n_1)}$$

where λ is the free-space wavelength of the fundamental wave. Suppose that a light with wavelength 1000 nm is passed through a KDP crystal along its optic axis. Given that $n_o = 1.509$ at $\lambda = 1000$ nm and $n_o = 1.530$ nm at 2λ , what is the coherence length l_c ? Find the percentage difference between n_2 and n_1 for a coherence length of 2 mm.

- 6.29 Optical Kerr effect** Consider a material in which the polarization does not have the second-order term:

$$P = \epsilon_o \chi_1 E + \epsilon_o \chi_3 E^3 \quad \text{or} \quad P/(\epsilon_o E) = \chi_1 + \chi_3 E^2$$

The first term with the electric susceptibility χ_1 corresponds to the relative permittivity ϵ_r and hence to the refractive index n_o of the medium in the absence of the third-order term, *i.e.*, under low fields. The E^2 term represents the irradiance I of the beam. Thus, the refractive index depends on the intensity of the light beam, a phenomenon called the optical Kerr effect:

$$n = n_o + n_2 I \quad \text{and} \quad n_2 = \frac{3\eta\chi_3}{4n_o^2}$$

where $\eta = (\mu_o/\epsilon_o)^{1/2} = 120\pi = 37 \Omega$, is the impedance of free space.

- (a) Typically, for many glasses, $\chi_3 \approx 10^{-21} \text{ m}^2 \text{ W}^{-1}$; for many doped glasses, $\chi_3 \approx 10^{-18} \text{ m}^2 \text{ W}^{-1}$; for many organic substances, $\chi_3 \approx 10^{-17} \text{ m}^2 \text{ W}^{-1}$; for semiconductors, $\chi_3 \approx 10^{-14} \text{ m}^2 \text{ W}^{-1}$. Calculate n_2 and the intensity of light needed to change n by 10^{-3} for each case.
- (b) The phase ϕ at a point z is given by

$$\phi = \omega_o t - \frac{2\pi n}{\lambda} z = \omega_o t - \frac{2\pi(n_o + n_2 I)}{\lambda} z$$

It is clear that the phase depends on the light intensity I and the change in the phase along Δz due to light intensity alone is

$$\Delta\phi = \frac{2\pi n_2 I}{\lambda} \Delta z$$

As the light intensity modulates the phase of the wave, this is called **self-phase modulation**. Obviously light is controlling light. When the light intensity is small $n_2 I \ll n_o$, the instantaneous frequency

$$\omega = \partial\phi/\partial t = \omega_o$$

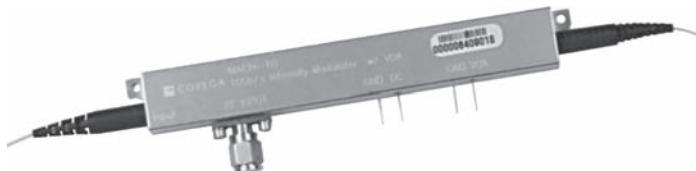
Suppose we have an intense beam and the intensity I is time dependent, $I = I(t)$. Consider a pulse of light traveling along the z -direction and the light intensity vs. t shape is a “Gaussian” (this is approximately so when a light pulse propagates in an optical fiber, for example). Find the instantaneous frequency ω . Is this still ω_o ? How does the frequency change with “time,” or across the light pulse? The change in the frequency over the pulse is called **chirping**. Self-phase modulation therefore changes the frequency spectrum of the light pulse during propagation. What is the significance of this result?

- (c) Consider a Gaussian beam in which the intensity across the beam cross-section falls with radial distance in a Gaussian fashion. Suppose that the beam is made to pass through a plate of nonlinear medium. Explain how the beam can become **self-focused**. Can you envisage a situation where diffraction effects trying to impose divergence are just balanced by self-focusing effects?

- 6.30. Jones vector** Consider a light wave that has the Jones vector \mathbf{J}_{in} , and is put through an optical modulation that has a transfer matrix \mathbf{T} ,

$$\mathbf{J}_{\text{in}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{T} = \begin{bmatrix} j & 0 \\ 0 & 1 \end{bmatrix}$$

What do these represent, and what is the output Jones vector \mathbf{J}_{out} ? If we take the output and put it through another optical modulation that has the same transmission matrix, what is the output Jones vector?



Twenty GHz analog Mach-Zehnder intensity modulator for use at 1550 nm, with half-wave voltage of 3.5 V. The modulator uses titanium-diffused Z-cut LiNbO₃. (Courtesy of Thorlabs.)



Fiber-coupled acousto-optic modulator. (Courtesy of Gooch & Housego.)

APPENDIX A

Gaussian Distribution

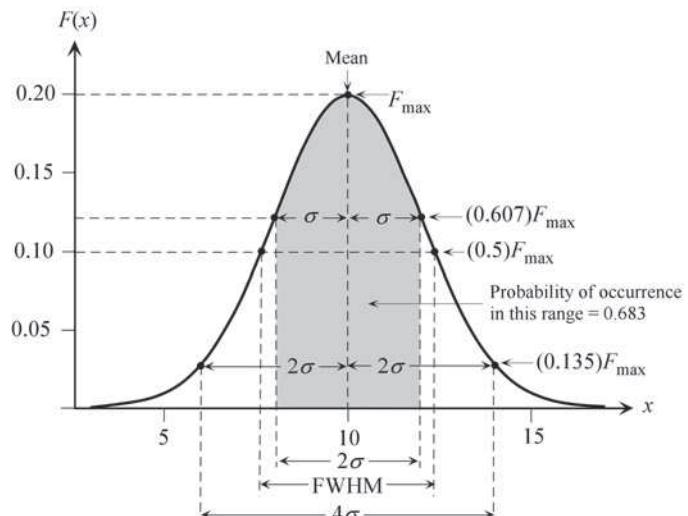
The **Gaussian distribution** refers to the *normal distribution* found in statistics; that is, it is a particular probability density function, denoted $F(x)$, of a variable x such that $F(x)$ is continuous and has a symmetrical shape about a mean value \bar{x} with the functional form given by

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x - \bar{x})^2}{2\sigma^2}\right]$$

in which σ^2 is the **variance**, *i.e.*, the mean value of $(x - \bar{x})^2$ found by integrating $(x - \bar{x})^2 F(x)$ over all x (from $-\infty$ to $+\infty$); σ is the **standard deviation**. $F(x)$ is called the **Gaussian function**. The integral of $F(x)$ over all x is unity, *i.e.*, $\int F(x) dx = 1$. Suppose that we observe a property x of a process or a phenomenon that has a Gaussian probability density function, then we can find the total probability ΔP of the occurrence of x in an interval Δx from x_1 to x_2 from

$$\Delta P = \int_{x_1}^{x_2} F(x) dx$$

Put differently, the probability dP of observing x with a value in x to $x + dx$ is $F(x)dx$. The Gaussian function is widely used in representing many observable quantities. For example, the output spectrum from gas lasers usually has a Gaussian distribution. The maximum of the Gaussian distribution is $F_{\max} = F(\bar{x}) = 0.399/\sigma$. When $x = \sigma$, one standard deviation away from the peak, $F(\sigma) = 0.242/\sigma$, and $F(\sigma)/F_{\max} = 0.607$. In moving 2σ from the peak, the Gaussian drops from its maximum by a factor of 0.135. The full width at half maximum, **FWHM**, is the width $\Delta x = |x_2 - x_1|$, where $F(x_1) = F(x_2) = (1/2)F_{\max}$, and $\text{FWHM} = \sigma 2\sqrt{2 \ln(2)} = 2.355\sigma$, or $\sigma = 0.425\Delta x$.



The Gaussian distribution that has $\bar{x} = 10$ and $\sigma = 2$. The shaded area is the integration from $x = \bar{x} - \sigma$ to $x = \bar{x} + \sigma$ and has a probability of 0.683.

Selected Elementary Properties of the Gaussian Distribution

Mean	\bar{x}	Total area	1
Variance	σ^2	Probability in $\bar{x} \pm \sigma$	0.683
Maximum, F_{\max}	$(\sigma\sqrt{2\pi})^{-1}$	Probability in $\bar{x} \pm 2\sigma$	0.955
$F(\sigma)$	$0.607F_{\max}$	FWHM	$\sigma 2\sqrt{2 \ln(2)} = 2.355\sigma$



A UV LED that emits at 315 nm. (*Courtesy of Chris Collins.*)

APPENDIX B

Solid Angles

The **solid angle** is an angle Ω in three dimensions that an area S on a sphere subtends at the center of the sphere whose magnitude is given by $\Omega = S/r^2$, as illustrated in Figure B.1. More rigorously, the solid angle is defined in terms of an **element of solid angle** $d\Omega$, defined by

$$d\Omega = \frac{dS}{r^2}$$

where dS is the elemental area on a sphere of radius r as shown in Figure B.2 (a). The units for the solid angle are **steradians**, abbreviated to sr. Steradians are essentially *square radians*. The solid angle is a dimensionless quantity.

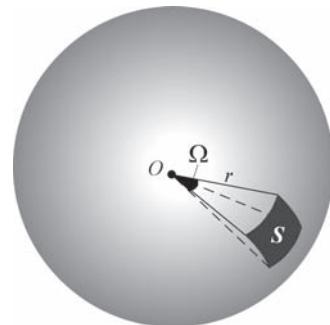


FIGURE B.1 The definition of the solid angle Ω in terms of the area S on a sphere of radius r .

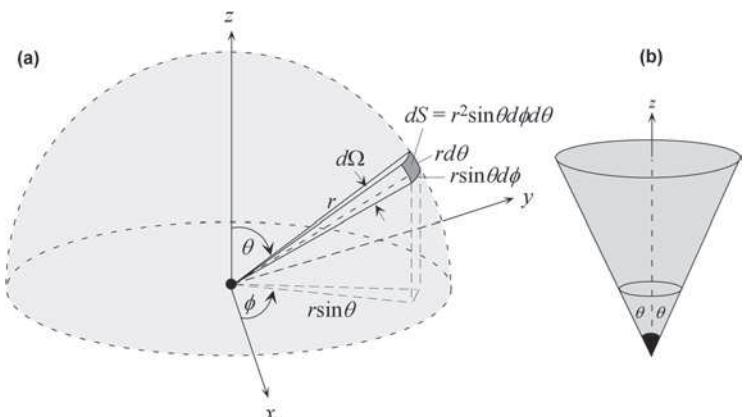


FIGURE B.2 (a) The definition of the elemental solid angle $d\Omega$ in terms of spherical coordinates r , θ , and ϕ . (b) The solid angle of a cone with an apex angle 2θ .

It is useful to express the differential solid angle $d\Omega$ in spherical coordinates. The area $dS = r^2 \sin \theta d\phi d\theta$ so that

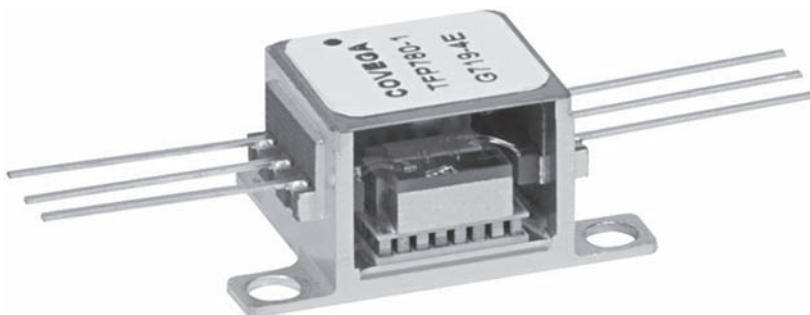
$$d\Omega = \sin \theta d\phi d\theta$$

For a whole sphere, $S = 4\pi r^2$ so that $\Omega = 4\pi$ sr, and for a hemisphere Ω is 2π sr. The surface area S for 1 sr is about 8% of the total spherical surface area.

The solid angle that corresponds to a cone of apex angle 2θ is of particular importance, especially when LED emission distribution is considered as in Figure B.2 (b). In this case, light is emitted from a point into a cone centered around the z -axis, and the angle between the z -axis and the side of the cone is θ . Integrating $d\Omega$ above, we find

$$\Omega = \int_0^{2\pi} \int_0^\theta \sin \theta d\phi d\theta = 2\pi(1 - \cos \theta) \approx \pi\theta^2$$

where the approximate sign is for small angles only (θ in radians). A cone of apex angle 2θ subtends a solid angle Ω given by the above expression. Thus, 1 sr is equivalent to an apex angle of 65.5° . The definitions of many radiometric and photometric quantities involve a solid angle.



Covega Fabry-Perot semiconductor optical gain chip for use in external cavity tunable diode laser applications in the 750–790 nm range. The package has a thermoelectric (TE) cooler and a thermistor for temperature control of the chip. (Courtesy of Thorlabs.)

APPENDIX C

Basic Radiometry and Photometry

The science of radiation measurement—for example, the measurement of emitted, absorbed, reflected, and transmitted radiation energy—is called **radiometry**; radiation is understood to mean electromagnetic energy in the optical frequency range (ultraviolet, visible, and infrared). **Photometry**, on the other hand, is a subset of radiometry in which radiation is measured with respect to the spectral responsivity of the eye, that is, over the visible spectrum and by taking into account the spectral visual sensitivity of the eye under normal light-adapted conditions, *i.e.*, *photopic* conditions.

The **flux** (Φ) in radiometry has three related definitions—**radiant**, **luminous**, and **photon flux**—which correspond to the rate of flow of radiation energy, perceptible visual energy, and photons. (Notice that, in radiometry, these fluxes are not defined in terms of flow *per unit area*.) For example, **radiant flux** is the energy flow per unit time in the units of Watts. Radiometric quantities, such as radiant flux, Φ_e (radiant energy flow per unit time), usually have a subscript *e*, and invariably involve energy or power. Radiometric *spectral* quantities, such as **spectral radiant flux** Φ_λ , refer to the radiometric quantity per unit wavelength, *i.e.*, $\Phi_\lambda = d\Phi_e/d\lambda$ is the radiant flux per unit wavelength.

The radiant flux Φ_e is simply referred to as the **optical power** P_o in many textbooks on optoelectronics due to the common use of P in representing power.

The **luminous flux** or **photometric flux** Φ_v , is the visual “brightness” of a source as observed by an average daylight-adapted eye, and is proportional to the radiant flux (radiation power emitted) of the source and the efficiency of the eye in detecting the spectrum of the emitted radiation.¹ While the eye can see a red color source, it cannot see an infrared source and the luminous flux of the infrared source would be zero. Similarly, the eye is less efficient in the violet than in the green region, and less radiant flux is needed to have a green source at the same luminous flux as the violet source. Luminous flux Φ_v is measured in **lumens** (lm) (also defined in Chapter 3), and at a particular wavelength it is given by

Luminous
flux

$$\Phi_v = \Phi_e \times K \times V(\lambda) \quad (\text{C.1.1})$$

where Φ_e is the radiant flux (in watts), K is a conversion constant (standardized to be 633 lm W^{-1}), and $V(\lambda)$ is the **luminous efficiency** (sometimes also called the *luminous efficacy*) of the daylight-adapted eye, which is unity at 555 nm. The luminous efficiency of the eye is written as $V(\lambda)$ to highlight its dependence on the wavelength. $V(\lambda)$ is also called the **visibility function**. The dependence of $V(\lambda)$ on λ is shown in Figure 3.41.

By definition, a 1 W light source emitting at 555 nm (amber green, where $V = 1$) emits a luminous flux of 633 lm. The same 1 W light source at 650 nm (red), where $V = 0.11$, emits only 70 lm. When we buy a light bulb, we are essentially paying for lumens because it is luminous flux that the eye perceives. A typical 100 W incandescent lamp provides roughly 1700 lm. Fluorescent tubes provide more luminous flux output than incandescent lamps for the same electrical power input as they have more spectral emission in the visible region and make better use of the eye’s spectral sensitivity.

¹The subscript *v* represents *visible*.

The **radiant intensity** is the radiant flux (radiant energy flowing per unit time) per unit solid angle; it is a radiometric quantity with the units of W sr^{-1} . It is a particularly useful quantity for point sources. If $d\Phi_e$ is the radiant flux within a solid angle $d\Omega$ in a particular direction θ , ϕ , with respect to spherical coordinates placed at the source, as in Figure C.1 (a), then

$$I_e = \frac{d\Phi_e}{d\Omega} = r^2 \frac{d\Phi_e}{dS} \quad (\text{C.1.2}) \quad \begin{matrix} \text{Radiant} \\ \text{intensity} \end{matrix}$$

The **luminous intensity** is the luminous flux per unit solid angle along a particular direction, and measures the visual **brightness** of radiation as observed by a daylight-adapted eye within a unit solid angle. If $d\Phi_v$ is the luminous flux within a solid angle $d\Omega$ for a direction θ , ϕ , as in Figure C.1 (a), then luminous intensity is defined as

$$I_v = \frac{d\Phi_v}{d\Omega} = r^2 \frac{d\Phi_v}{dS} \quad (\text{C.1.3}) \quad \begin{matrix} \text{Luminous} \\ \text{intensity} \end{matrix}$$

The units for the luminous intensity are lumens per steradian or **candela** (cd). Both the radiant and luminous intensity, I_e and I_v , usually depend on the angles θ and ϕ . Suppose that a source emits over 4π steradians and emits equally in all directions. This would be a point source. If the source has a luminous flux of 1 lm, then its luminous intensity is $1/(4\pi)$ cd. A **1 candela** point source emits 4π lumens into space in all directions. It is important to note that the candela is a visual photometric unit (an SI unit) for measuring the light intensity of a source which emits a luminous flux of 1 lm in a particular direction within one unit of solid angle, i.e., $1 \text{ cd} = 1 \text{ lm sr}^{-1}$, at a reference frequency $\nu = 540 \times 10^{12} \text{ Hz}$ ($\lambda \approx 555 \text{ nm}$).² Thus,

$$1 \text{ cd} = \left(\frac{1}{633} \right) \text{W sr}^{-1} \quad (\text{C.1.4}) \quad \begin{matrix} \text{One} \\ \text{candela} \end{matrix}$$

Conventionally, the intensity is plotted on a polar plot, as illustrated in Figure C.1 (b). A point on this plot has an angle θ with respect to z , which is usually taken along the maximum intensity. The distance from the origin is either I_v or proportional to I_v . Such polar plots represent

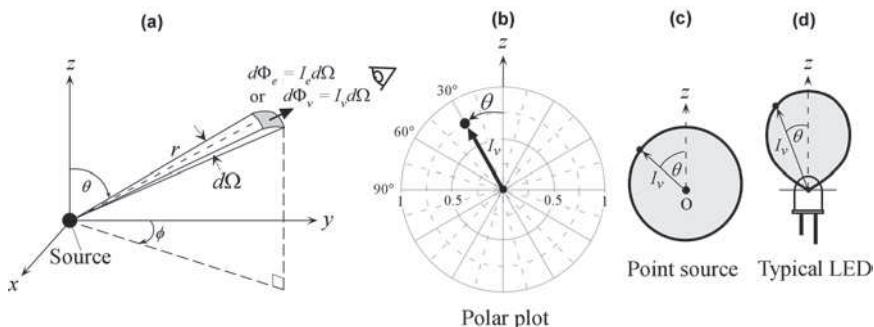


FIGURE C.1 (a) Radiant or luminous flux that is within an elemental solid angle $d\Omega$ and the definition of radiant or luminous intensity. (b) Representation of luminous intensity in polar coordinates. The point shown has a luminous intensity of 0.75 cd at an angle of 30° to z . (c) Isotropic emitter. (d) Typical angular dependence of luminous intensity from a single LED.

²Definition by the 16th Conférence Générale des Poids et Mesures—General Conference on Weights and Measures, 1979.

the angular distribution of intensity from a source. For a **point source**, $I_v(\theta)$ is a circle, as illustrated in Figure C.1 (c), that is $I_v(\theta) = \text{constant}$. An isotropic source emits luminous intensity that is the same in all directions as mentioned above. For a **Lambertian source**, the intensity depends on θ as

Lambertian source

$$I_v(\theta) = I_v(0)\cos\theta \quad (\text{C.1.5})$$

For an exponential intensity source, I_v falls exponentially (very rapidly) with θ , and can be approximated by using

$$I_v(\theta) = I_v(0)(\cos\theta)^n$$

where n is usually greater than unity. If $I_v(\theta) = I_v(0)(1/2)$ when $\theta = \theta_{1/2}$, then $n = \ln(0.5)/\ln[\cos(\theta_{1/2})]$. Figure C.1 (d) shows a typical example on the angular dependence of the luminous intensity from a single LED device.

The radiant intensity can also be a function of wavelength but, in this case, we must use spectral radiant intensity to represent the distribution of intensity over wavelengths. The **spectral radiant intensity** $I_{e\lambda}$ is defined by

Spectral radiant intensity

$$I_{e\lambda} = \frac{dI_e}{d\lambda} \quad (\text{C.1.6})$$

Similarly, the **spectral luminous intensity** $I_{v\lambda}$ is defined by

Spectral luminous intensity

$$I_{v\lambda} = \frac{dI_v}{d\lambda} \quad (\text{C.1.7})$$

Clearly, if we integrate the spectral intensity $I_{e\lambda}$ (or $I_{v\lambda}$) over all wavelengths, we would get the total intensity I_e (or I_v), which is sometimes called the integrated intensity. An integration of spectral intensity over a limited wavelength range $\Delta\lambda$ would, of course, only give the total intensity within $\Delta\lambda$.



Piezoelectric transducer controlled Fabry–Perot etalons. Left has a 70 mm and the right has a 50 mm clear aperture. The piezoelectric controller maintains the reflecting plates parallel while the cavity separation is scanned. (The left etalon has a reflection of interference fringes on its surface plate that are actually on the adjacent computer display; not shown.) (Courtesy of IC Optical Systems Ltd.)

APPENDIX D

Useful Mathematical Formulae

QUADRATIC EQUATION

$$ax^2 + bx + c = 0$$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

COMPLEX NUMBERS

$$j = (-1)^{1/2}$$

$$j^2 = -1$$

$$\exp(j\theta) = \cos\theta + j\sin\theta$$

$$Z = a + jb = re^{j\theta}$$

$$r = (a^2 + b^2)^{1/2} \quad \tan\theta = b/a$$

$$Z^* = a - jb = re^{-j\theta}$$

$$\text{Re}(Z) = a \quad \text{Im}(Z) = b$$

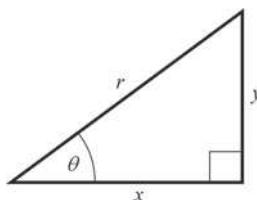
$$\text{Magnitude}^2 = |Z|^2 = ZZ^* = a^2 + b^2$$

$$\text{Argument} = \theta = \arctan(b/a)$$

$$\cos\theta = \frac{1}{2}[e^{j\theta} + e^{-j\theta}]$$

$$\sin\theta = \frac{1}{2j}[e^{j\theta} - e^{-j\theta}]$$

TRIGONOMETRY



$$\sin\theta = y/r$$

$$\tan\theta = y/x = \sin\theta/\cos\theta$$

$$\cos\theta = x/r$$

$$r^2 = x^2 + y^2$$

$$\sin(\pi/2 \pm \theta) = \cos\theta$$

$$\sin^2\theta + \cos^2\theta = 1$$

$$\sin 2\theta = 2\sin\theta\cos\theta$$

$$\cos 2\theta = 1 - 2\sin^2\theta = 2\cos^2\theta - 1$$

$$\sin(A + B) = \sin A \cos B + \cos A \sin B$$

$$\cos(A + B) = \cos A \cos B - \sin A \sin B$$

$$\sin A + \sin B = 2\sin\left[\frac{1}{2}(A + B)\right]\cos\left[\frac{1}{2}(A - B)\right]$$

$$\cos A + \cos B = 2\cos\left[\frac{1}{2}(A + B)\right]\cos\left[\frac{1}{2}(A - B)\right]$$

EXPANSIONS

$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \dots$$

$$(1 + x)^n = 1 + nx + \frac{n(n - 1)}{2!}x^2 + \frac{n(n - 1)(n - 2)}{3!}x^3 + \dots$$

$$f(x_o + \Delta x) = f(x_o) + \Delta x \left[\frac{df}{dx} \right]_{x_o} + \frac{(\Delta x)^2}{2} \left[\frac{d^2 f}{dx^2} \right]_{x_o} + \dots$$

$$\text{Small } x \quad (1 + x)^n \approx 1 + nx \quad \sin x \approx x \quad \tan x \approx x \quad \cos x \approx 1$$

$$\text{Small } \Delta x \text{ in } x = x_o + \Delta x, \quad f(x) \approx f(x_o) + \Delta x \left[\frac{df}{dx} \right]_{x_o}$$

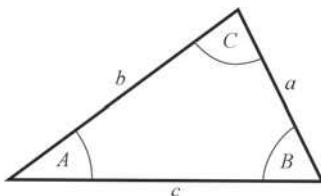
GEOMETRY

$$\text{Area of a circle} = \pi r^2$$

$$\text{Surface area of a sphere} = 4\pi r^2$$

$$\text{Circumference of a circle} = 2\pi r$$

$$\text{Volume of a sphere} = (4/3)\pi r^3$$



$$a^2 = b^2 + c^2 - 2bc \cos A$$

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C}$$

USEFUL RELATIONSHIPS

$$\ln x = 2.303 \log x$$

$$\log x = 0.434 \ln x$$

$$a^x = e^{(\ln a)x}$$

HYPERBOLIC FUNCTIONS

$$\sinh x = (e^x - e^{-x})/2$$

$$\tanh x = (e^x - e^{-x})/(e^x + e^{-x})$$

$$\cosh x = (e^x + e^{-x})/2$$

$$\tanh x = \sinh x / \cosh x$$

APPENDIX E

Notation and Abbreviations

A	area; cross-sectional area
A_1, A_2, A_3	coefficients in the Sellmeier equation (Chapter 1)
A_{21}	Einstein coefficient for spontaneous emission
A_R	Rayleigh attenuation constant in $\alpha_R = A_R/\lambda^4$
a	half dimension of the core region of an optical waveguide; fiber core radius; lattice parameter of a crystal; side of a cubic crystal
a (subscript)	acceptor, <i>e.g.</i> , N_a = acceptor concentration
AC	alternating current
AM	air mass; describes the integrated solar intensity and depends on location
APD	avalanche photodiode
AR	antireflection coating
B	maximum bit-rate; frequency bandwidth; direct recombination capture coefficient
\mathbf{B}, B	magnetic field vector (\mathbf{T}), magnetic field, magnetic flux density
B_{12}, B_{21}	Einstein absorption and stimulated emission coefficients for the rate transitions between energy levels 1 and 2
b	normalized propagation constant (Chapter 2)
BJT	bipolar junction transistor
BL	bit-rate \times distance product ($\text{Gb s}^{-1} \times \text{km}$)
C	capacitance
C_{dep}	depletion region capacitance of a pn junction; junction capacitance of <i>pin</i>
C_{diff}	diffusion capacitance of forward biased pn junction
C_t	terminal (total) capacitance across the terminals of a photodetector
c	speed of light in free space ($3 \times 10^8 \text{ m s}^{-1}$)
CB	conduction band
CCD	charge coupled device
CMOS	complementary metal-oxide semiconductor chip
CVD	chemical vapor deposition
CW	continuous wave
D	diameter of a circular aperture; diffusion coefficient ($\text{m}^2 \text{ s}^{-1}$); detectivity ($\text{Hz}^{1/2} \text{ W}^{-1}$); thickness
\mathbf{D}, D	electric displacement vector (C m^{-2})
D	direct (bandgap)
D^*	specific detectivity ($\text{m Hz}^{1/2} \text{ W}^{-1}$ or Jones)

D_{ch}	chromatic dispersion coefficient ($\text{ps nm}^{-1} \text{km}^{-1}$)
D_e, D_h	electron and hole diffusion coefficients ($\text{m}^2 \text{s}^{-1}$)
D_m	material dispersion coefficient ($\text{ps nm}^{-1} \text{km}^{-1}$)
D_p	profile dispersion coefficient ($\text{ps nm}^{-1} \text{km}^{-1}$)
D_{PMD}	polarization mode dispersion coefficient ($\text{ps km}^{-1/2}$)
D_w	waveguide dispersion coefficient ($\text{ps nm}^{-1} \text{km}^{-1}$)
d	differential in mathematics; distance; diameter; thickness or width
d (subscript)	donor, <i>e.g.</i> , N_d = donor concentration
DBR	distributed Bragg reflector
DC	direct current
DFB	distributed feedback
DH	double-heterostructure
DOS	density of states, number of electronic states per unit energy per unit volume in a given energy band
DWDM	dense wavelength division multiplexing
E	electric field (in Chapters 1, 2, and 6); energy (in Chapters 3, 4, and 5)
E	electric field (in Chapters 3, 4, and 5)
E_a	applied external field (non-optical field) in an electro-optic crystal
E_a, E_d	acceptor and donor energy levels
E_c, E_v	conduction band edge energy, valence band edge energy
E_F	Fermi energy with respect to the bottom of the valence band
E_{Fi}	Fermi energy in the intrinsic semiconductor
E_{Fn}	Fermi energy in the n -side of a pn junction
E_{FO}	Fermi energy in metals at 0 K
E_{Fp}	Fermi energy in the p -side of a pn junction
E_g, E_{go}	bandgap energy and bandgap energy at 0 K (given in eV)
E_n	energy corresponding to quantum number n
E_{ph}	photon energy, $h\nu$ (usually stated in eV)
e	electronic charge ($1.60218 \times 10^{-19} \text{ C}$)
e	Napierian base, 2.71828...
e (subscript)	electron, <i>e.g.</i> , μ_e = electron drift mobility
eV	electron volt
$\exp(x)$	exponential function of x
EDCL	external cavity diode laser
EDFA	erbium doped fiber amplifier
EDQE	external differential quantum efficiency
EE	extraction ratio or efficiency
EHP	electron and hole pair
ELED	edge light emitting diode

EM	electromagnetic
EMF	electromagnetic force (V)
EO	electro-optic
EQE	external quantum efficiency
F	Finesse of a Fabry-Perot cavity (Chapter 1); excess noise factor (Chapter 5)
$f(E)$	Fermi-Dirac function
f_{op}	optical bandwidth
f_r	frequency of relaxation oscillations in a laser diode
FBG	fiber Bragg grating
FF	fill factor of a solar cell
FTIR	frustrated total internal reflection
FWHP	full width at half power
G	gain; rate of generation ($\text{m}^{-3} \text{s}^{-1}$); photoconductive gain (Chapter 5)
G_{op}	net optical gain
G_{ph}	rate of photogeneration per unit volume ($\text{m}^{-3} \text{s}^{-1}$)
$G_{\text{thermal}}, G_{\text{th}}$	thermal generation rate of electron and hole pairs per unit volume ($\text{m}^{-3} \text{s}^{-1}$)
g	optical gain coefficient due to stimulated emissions exceeding absorption (m^{-1})
$g(E), g(E)$	density of states; number of electronic states (possible electron wavefunctions) per unit energy per unit volume ($g(E)$ in Chapter 3 and $g(E)$ in Chapter 4)
$g(v)$	optical gain lineshape; optical gain vs. frequency curve shape in Chapter 4
g_{th}	threshold optical gain coefficient for achieving lasing emission from a laser device (for laser oscillations)
GRIN	graded index
H	height of the optical cavity in a laser diode
\mathbf{H}, H	magnetic field intensity (strength) or magnetizing field (A m^{-1})
h	Planck's constant ($6.6261 \times 10^{-34} \text{ J s}$)
h (subscript)	hole, e.g., μ_h = hole drift mobility
\hbar	Planck's constant divided by 2π , $\hbar = h/2\pi$ ($1.0546 \times 10^{-34} \text{ J s}$)
HF	high frequency
I	electric current (A); irradiance or intensity (W m^{-2}) (in Chapters 1 and 6)
I	irradiance or intensity of light (W m^{-2}) (I used in Chapters 1 and 6)
I	indirect (bandgap)
I_d	dark current
I_{diff}	forward diffusion (Shockley) current of a pn junction
I_{diode}	the diode current; the current in an ideal pn junction
I_{ph}	photocurrent; multiplied APD photocurrent
I_{pho}	unmultiplied APD photocurrent

I_{recom}	forward recombination current in a <i>pn</i> junction
I_{ro}	reverse recombination current in a <i>pn</i> junction
I_{sc}	short circuit current under illumination (solar cells)
I_{sro}	reverse surface recombination current
I_{so}	reverse Shockley saturation current in a <i>pn</i> junction
I_{surf}	surface currents in a <i>pn</i> junction, <i>e.g.</i> , through the surfaces of the crystal
I_{th}	threshold current
i (subscript)	initial; intrinsic, <i>e.g.</i> , n_i = intrinsic concentration
i_n	noise current, rms fluctuations in the photodetector current
i_{ph}	photocurrent
IC	integrated circuit
IDQE	internal differential quantum efficiency
IQE	internal quantum efficiency
J	current density
J_{gen}	reverse current in a <i>pn</i> junction due to thermal generation of EHPs in the SCL
J_{recom}	forward recombination current density in a <i>pn</i> junction
J_{rev}	total reverse current in a <i>pn</i> junction
J_{th}	threshold current density
j	imaginary constant, $\sqrt{-1}$
K	extinction coefficient; the imaginary part of the complex refractive index; Kerr coefficient
\mathbf{K}, K	phonon (lattice vibration) wave vector
k	ratio of the hole ionization coefficient to that for electrons
\mathbf{k}, k	wave vector; propagation constant; free-space propagation constant, $k = 2\pi/\lambda$
k_B	Boltzmann constant ($k_B = R/N_A = 1.3807 \times 10^{-23} \text{ J K}^{-1}$)
KDP	potassium dihydrogen phosphate
KE	kinetic energy
L	distance, length; optical cavity length; inductance
L_o	transfer distance for full light transfer from one to another coupled optical guide
L_p	pump length of an Er-doped fiber amplifier (EDFA)
L_e, L_h	electron and hole diffusion lengths: $L_e = (D_e \tau_e)^{1/2}$ and $L_h = (D_h \tau_h)^{1/2}$
ℓ	orbital angular momentum quantum number of an electron, $0, \pm 1, \pm 2, \dots, n - 1$; n is the principal quantum number
l, l_n, l_p	length, lengths of the neutral n and p regions outside depletion region in a <i>pn</i> junction
$\ln(x), \ln x$	natural logarithm of x
LASER	light amplification by stimulated emission of radiation
LED	light emitting diode

LF	low frequency
LD	laser diode
LP	linearly polarized
M	number of modes (Chapter 2); multiplication in avalanche effect (Chapter 5); mass
M^2	a parameter that gauges the quality of the laser beam in terms of how close it is to a Gaussian beam; it measures the deviation of the real laser beam from true Gaussian characteristics
m	integer; mode number; mass
m	numerical constant characterizing the photon energy width of the LED output spectrum
m_e	mass of the electron in free space (9.10939×10^{-31} kg)
m_e^*, m_h^*	electron and hole effective masses in a crystal
m_l	magnetic quantum number for an electron
m_s	spin magnetic quantum number ($\pm 1/2$)
MBE	molecular beam epitaxy
MFD	mode field diameter, $2w_o$
MMF	effective mass of a hole in a crystal (in the valence band)
multimode fiber	
MQW	multiple quantum well
N	number of molecules per unit volume; n -type semiconductor with a wider bandgap
N	photon concentration
N	complex refractive index
N_A	Avogadro's number (6.022×10^{23} mol $^{-1}$)
NA	numerical aperture
N_a, N_d	acceptor and donor dopant concentrations in a semiconductor (m $^{-3}$)
N_c, N_v	effective density of states at the conduction and valence band edges, respectively, (m $^{-3}$)
N_g	group index, $N_g = c/v_g$
N_{MAC}	MAC value or MAC-number of a fiber
$N_{\text{ph}}, N_{\text{ph}}$	coherent photon concentration in the optical cavity; N used when N refers to another quantity in a chapter
n	refractive index
	electron concentration in the conduction band (in Chapters 3, 4, and 5)
\bar{n}	average refractive index
n	refractive index (used when n represents the electron concentration as in Chapters 3, 4, and 5); quantum number for energies in a quantum well
n_1, n_2	the refractive indices of the core and cladding, respectively
n_2	nonlinear refractive index, <i>i.e.</i> , the change in n , $\Delta n = n_2 I$

n_e, n_o	refractive index for the extraordinary and ordinary waves in a birefringent crystal
n_i	intrinsic concentration (m^{-3})
n_L, n_R	refractive index experienced by left-handed circularly polarized and right-handed circularly polarized waves in a crystal
n_n, p_p	instantaneous majority carrier concentrations (e.g., electrons in an n -type semiconductor)
n_{no}, p_{po}	equilibrium majority carrier concentrations (e.g., electrons in an n -type semiconductor)
n_p, p_n	instantaneous minority carrier concentrations (e.g., electrons in an p -type semiconductor)
n_{po}, p_{no}	equilibrium minority carrier concentrations (e.g., electrons in an p -type semiconductor)
$n_p(0)$	electron concentration just outside the depletion region in the p -side of a pn junction,
n_{th}	threshold electron concentration in the conduction band for threshold gain
NEP	noise equivalent power, 1/Detectivity
NF	noise figure, usually expressed in dB
NRZ	nonreturn to zero (pulses)
OVD	outside vapor deposition
P	power, energy flow per unit time; p -type doping in a larger bandgap material; polarization (Chapter 2)
P	polarization vector; polarization per unit volume in a dielectric (Chapter 6)
P_a	acoustic power (W)
P_f, P_i	final and initial optical power
P_o	optical power (W)
$P_{o\lambda}$	spectral optical power, $dP_o/d\lambda$, that is, optical power per unit wavelength (W nm^{-1})
P_p, P_{pin}	optical pump power; input optical pump power
$P_{\text{sin}}, P_{\text{sout}}$	optical signal power (W) at input and output
p	hole concentration in the valence band
p_e	photoelastic coefficient
pin	p -type/intrinsic/ n -type semiconductor photodiode
$p_n(0)$	hole concentration just outside the depletion region in the n -side of a pn junction
PBG	photonic bandgap
PD	photodiode, photodetector
PE	potential energy
Q	charge; quality factor
q	an integer; diffraction order; charge
QE	quantum efficiency

R	reflectance (fractional reflected light intensity at a dielectric–dielectric interface); spectral responsivity, or responsivity or radiant sensitivity of a photodetector (A/W)
R	resistance; reflectance (depending on context)
R_c	critical radius of curvature for a fiber that characterizes the bending loss in $\alpha_B = A \exp(-R/R_c)$
R_L	load resistance (external to device)
R_p, R_s	equivalent parallel and series resistances of a solar cell
r, r_{22}	Pockels coefficient
r	radial distance
\mathbf{r}	position vector
r_d	dynamic resistance or incremental resistance of a <i>pn</i> junction
r	EM wave reflection coefficient
Re	real part
rms	root mean square
RTZ	return-to-zero as in RTZ data rate; NRZ is nonreturn to zero rate
RZ	return to zero (pulses)
S	strain, cross section
\mathbf{S}, S	Poynting vector that quantifies the rate of electromagnetic energy flow
S_0	slope of the chromatic dispersion (D_{ch}) vs. wavelength (λ) behavior at the wavelength λ_0 at which $D_{ch} = 0$
s, p, f	atomic subshells
SAGM	separate absorption, grading, and multiplication
SAM	separate absorption and multiplication
SCL	space charge layer, depletion region, of a <i>pn</i> junction
SHG	second harmonic generation
SLED	surface emitting LED
SMF	single-mode fiber
SNR, S/N	signal to noise ratio
SQW	single quantum well
T	temperature in Kelvin
T	transmittance, fractional intensity transmitted through a dielectric–dielectric interface
T_f	fictive temperature of glass (K)
t	time; thickness
t	transmission coefficient
t_d	delay time
t_{drift}	drift time of a carrier
t_h	transit time of holes
t_e	transit time of electrons

TA	transmission axis of a polarizer
TE	transverse electric field modes
TEM	transverse electric and magnetic
TIR	total internal reflection
TM	transverse magnetic field modes
U_k	periodic function that modulates the amplitude of the electron wavefunction in a crystal
UV	ultraviolet
V	V -number, V -parameter, normalized thickness; applied voltage, potential energy in the Schrödinger equation (Chapter 3); the Verdet constant (Chapter 6)
$V(\lambda)$	relative luminous efficiency (or the relative sensitivity) of an average light-adapted eye; also called the luminosity function or visibility function
V_{br}	avalanche breakdown voltage
$V_{\lambda/2}$	applied voltage in a Pockels cell that introduces a relative phase change $\Delta\phi$ of π
V_o	built-in voltage or potential in a pn junction
V_{oc}	open circuit voltage of a solar cell
V_r	reverse bias voltage
V_{th}	thermal voltage, $V_{\text{th}} = k_B T/e = 0.026 \text{ V}$
v	velocity
v_d	drift velocity
v_e, v_h	electron and hole drift velocities
v_g	group velocity
v_{th}	thermal velocity of an electron in the conduction band and a hole in the valence band
VB	valance band
VCSEL	vertical cavity surface emitting laser
W	width; thickness; width of depletion region in a pn junction
W_n, W_p	widths of depletion region on the n -side (W_n) and on the p -side (W_p) of the pn junction
W_o	width of depletion region of a pn junction with no applied voltage
w	width; mode field radius, <i>i.e.</i> , $2w =$ mode field diameter
w_o	half-waist of a Gaussian beam (Chapter 1); mode field radius (MFD/2) (Chapter 2)
w_{or}	half-waist of a real laser beam; the minimum diameter of a laser beam
WDM	wavelength division multiplexing
x	distance; excess avalanche noise index for APDs (Chapter 5)
z_o	Rayleigh range of a Gaussian beam
$\hat{\mathbf{x}}$	unit vector along the x -direction

$\hat{\mathbf{y}}$	unit vector along the y -direction
$\hat{\mathbf{z}}$	unit vector along the z -direction
\perp (subscript)	perpendicular
\parallel (subscript)	parallel
α	optical attenuation; attenuation coefficient for the electric field in medium 2; absorption coefficient of a semiconductor
α_{dB}	attenuation in dB
α_e, α_h	electron and hole ionization coefficients (m^{-1})
α_{FIR}	fundamental infrared attenuation in optical fibers (m^{-1})
α_R	Rayleigh scattering attenuation (m^{-1})
α_s	loss coefficient in a laser optical cavity, except from the end mirrors (m^{-1})
α_t	total attenuation coefficient in the optical cavity of a semiconductor laser (m^{-1})
α_{\max}	maximum acceptance angle for an optical fiber; total acceptance angle is $2\alpha_{\max}$
β	propagation constant; base-to-collector bipolar transistor current gain
β_m	propagation constant along a planar dielectric waveguide for mode m
β_T	isothermal compressibility (m^2N^{-1})
Γ	confinement factor that characterizes the confinement of the optical mode to the active region in which there is optical gain (maximum $\Gamma = 1$) (Chapter 4); overlap factor that describes the extent of overlap between the applied field region and the optical mode in a waveguide (maximum $\Gamma = 1$) (Chapter 6)
γ	graded index profile constant; coefficient of index grading
γ_o	optimum graded index profile for lowest intermode dispersion in a graded index fiber
Δ	difference, change; normalized index difference, $\Delta = (n_1 - n_2)/n_1$
$\Delta\tau$	propagation time difference along an optical guide, group delay difference along a fiber
$\Delta\tau_{1/2}$	time width of the output light pulse between half intensity points
Δv_{rms}	the width between rms points of the intensity vs. frequency spectrum
$\Delta\beta$	phase mismatch per unit length between two propagating waves
$\Delta\phi$	phase change
$\Delta\nu$	spectral width (in frequency) of the output spectrum of a light source. This can be between half-intensity points ($\Delta\nu_{1/2}$) or between rms points; frequency bandwidth of optical gain in a medium
$\Delta\lambda$	spectral width (in wavelength) of the output spectrum (intensity vs. wavelength) of a light source. This can be between half-intensity points ($\Delta\lambda_{1/2}$) or between rms points; wavelength interval
$\Delta\lambda_m$	wavelength separation of allowed modes in an optical cavity
Δn	$n - n_o$, excess electron concentration (above equilibrium)
ΔN_{th}	$(N_2 - N_1)_{\text{th}}$, threshold population for threshold gain (lasing emission)

$\Delta\theta$	divergence angle due to diffraction effects
δ	small change; penetration depth of an EM wave from one medium to another; phase difference; profile dispersion parameter in optical fibers, $\delta = -\left(\frac{n_1\lambda}{N_{g1}\Delta}\right)\frac{d\Delta}{d\lambda}$
$\delta\lambda_m$	spectral wavelength width of an allowed mode in an optical cavity
$\epsilon = \epsilon_o\epsilon_r$	permittivity of a medium where ϵ_o = permittivity of free space or absolute permittivity, ϵ_r = relative permittivity or dielectric constant; strain (Chapter 2)
ϵ_n	electron energy in a quantum well with respect to the bottom of the conduction band, n is a quantum number 1, 2,...
$\epsilon'_{n'}$	hole energy in a quantum well with respect to the bottom of the valence band, n' is a quantum number 1, 2,...
η	efficiency; quantum efficiency; ideality factor of a diode ($\eta = 1$ for minority carrier diffusion controlled and $\eta = 2$ for recombination controlled diode current)
η_{EDQE}	external differential quantum efficiency
η_{EE}	extraction efficiency
η_{EQE}	external quantum efficiency
η_{IQE}	internal quantum efficiency
η_{IDQE}	internal differential quantum efficiency
η_i	internal quantum efficiency
η_{LE}	luminous efficacy
η_{PCE}	power conversion efficiency
η_{slope}	slope efficiency of a laser diode
θ	angle; angular spherical coordinate
θ_c	critical angle for total internal reflection
θ_i	angle of incidence
θ_p	Brewster's polarization angle
κ	propagation constant along the transverse direction of a planar dielectric waveguide; coupling coefficient in Bragg gratings
Λ	periodicity of a grating type structure such as a fiber Bragg grating
λ	free-space wavelength
λ_0	the wavelength λ_0 at which chromatic dispersion coefficient $D_{ch} = 0$
λ_B	Bragg wavelength of a grating type structure that satisfies the Bragg condition
λ_c	critical or cutoff wavelength; for $\lambda > \lambda_c$ the fiber becomes a SMF
λ_g	bandgap wavelength; threshold wavelength for photoexcitation; wavelength involved in photon emission from direct recombination across the bandgap
λ_m	mode wavelength
λ_o	central wavelength of a spectrum; free-space wavelength when emphasis is on free space

λ_{QW}	wavelength of emission from a QW (quantum well)
λ_p	pump wavelength
λ_s	signal wavelength
$\mu = \mu_o \mu_r$	magnetic permeability (H m^{-1}); μ_o = absolute permeability, μ_r = relative permeability
μ_d	drift mobility ($\text{m}^2 \text{V}^{-1} \text{s}^{-1}$)
μ_h, μ_e	hole drift mobility, electron drift mobility ($\text{m}^2 \text{V}^{-1} \text{s}^{-1}$)
μm	micron, micrometer
π	pi, 3.14159...
ρ	resistivity ($\Omega \text{ m}$); charge density (C m^{-3})
ρ_{net}	net space charge density
$\rho(v)$	photon energy density per unit frequency (Chapter 4)
$\rho_{\text{eq}}(v)$	Planck's black body equilibrium radiation distribution as a function of frequency (Chapter 4)
σ	conductivity; rms deviation of dispersion of a light pulse going through a fiber
$\sigma_{\text{ab}}, \sigma_{\text{em}}$	absorption cross section (m^2), emission cross section (m^2)
$\sigma_{\text{intramode}}$	dispersion that occurs within a given mode (<i>e.g.</i> in the fundamental mode) in a fiber, usually due to the finite spectral width of the excitation source
$\sigma_{\text{intermode}}$	dispersion in a multimode fiber due to different modes traveling with different group velocities
τ	group delay time in a fiber (Chapter 2); recombination time of a charge carrier; lifetime (τ_e for electrons, τ_h for holes); effective carrier recombination time (Chapter 4); time constant
τ_g	group delay time per unit length, $\tau_g = 1/v_g$ (Chapter 2); mean time to thermally generate an electron hole pair in a semiconductor (Chapter 3)
τ_h, τ_e	lifetime of holes and electrons in a semiconductor; average recombination time of holes and electrons
τ_{nr}	nonradiative lifetime; $1/\tau_{\text{nr}}$ = rate of nonradiative recombinations
τ_{ph}	photon lifetime in the optical cavity or time it takes for the photon to be lost from the cavity
τ_R, τ_F	rise and fall time of the light output pulse from an LED
τ_r	radiative lifetime; $1/\tau_r$ = rate of radiative recombinations
τ_{sp}	average time for an electron hole pair recombination by spontaneous emission
$\tau_t, \tau_{th}, \tau_{te}$	effective transit time of minority carriers in a <i>pn</i> junction type device; τ_{th} for holes in the <i>n</i> -side and τ_{te} for electrons in the <i>p</i> -side.
v	frequency
v_m	mode frequency
v_p	pump frequency
v_o	central or peak frequency
Φ	phase change; work function; photon flux

Φ_B	Schottky barrier height
Φ_m	metal work function
Φ_p	work function for a <i>p</i> -type semiconductor
Φ_{ph}	photon flux in a given direction
Φ_{pin}	input pump photon flux
Φ_n	work function for an <i>n</i> -type semiconductor
Φ_{sin}, Φ_{sout}	signal photon flux at input and output
Φ_v	luminous flux (lumens, lm)
ϕ	angular spherical coordinate; phase ($\phi = \omega t - kz + \phi_o$); phase change
ϕ_o	phase constant
χ	electron affinity; electric susceptibility
χ_1, χ_2, χ_3	linear, second-order, and third-order electric susceptibilities
$\psi(x)$	spatial dependence of the electron wavefunction under steady state conditions satisfying the Schrödinger equation
$\psi_k(x)$	Bloch wavefunction; electron wavefunction in a crystal with a wave vector k
Ω	phonon angular frequency solid angle
ω	angular frequency ($2\pi\nu$) of light

INDEX

A

Absorption, 59
 coefficient, 60, 388
 graph vs. wavelength, 389
 indirect, 390
 lattice, 161
Acceptance angle, fibers, 133
Acceptors in semiconductors, 205
Acousto-optic effect, 492
 figure of merit and table, 497
Acousto-optic modulator, 493, 497
 Bragg regime and diffraction, 495, 511
 Raman–Nath regime, 493
Active Matrix Array, 433–434
Active pixel sensor, 434
Air mass, 442
Airy disk, 78
 angular radius, 78
Airy, George, 74, 107
Airy rings, 74, 75, 77
Alloys, semiconductors
 III-V alloys, 253, 273
 III-V bandgap diagram, 171
 III-V ternary alloys, 253
Amorphous structure broadening, 321
Amplified spontaneous emission (ASE),
 301, 303
Amplifier, *see* Optical amplifier
Anisotropy, optical, 27, 462, 463
 refractive index, 27, 461
 table of crystals, 462
Antireflection coating (AR), 55–56,
 381, 437
Argon-ion laser, 375
Attenuation, 59
 coefficient, 59, 159
 optical fibers and dB, 159–167
 table for fibers, 163
 vs. bend radius, 166
Avalanche, 403
 breakdown voltage, 404
 multiplication factor, 404
 noise, 430
Avalanche photodiode (APD), 402
 absorption region in APD, 409
 guard ring, 404
 internal gain, 404
 noise, 430
 photogeneration, 409
 primary photocurrent, 404, 406
 principle, 417
 responsivity of InGaAs, 406
 separate absorption
 and multiplication (SAM), 409–411
 grading and multiplication (SAGM), 409
Si reach-through, 402
silicon, 406
superlattice, 411

B

Bandgap, energy gap, 198
 III-V bandgap diagram, 273
Bandwidth, 146
 optical, electrical, 150–151
Bandwidth theorem, 65
Bardeen, John, 194
Bayer color image sensor, 433
Beam
 circular cross-section, 192
 diameter, 23
 displacement, *see* Lateral displacement
 of light
 divergence, 24
 Gaussian, 78–79, 305, 308
 light, 22
 self-focused, 512
 splitter cube, 51
Becker, P. C., 292, 479
Bending losses, 164
Bennett, William Jr., 306
Biaxial crystals, 462
Birefringence, 461
 calcite, 466
 circular, 472
 induced, 480
 quartz, 469
Birefringent
 circularly, media, 474
 crystal, 461
 optical devices, 468
 prisms, 472
 retarding plates, 468
Bit rate, 146, 151
 capacity, 146
 intersymbol interference, 147
 nonreturn-to-zero (NRZ), 151
 return-to-zero (RZ), 147
Black body radiation law, 287
Bloch wave in a crystal, 94, 211
Boundary conditions, electromagnetism, 42, 43
Bragg angle, 495
Bragg diffraction condition, 83, 276, 301, 511
Bragg fibers, 175–176
 gratings and sensors, 179–183
Bragg reflector, 57
Bragg wavelength, 179, 355
Bragg, William Lawrence, 81, 83
Brewster angle, 45, 46
 reflection and transmission, 53–54
Brewster, David, 45
Brewster window, 305, 319
Broadening
 amorphous structure, 321
 collision, 320
 homogeneous, 320
 inhomogeneous, 321
 lifetime, 320

- Broadening (*continued*)
 optical gain curve, 319
 pressure, 320
 Burrus device, 263
- C**
- Calcite, 462, 466
 principal section, 466
 rhomb, 466
 specific rotary power, 473
 Carrier confinement, 331
 Cauchy equation and diamond, 30
 Cavity lifetime, 316–317, 353
 Centrosymmetric crystals, 479
 Charge-coupled device (CCD), 435–437
 Frame architecture, 436–437
 full frame architecture, 436–437
 Interline transfer architecture, 436–437
 sensor, 432
 Chirp, chirping, 174, 512
 Chromatic dispersion, 141–142
 Cladding, 111
 attenuation, 118
 layers, 263
 Coefficient of index grading, 153
 Coherence
 length, 64
 mutual temporal, 66
 perfect, 63
 spatial, 63, 66
 temporal, 63
 time, 64
 Collision broadening, 320
 Compensation doping, 206
 Complementary metal oxide semiconductor (CMOS)
 sensor, 432, 433
 Complex propagation constant, 60
 Complex refractive index, 61
 of InP, 62–63
 Complex relative permittivity, 60
 Conduction band (CB), 198, 212
 electron concentration, 204
 Conductivity, semiconductors, 204
 Confining layers, 254, 263
 Conjugate image, 372
 Constructive interference, 68
 Coupled waveguides, 489
 Coupling coefficient, 181
 Cross-phase modulation, 174
 Crystal momentum, 212
 Current-Voltage Convention, 385–386
- D**
- Degenerate semiconductor, 207
 Dense wavelength division multiplexing (DWDM), 172
 nonlinear effects, 173–175
 Density of states (DOS), 196
 effective, 201
 Depletion region (layer), 215, 381
 capacitance, 227, 396
 Depth of focus, 24
- Destructive interference, 68
 Desurvire E., 292
 Detectivity, 427
 Dichroism, 467
 Dielectric mirrors, 56, 58, 360
 Diffraction, 74
 Fraunhofer, 74, 75
 Fresnel, 74
 diagram, 75
 order, 355
 pattern, 74
 circular aperture, 78
 rectangular aperture, 77
 single slit, 76
 principles, 74
 Diffraction grating, 80
 blazed (echellette), 83
 reflection, 81
 transmission, 81
 Diffraction grating equation, 81
 single slit, 77
 Diffusion (storage) capacitance, 230
 Diffusion current, 220
 Diffusion flux, 219
 Diffusion length, 219
 Diode equation, 224
 Diode ideality factor, 224
 Direct recombination, 231
 Dispersion, 141, 151
 bandwidth relations, 151
 chromatic, 156, 178
 coefficients, 156
 diagram, 120
 graph of ω vs. β , 122
 intermode, intermodal, 121, 126
 intramode, intramodal, 122
 material, 123, 135, 140, 187
 coefficient, 136
 graph vs. wavelength, 137
 modal, 120, 121
 polarization mode, 139
 profile, 139, 188
 profile-coefficient, 119
 relation, 28
 single mode fiber, 135
 table, 144
 total, 137, 141
 vs. wavelength, 137, 141, 143
 Dispersive medium, 32
 Distributed Bragg reflector, 181, 257, 354, 360
 Donors in semiconductors, 206
 Doppler broadened linewidth, 307, 309
 Doppler effect, 306
 Doppler shift, 496
 Double-heterostructure (DH), 242
 active layer, 331–332
 buried, 334
 contacting layer, 333
 device, 331
 Drift mobility, 204, 397
 Drift velocity, 271, 397–398
 graph vs. field, 398

E

EDFA, *see* Erbium doped fiber amplifier

Effective density of states, 201, 202

Effective mass, 199

Efficiency

detector, 391

fiber coupling efficiency, 277

He-Ne laser, 306

laser diode external power, 343

laser diode slope, 343

optical amplifier, 295

quantum efficiency, 343

external, 340, 392

external differential, 343

internal, 277, 344

Einstein coefficients, 286, 287

E vs. k diagrams, 210

direct bandgap, 212

indirect bandgap, 213

Electrical bandwidth of fibers, 146, 149, 269

Electromagnetic (EM) wave, *see* Wave

Electron affinity, 198

Electron ionization coefficient, 407–408

Electro-optic effects, 478

Kerr, 485

linear, Pockels, 479

Energy band, 195, 196

Energy band diagram, 196, 208, 383, 385

field applied, 208

GaAs laser, 377

heterojunction, 242, 333

pn junction, 236

Energy level, 195

long-lived, 293

Epitaxial growth, 254

Erbium doped fiber amplifier (EDFA), 292, 299

characteristics, efficiency and gain saturation, 296

energy diagram, 293

gain-flattened, 300–303

Excess carrier distribution, 231

External photocurrent, 386–388

time spread, 388

External quantum efficiency (EQE), 258

Extinction coefficient, 61

Extraction efficiency (EE), 259

Extraordinary wave, 463

Extrinsic semiconductors, *see* Semiconductors

F

Fabry, Charles, 69

Fabry–Perot interferometer, 84, 85

Fabry–Perot laser amplifier, 364

Fabry–Perot optical resonator, 70, 308

transmitted light, 72

Fall time, 269

Faraday effect (rotation), 499, 511

Faraday, Michael, 123

Fermi–Dirac function, 195, 197, 200, 244, 329

Fermi energy, 195, 203

Fermi level, 201, 202

Fiber, *see* Optical fibers

Fiber Bragg grating (FBG), 179

Fill factor (FF), 440; *see also* Solar cell

Finesse, 72

First Brillouin zone, 94

Flat-band voltage, *see* Reach-trough voltage

Four-level laser system, 285–286

Four photon mixing, 173

Franken, Peter A., 502

Franck–Condon principle, 265

Fraunhofer, Joseph, 75

Free carrier absorption, 271

Free spectral range, 70

Fresnel, Augustin Jean, 17

Fresnel prism, 508

Fresnel's equations, 42, 43, 104

Fresnel's optical indicatrix, 463

Full width at half maximum, *see* FWHM

Full width at half power, *see* FWHP

FWHM, 147

of a gas laser, 307

FWHP, 147

G

Gabor, Dennis, 370

Gain coefficient in Nd³⁺ doped glass fiber, 291

Gain saturation, 298

Gaussian beam, 37, 99, 308, 318

power and irradiance, 37

Gaussian broadening, 321

Gaussian dispersion, 150–151

Gaussian pulse, 149

Glan–Focault prism, 508

Glass preform, 170

Goos Haenchen phase shift, 104

Graded index fiber, 152–157; *see also* Optical fiber

dispersion and bit rate, 157

properties – table, 156

Graded index (GRIN) rod lens, 151, 155, 264

Group delay, 136

Group index, 30

graph vs. wavelength, 31

Group velocity, 31, 120, 130–131, 185

H

Half-wave plate retarder, 469

quartz, 469

Half-wave voltage, 482

Helical ray path, 124, 153

Helium–Neon laser, 303, 315, 317

characteristics – table, 374

efficiency, 306

energy diagram, 304, 374

modes, 307

principle of operation, 304

Herriott, Donald, 306

Heterojunction, 238

photodiodes, 409

phototransistor, 418

Heterostructure, 241, 242

device, 242, 261, 333, 337, 364

High (strong) injection, 230–232, 272

- Hole, 199
 diffusion current, 219
 diffusion length, 219
 ionization coefficient, 407
- Holey fiber, 160; *see also* Photonic crystal fiber
- Hologram, 370
- Holography, 370, 372
- Homogeneous broadening, 320
- Homojunction, 240, 329
- Huygens, Christiaan, 456
- Huygens-Fresnel principle, 75–76
- I**
- Image sensors, 431–437
 pixels, 431
- Impact-ionization, 403
- Index matching, 503
- Induced transition, *see* Stimulated emission
- InGaAsP on InP substrate, 335
- Inhomogeneous broadening, 321
- Injection, 217
 of excess minority carriers, 217
 pumping, 329
 strong, 231
 weak, 231
- Instantaneous irradiance (intensity), 35
- Integrated optics, 486
- Intensity of light, 37, 48
- Interference, 67
 incoherent, 89
- Interference fringes, 87
- Interferometers, 84–86
- Internal quantum efficiency (IQE), 258
- Intersymbol interference, 147
- Intrinsic concentration n_i , 202, 205
- Intrinsic semiconductors, *see* Semiconductors
- Irradiance, 34, 35
 reflected light, 48
- Isoelectronic impurities, 253
- Isotropic, 27
- J**
- Javan, Ali, 306
- Jones matrices, 506
- Jones units, 427
- Jones vector, 506
- K**
- Kao, Charles, 110
- Kerr coefficients – table, 485
- Kerr effect, 484–486
- Kerr effect modulator, 486
- Kerr, John, 478
- Kramers-Kronig relations, 62
- L**
- LASER, 265, 266–269; *see also* Laser diode
 active medium, 305
 active region, 328, 334
 amplifiers, 364
 distributed Bragg reflection (DBR), 354–355
 distributed feedback (DFB), 354, 355
 efficiency of the He-Ne, 306
- gain guided, 334
- gas, 303
 modes, 308
- He-Ne, 303, 315
 characteristics – table, 374
 efficiency, 306
 principle of operation, 304
- modes, 308, 310, 317, 318
- multiple quantum well (MQW), 338
- Nd³⁺:YAG laser, 285
- optical cavity, 284, 303, 313
- oscillation conditions, 311
- oscillator, 314
- output spectrum of a gas, 306
- output wavelength variations, 346
- principles, 282
- ruby laser, 283
- semiconductor, 364
- single frequency solid state, 354
 guiding layer, 355
- threshold gain, 312
- vertical cavity surface emitting (VCSE), 360
- Laser diode (LD)
 active region, 334
 buried heterostructure, 335
 characteristics, 340
 conversion efficiency, 344
 direct modulation of, 367–370
 distributed Bragg reflector (DBR), 354
 distributed feedback (DFB), 355
 divergence of output, 341
 double heterostructure, 348
 edge emitting, 334
 equation, 348
 external cavity, 358
 external power efficiency, 378
 Fabry-Perot, 334
 gain guided, 334
 heterostructure, 331
 homojunction, 331
 index guided, 335
 inversion layer, 328
 modes, 336
 modulation, 367
 principles, 327
 rate equations, 348
 single frequency, 354
 single mode, 335
 slope efficiency, 343
 stripe geometry, 333
 transient response, 368
 delay time, 369
 relaxation oscillation, 368
- tables, 345, 356
- threshold current, 330, 331
- transparency current, 330
- vertical cavity, surface emitting, 360
- Lasing emission, 283
- Lateral displacement of light, 41
- Lattice vibrations, 511
- LED, *see* light emitting diode, LED
- Lifetime broadening, 320
- Light absorption, 59

Light emitting diode, LED, 195, 240
 active region, 263
 AlGaAs, 272
 characteristics, 261–262
 cut-in voltage, 261
 edge emitting (ELED), 263–264, 277
 efficiencies and luminous flux, 258–260
 electronics, 267–269
 energy band diagram, 240
 external quantum efficiency (EQE), 258
 fiber coupling efficiency, 277
 for optical communications, 267
 heterojunction, 262
 high intensity, 249
 internal quantum efficiency (IQE), 258
 linewidth, 242, 245
 luminous efficacy, 260
 materials, 253–254
 optical fiber communications, 262–264
 output spectrum, 262
 phosphors and white, 265–267
 power conversion efficiency, 259
 principles, 240
 structure, 254–257
 superluminescent and resonant cavity, 366–367
 surface emitting (SLED), 263, 277
 turn-on voltage, 161
 wavelengths, 257
 white, 265
 Lineshape function, 312
 Linewidth, 140
 Liquid crystal display (LCD), 474–478
 twisted nematic field effect, 475
 Lithium niobate (LiNbO_3)
 acousto-optic, 492
 phase modulator, 487, 509
 Pockels effect, 479
 properties, 463
 Load line, 440
 Longitudinal axial modes, 318
 Lorentzian lineshape, 320
 Loss coefficient, 313
 Luminescence, 265
 Luminosity function, 259
 Luminous
 efficacy of light source, 260
 flux, 259

M

Mach–Zehnder interferometer, 85
 Mach–Zehnder modulator, 86, 489
 Macrobending loss, 164
 Magneto-optic effects, 499
 Magneto-optic isolator, 511
 Magneto-optic modulator, 511
 Maiman, Theodore Harold, 379
 Majority carriers, 206
 Malus's law, 460
 Mass action law, 202
 Matrix emitter, 363
 Maximum acceptance angle, 133
 Maxwell's wave equation, 22, 26
 Mean thermal generation time, 226

Meinel and Meinel equation, 454
 Metalsemiconductor-metal (MSM) structure, 416
 photodiodes, 416
 Meridional ray, 124–125, 153
 Microbending loss, 164
 graph vs. bend radius, 166
 Microlaser, 363
 Midwinter, John, 50
 Mie scattering, 92
 Minority carriers, 206
 diffusion, 219
 diffusion length, 393
 Modal index, 179
 Mode
 cavity, 71, 308, 330
 coherently coupled, 356
 field diameter (MFD), 128
 hop in lasers, 342
 intensity of transmitted, 73
 linearly polarized, 125
 number, 115, 308
 of propagation, 115
 resonator, 73
 TE, 116, 184–185, 319
 TM, 116, 184–185, 319
 transverse, 319
 Mode field diameter (MFD), 128–129
 Mode locking, 326–327
 Modulated directional coupler, 492
 Modulation of light, 457–513
 Monochromatic wave, 30
 MQW, *see* Multiple quantum well (MQW)
 Multimode fiber, 128
 Multiple interference, 77
 Multiple quantum well (MQW), 252, 338
 Multiple reflections
 in plates and incoherent waves, 89–90
 thin films, 86
 Multiplication region, 409
 Mutual temporal coherence, 67

N

Natural broadening, 320
 Negative absolute temperature, 288
 Net round-trip optical gain, 313
 Noise
 current, 425
 equivalent power (NEP), 426
 NEP graph, 428
 excess avalanche, 453
 excess, 403
 noise figure, 302
 in an APD, 430
 in photodetectors, 424
 of an ideal photodetector, 428
 photon noise, 425
 quantum, 425, 429
 shot, 425
 Noncentrosymmetric crystals, 479
 Nondegenerate semiconductors, 207
 Nonlinear optics, 501
 Nonlinear effects, 173, 478, 502
 Non-thermal equilibrium, 288

Normalized frequency, 116, 126
 Normalized index difference, 124, 126
 Normalized propagation constant, 130
 Normalized thickness, 116
 Numerical aperture, 133, 134, 154

O

Optical activity, 472, 500
 dextrorotatory, 473
 levorotatory, 473
 Optical amplifier, 283
 erbium doped fiber amplifiers (EDFA), 292
 semiconductor optical amplifier, 364
 table, EDFA, 302
 Optical anisotropy, 461–462
 Optical bandwidth, 146, 150, 268
 Optical cavity, 288, 303
 intensity, 71
 maximum intensity, 71
 modes, 70
 in lasers, 317
 resonant frequencies, 70, 330
 Optical divergence, 23
 Optical confinement, 331
 Optical fiber, 111
 all wave fiber, 162
 attenuation, 158, 160
 cutoff wavelength, 127
 doubly clad, 143
 dispersion flattened, 143
 dispersion shifted, 138
 drawing, 168–171
 schematic, 168
 fundamental mode, 125
 graded index, 151–158
 light emitters, 262
 MAC number, 166
 manufacture, 168
 laydown stage, 169
 multimode, 127, 134, 186
 nonzero dispersion shifted, 143
 propagation constant, 130
 propagation of light, 125
 reduced slope, 142
 sensor, 182
 single mode, 127, 128, 131, 186
 single mode cutoff, 127
 step index, 123
 types of, 143
 V-number, 125
 zero dispersion shifted, 142
 Optical fiber amplifier, 292
 erbium ion doped, 292
 gain efficiency, 297
 Optical field, 20
 Optical frequencies, 27
 Optical gain, 267, 275, 307
 coefficient, 311–312
 lineshape, 307
 net round-trip, 313
 threshold, 313
 Optical indicatrix, 464
 Optical isolator, 500

Optical laser amplifiers, 364
 Optical modulators
 coupled waveguide, 489
 integrated, 486
 intensity, 483
 Kerr effect, 486
 Mach-Zehnder, 489
 Pockels cell, 484
 Optical pumping, 283–284
 Optical resonator, 69
 Optical tunneling, 49, 50
 Optically isotropic, 27, 461
 Optic axis, 462
 Optimal profile index, 154
 Output power and photon lifetime in the cavity, 315–316
 Output spectrum
 LED, 242
 gas laser, 306
 laser diode, 331, 342–343
 white LED, 265
 Outside vapor deposition, 168, 169, 170
 schematic, 169

P

Passive pixel sensor, 434
 Penetration depth, 47, 389
 Perot, Alfred, 69
 Phase change, 44
 in TIR, 46, 50
 Phase condition in lasers, 317
 Phase matching, 503
 Phase matching angle, 503
 Phase mismatch, 491
 Phase modulation, 486
 Phase modulator, 481
 LiNbO₃, 484
 Phase of a wave, 19
 Phase velocity, 21
 Phonon, 213, 391, 511
 energy and momentum, 391
 Phosphors, 265
 Photoconductive detectors, 418
 Photoconductive gain, 418
 Photoconductivity, 420
 Photocurrent, 382, 369, 422
 Photodetection Modes, 383–385
 Photodetector, 381
 neutral region, 382
 quantum efficiency, 391
 Shockley–Ramo theorem, 388
 responsivity, 391
 Photodiode, 381, 399
 circuits, 421–424
 materials, 389
 photocurrent, 386
pn junction, 381
 diagram, 382
 responsivity, 391
 space charge layer (SCL), 382
 spectral responsivity, 391
 table, 389
 Photoelastic effect, 175, 492
 Photogeneration, 382–383

- Photon amplification, 281–282
 Photon cavity lifetime, 316; *see also* Cavity lifetime
 Photon confinement, 335
 Photon flux, 258
 Photon lifetime, *see* Cavity lifetime
 Photonic bandgap, 93, 94
 guided, 178
 Photonic crystal fiber (PCF),
 176–179
 Photonic crystals, 92–98
 Phototransistor, 417
 heterojunction, 418
 Photovoltaic devices, *see* Solar cell
 Photovoltaic mode of operation, 386
 Piezoelectric effect, 498
p-i-n photodiode, 395, 396
 characteristics – table, 405
 junction capacitance, 397
 NEP of a Si, 428
 photocarrier diffusion, 399
 response time, 397
 responsivity, 400
 speed, 399
 steady state photocurrent, 401
 transient photocurrents, 448
 transit time, 397
 Pixel image sensor, 434
 Planck's radiation distribution law, 287
p-n junction, 214
 band diagram, 334
 graph, 335
 built-in field, 216
 built-in potential, 216, 217
 current density, 226, 148
 depletion region, 215
 diffusion (storage) capacitance, 230
 direct bandgap, 152
 dynamic (incremental) resistance, 229
 energy band diagram, 236
 forward bias, 221, 226
 currents, 222
 diagram, 217
 recombination current, 222
 law of the junction, 218, 220
 metallurgical junction, 214
 properties – graph, 214
 reverse bias, 225, 236
 diagram, 225
 total current, 226
 reverse saturation, 226
 space charge layer (SCL), 215, 227
 Pockels cell modulator, 484
 longitudinal, 481
 transverse, 481
 Pockels coefficients – table, 485
 Pockels effect, 478
 transverse, 481
 Pockels, Friedrich Carl Alwin, 479
 Pockels phase modulator, 481
 Point defect, 97
 Polarization
 angle, 29; *see also* Brewster angle
 anisotropic, 507
 circular, 458
 right, 258
 dispersion, 138
 dispersion effects, 138
 elliptical, 459
 induced, 502
 linear, 459
 modulation, 486
 modulator, 482
 of EM wave, 457, 460
 state, 457
 Polarization transmission matrix, 506
 Population inversion, 281, 288, 328
 threshold, 313
 Poynting vector, 34–38
 Preform, 168
 Pressure broadening, 320
 Prism, 41, 60
 birefringent, 471
 Profile effects–fibers, 138
 Profile index, 153
 optimal, 153
 Propagation constant, 19, 130
 complex, 60
 Propagation vector, 21
 Pseudo photonic bandgap (PBG), 95
 Pulsed Lasers, 323–327
 Pumping, 283, 280
 erbium doped fiber amplifier (EDFA), 296
 gas discharge (collision), 304
 injection, 329
 Pyroelectric detectors, 381
- Q**
- Q*-Switching, 323–325
 Quantum efficiency, 391
 detector, 391
 external differential, 343
 external, 343, 391
 internal, 258, 328
 Quantum noise, 425
 Quantum well, 249
 devices, 249
 GaAs, 339–340
 high intensity LEDs, 249–252
 laser diode, 337
 Quaternary alloy, 253
 Quarter-wave plate retarder, 469
 Quartz, 462, 469
- R**
- Radiant flux, 258
 Radiant sensitivity, 392
 Raman–Nath regime, 493
 Ramo's theorem, *see* Shockley–Ramo theorem
 Rate equations, laser diodes, 348
 Rayleigh criterion, 79
 Rayleigh range, 24
 Rayleigh scattering, 91
 Rayleigh, John William Strutt, 91
 Rayleigh scattering limit, 162
 Reach-trough voltage, 416
 Real image, holography, 371, 372

- Recombination, 199
 center, 213
 current, 224
 direct, 230
 direct capture coefficient, 231
 indirect, 232
- Recombination lifetime, 230
 definition, 230
 excess minority carrier, 231
 mean, 222
 weak injection, 231
- Reflectance, 48, 71
 bandwidth, 57
 at normal incidence, 48
- Reflection
 amplitude, 42
 at normal incidence, 46, 49
 coefficients, 43
 graph, 47
 external, 46, 52
 at normal incidence, 46
 frustrated total internal, 50, 51
 internal, 46, 52
 graph, 45
 multiple reflections, 86
 and transmission at the Brewster angle, 53
 total internal, 38, 39, 40, 46, 50
- Refracted light, 38
 from a less dense medium, 51
- Refractive index, 27
 complex, 61
 dispersion, 26
 field induced change, 478
 graph vs. wavelength, 27
 nonlinear, 174
 optical crystals, 462
 table, 28
- Resolving power
 angular limit of resolution, 79
 diffraction grating, 82
 imaging systems, 79
- Response time, 397
- Responsivity, 392
 graphs, 393, 398
 table, 405
- Reststrahlen absorption, 62
- Retarding plates, 468
- Return-to-zero (RTZ) data
 rate, 147
- Rise time, 269
- S**
- Saturation drift velocity, 400
- Scattering
 anisotropic, 507
 light, 161, 313
 Rayleigh, 90, 91
- Schottky barrier height, 414
- Schottky-junction photodetector, 413–416
- SCL, *see* Space charge layer, 215
- Second harmonic generation, 501
- Self-phase modulation, 174
- Sellmeier and Cauchy coefficients, 29
- Sellmeier dispersion equation, 13, 99
- Semiconductors, 195–213
 charge neutrality, 215
 compensation doping, 206
 conductivity, 204
 degenerate, 207
 direct bandgap, 210–213, 232
 extrinsic, 203
 III-V compound, 273
 indirect bandgap, 210, 213
 intrinsic, 202
 non-degenerate, 201, 207
n-type, 203
 optical amplifier, 364–365
p-type, 203
 rate equations, 348
 statistics, 200
- Shockley–Ramo theorem, 386, 388
- Shockley equation, 220
- Shot noise, 425
- Shunt resistance, 423
- Side mode suppression ratio, 355
- Silica (SiO_2)
 Germania ($\text{SiO}_2\text{-GeO}_2$), 101, 169
 refractive index, 32, 101, 169
- Signal to noise ratio (SNR), 302, 426
 of a receiver, 429
- Simpson, J. R., 292
- Single frequency lasers, 354
 cleaved coupled cavity, 356
 distributed Bragg reflection (DBR), 354
 distributed feedback (DFB), 355
- Single-mode fiber, 128
- Single quantum well (SQW), 337
 energy levels, 337
 laser, 337
- Skew ray, 124
- Snell's law, 38–41
- Snell, Willebrord van Roijen, 38
- Solar cell, 437–444
 antireflection coating, 55
 equivalent circuit, 441
 properties – table, 442
- Solar constant, *see* Air-mass
- Soleil–Babinet compensator, 470, 471
- Solid state photomultiplier, 412
- Space charge layer (SCL), 199; *see also* Depletion region
 width and voltage, 228
- Specific rotatory power, 473
- Spectral hole burning, 323
- Spectral intensity, 274

Spectral responsivity, 392
 Spectral width, 72, 73
 of a wave train, 64
 Spontaneous emission, 282
 Spot size, 23
 SQW, *see* Single quantum well (SQW)
 Step-index fibers, 123–132
 Stimulated Brillouin scattering (SBS), 174
 Stimulated emission, 281, 282
 rate, 286
 Stokes shift, 266
 Stop band, 57
 Strong injection, *see* High injection
 Superlattice, 411
 Superposition of waves, 67

T

Tandem solar cell, 443
 Terminal capacitance, 423
 Ternary alloy, 253
 Thermal equilibrium, 287
 Thermal generation, 199
 in SCL, 226
 Thermal velocity, 202
 Thin films
 multiple reflections, 86
 optics, 86, 88–89
 Three-level laser system, 283
 Threshold concentration, 349
 Threshold current, 342, 253, 350
 Threshold wavelength, 388
 Total acceptance angle, 133, 134
 Total internal reflection, 46, 49, 50
 critical angle, 40
 Townes, Charles D., 288
 Transfer distance in coupled waveguides, 490
 Transit time, 397
 Transmission axis, 460
 Transmission coefficient, 42–44
 Transmittance, 48
 at normal incidence, 49
 Transverse electric field (TE), 42
 Transverse magnetic field (TM), 42
 Trench fiber, 166
 Truncated spherical lens, 263
 Turbidity, 92
 Twisted nematic liquid crystal cell, 475
 Tyndall, John, 123

U

Uniaxial crystals, 462
 negative, 462
 optic axis, 462
 positive, 462
 Unpolarized light, 45

V

Valence band (VB), 198
 hole concentration, 202
 Verdet constant, 499
 table, 501
 Vertical cavity surface emitting laser (VCSEL), 360
 Virtual image holography, 370, 371
 Visibility function, 259
 V-number, 116, 118, 120
 V-parameter, 116

W

Waist, beam, 23
 radius, 23
 Wave
 circularly polarized, 458, 469
 diverging, 22
 electromagnetic, 19
 elliptically polarized, 459
 energy density in an EM, 34
 evanescent, 40, 47
 attenuation, 47
 extraordinary, 463
 fields in an EM, 34
 linearly polarized, 45, 125, 292
 monochromatic plane, 19
 ordinary, 463
 plane electromagnetic, 19
 plane-polarized, 457
 sinusoidal, 64
 spherical, 23
 stationary or standing EM, 70
 Wave equation, 20, 22
 Wavefront, 19
 Gaussian light beam, 24
 Wavefront reconstruction, 372
 Waveguide, 111
 condition, 111–115
 cutoff wavelength, 116
 dielectric, 111–183
 dispersion, 120–123, 141, 187
 coefficient, 137
 diagram, 120
 modes, 117
 mode determination, 118
 multimode, 116
 planar, 120
 propagation constant, 113
 single mode, 116
 symmetric dielectric slab, 111
 diagram, 112
 possible modes, 114
 propagation constant, 114
 transverse propagation, 113
 Wavelength division multiplexing (WDM), 171

- Wave number, 19
Wave packet, 30
Wave vector, 21
 surface, 465
Weakly guiding fiber, 125
Wire grid polarizer, 507
- Wollaston prism, 471, 472, 508
Wood, Robert William, 82

Y

- Young's two slit experiment, 68

I don't really start until I get my proofs back from the printers. Then I can begin serious writing.
—JOHN MAYNARD KEYNES (1883–1946)¹

¹The *Guardian*, 8 June 1983, p. 22 as cited in A. L. Mackay, *Dictionary of Scientific Quotations*, Institute of Physics Publishing, Bristol, 1991, p. 140.

Visible Spectrum

Typical and approximate wavelength ranges, and color perception by an average person

Color	Violet	Blue	Green	Yellow	Orange	Red
λ (nm)	390–455	455–492	492–577	577–597	597–622	622–780

LED Colors

Typical wavelength ranges and colors as usually specified for LEDs

(Manufacturers' labeling can differ significantly)

Color	Violet	Blue	Aqua green	Emerald green	Green	Yellow	Amber	Orange	Red-orange	Red	Deep red	Infrared
λ (nm)	400–450	450–490	490–524	525–564	565–579	580–587	588–594	595–606	607–615	616–632	633–700	>700

Greek Alphabet

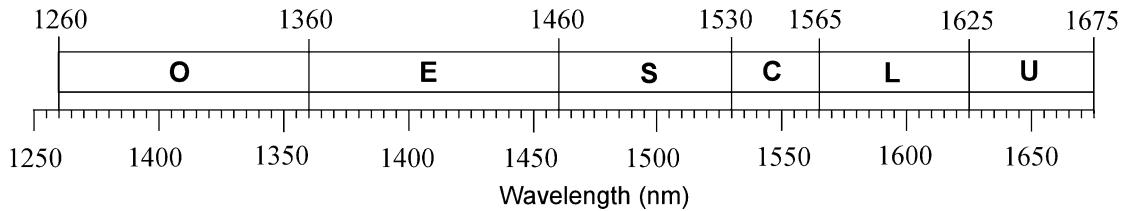
A, α	Alpha	I, ι	Iota	P, ρ	Rho
B, β	Beta	K, κ	Kappa	Σ , σ	Sigma
Γ , γ	Gamma	Λ , λ	Lambda	T, τ	Tau
Δ , δ	Delta	M, μ	Mu	Y, ν	Upsilon
E, ε	Epsilon	N, ν	Nu	Φ , ϕ	Phi
Z, ζ	Zeta	Ξ , ξ	Xi	X, χ	Chi
H, η	Eta	O, \o	Omicron	Ψ , ψ	Psi
Θ , θ	Theta	Π , π	Pi	Ω , ω	Omega

Equivalent Power Levels in dBm in Optical Communications

Power	100 pW	1 nW	10 nW	100 nW	1 μW	10 μW	100 μW	1 mW	10 mW	100 mW	200 mW	500 mW	1 W
dBm	-70	-60	-50	-40	-30	-20	-10	0	10	20	23	27	30

Bands in Optical Communications

O-band (Original)	E-band (Extended)	S-band (Short wavelengths)	C-band (Conventional)	L-band (Long wavelengths)	U-band (Ultralong wavelengths)
1260–1360	1360–1460	1460–1530	1530–1565	1565–1625	1625–1675



Physical Constants

c	Speed of light in vacuum	$2.9979 \times 10^8 \text{ m s}^{-1}$
h	Planck's constant	$6.6261 \times 10^{-34} \text{ J s}$
\hbar	$\hbar = h/2\pi$	$4.1357 \times 10^{-15} \text{ eV s}$
e	Electronic charge	$1.0546 \times 10^{-34} \text{ J s}$
ε_0	Absolute permittivity	$6.5821 \times 10^{-16} \text{ eV s}$
k_B	Boltzmann constant ($k_B = R/N_A$)	$1.60218 \times 10^{-19} \text{ C}$
$k_B T/e$	Thermal voltage at 300 K	$8.8542 \times 10^{-12} \text{ F m}^{-1}$
m_e	Electron mass in free space	$1.3807 \times 10^{-23} \text{ J K}^{-1}$
μ_0	Absolute permeability	0.02585 V
N_A	Avogadro's number	$9.10939 \times 10^{-31} \text{ kg}$
R	Gas constant ($N_A k_B$)	$4\pi \times 10^{-7} \text{ H m}^{-1}$
		$6.0221 \times 10^{23} \text{ mol}^{-1}$
		$8.31457 \text{ J mol}^{-1} \text{ K}^{-1}$

Useful Information

$$\pi = 3.1416 \quad e = 2.7183$$
$$1 \text{ \AA (Angstrom)} = 0.1 \text{ nm} = 10^{-10} \text{ m} \quad 1 \text{ eV} = 1.60218 \times 10^{-19} \text{ J}$$

Common Prefixes for Multiples of Ten

10^{-15}	10^{-12}	10^{-9}	10^{-6}	10^{-3}	10^{-2}	10^3	10^6	10^9	10^{12}
f	p	n	μ	m	c	k	M	G	T
femto	pico	nano	micro	milli	centi	kilo	mega	giga	tera
