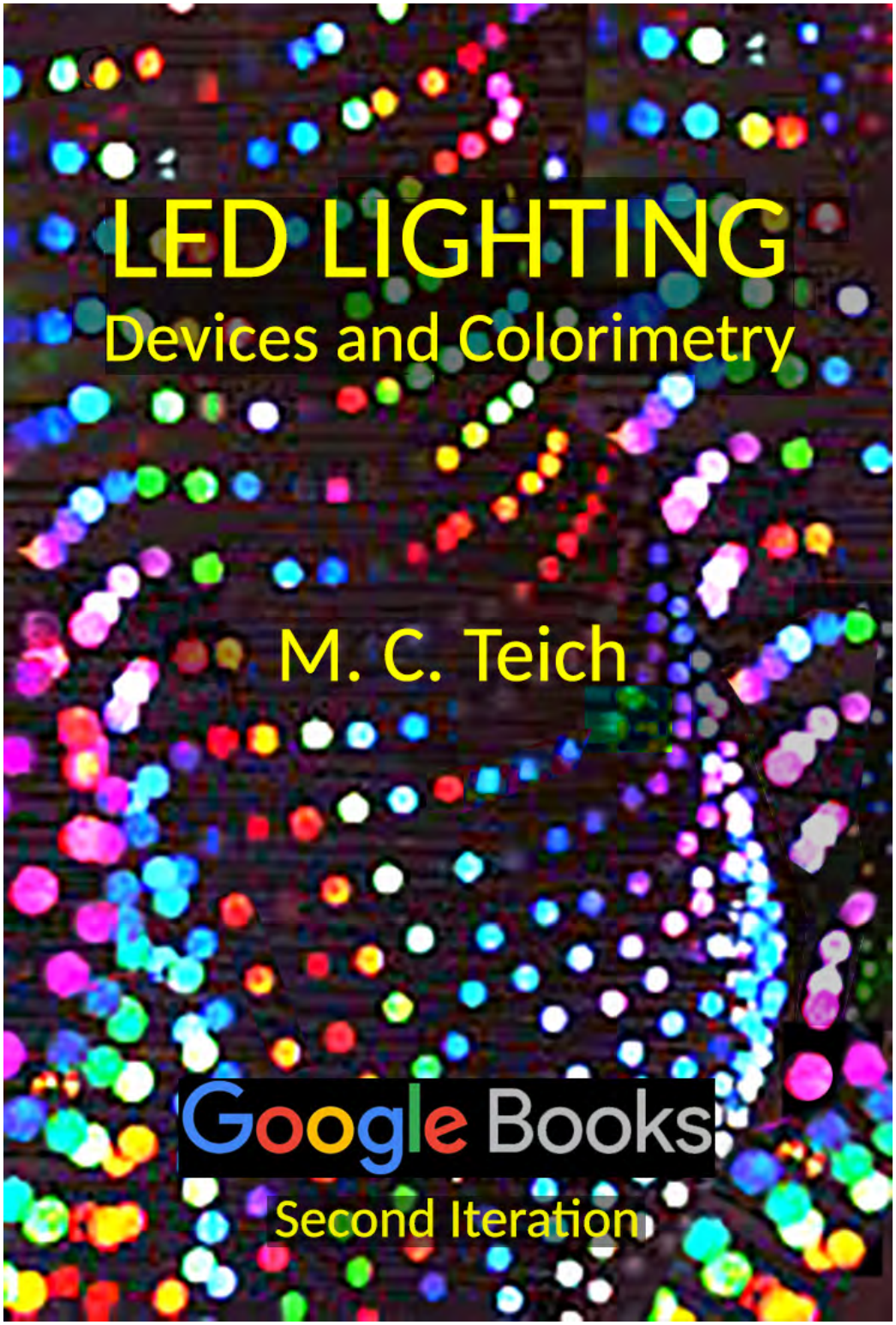

LED LIGHTING
Devices and Colorimetry
M. C. Teich
Google Books
Second Iteration

LED LIGHTING

Devices and Colorimetry

LED LIGHTING

Devices and Colorimetry

MALVIN CARL TEICH
Columbia University
Boston University

LED LIGHTING

Devices and Colorimetry

MALVIN CARL TEICH
Columbia University
Boston University

Google Books
Published: 2024
ISBN: 979-8-9901705-0-6

FOR

LORETTA AND SIDNEY

MARLENE AND ROBERT

CONTENTS

PREFACE xi

1 RAYS 1

1.1 Ray Optics 2
1.2 Mirrors 5
1.3 Planar Boundaries 8
1.4 Spherical Boundaries 12
1.5 Lenses 14
1.6 Optical Fibers 18
Bibliography 22

2 WAVES 24

2.1 Scalar Waves 26
2.2 Monochromatic Scalar Waves 27
2.3 Elementary Scalar Waves 30
2.4 Frequency and Wavelength 34
2.5 Optical Components 35
2.6 Electromagnetic Waves 44
2.7 Random Waves 51
Bibliography 59

3 PHOTONS 61

3.1 The Photon 63
3.2 Photon Energy, Frequency, and Wavelength 65
3.3 Photon Position and Time 66
3.4 Photon Streams 69
3.5 Randomness of Photon Flow 73
3.6 Photon-Number Statistics 75
3.7 Random Partitioning of Photon Streams 79
Bibliography 81

4 THERMAL LIGHT 84

4.1 Temperature and Equipartition of Energy 85
4.2 Occupation of Energy Levels 89
4.3 Interactions of Photons with Atoms 96
4.4 Spontaneous Emission 99
4.5 Absorption and Stimulated Emission 103
4.6 Line Broadening 108
4.7 Blackbody Radiation 111
4.8 Thermal Radiation 116
Bibliography 121

5 SEMICONDUCTOR PHYSICS 125

5.1 Energy Bands 127
5.2 Charge Carriers 131
5.3 Semiconductor Materials 134
5.4 Carrier Concentrations 141
5.5 Generation, Recombination, and Injection 148
5.6 Junctions and Heterojunctions 152
5.7 Quantum Wells and Multiquantum Wells 156
5.8 Quantum Dots 161
5.9 Organic and Perovskite Semiconductors 164
Bibliography 166

6 SEMICONDUCTOR PHOTONICS 169

6.1 Carrier Transitions in Bulk Semiconductors 170
6.2 Interband Transitions 171
6.3 Absorption, Emission, and Gain 175
6.4 Injection Electroluminescence 182
6.5 Quantum Wells and Multiquantum Wells 192
6.6 Quantum-Dot Single-Photon Emitters 193
6.7 Refractive Index 194
Bibliography 195

7 LIGHT-EMITTING DIODES 198

7.1 Photon Flux and Quantum Efficiency 200
7.2 Spatial, Spectral, and Temporal Properties 207
7.3 LED Materials and Device Structures 210
7.4 LEDs for Illumination 215
7.5 Quantum-Dot Light-Emitting Diodes (QLEDs) 218
7.6 Organic Light-Emitting Diodes (OLEDs) 220
7.7 Perovskite Light-Emitting Diodes (PeLEDs) 223
7.8 Laser Diodes and Light-Emitting Diodes 226
Bibliography 229

8 COLOR VISION 234

8.1 Visual Pathways 236
8.2 Eye 237
8.3 Retina 239
8.4 Photoreceptors 240
8.5 Trichromatic Vision 245
8.6 Opponent Channels 247
8.7 Non-Trichromatic Vision 250
8.8 Radiometric and Photometric Units 252
8.9 Luminous Efficacy and Efficiency 257
Bibliography 262

9 COLORIMETRY 265

9.1 Color Matching and Mixing 268
9.2 Grassmann's Laws 270
9.3 Complementary and Metameric Colors 271
9.4 Color Appearance 273
9.5 Color Spaces and Color Solids 276
9.6 Chromaticity Diagrams 285
9.7 Color Temperature 291
9.8 Correlated Color Temperature 296
9.9 Color Rendering Index 299
Bibliography 302

10 PHOSPHOR-CONVERSION LEDS 306

10.1 Monochromatic and White LED Light 308
10.2 Photoluminescence 310
10.3 Broadband and Narrowband Phosphors 312
10.4 Blended Phosphors 318
10.5 Discrete Cool-White PCLEDs 322
10.6 Discrete Warm-White PCLEDs 328
10.7 PCLED Filaments 331
10.8 Chip-on-Board PCLEDs 332
Bibliography 335

11 LED LIGHTING 338

11.1 Merits of LED Lighting 340
11.2 Single-Color LEDs 343
11.3 Additive Color-Mixing LEDs 345
11.4 Retrofit LED Lamps 348
11.5 Hybrid LEDs 353
11.6 LED Luminaires 356
11.7 OLED Light Panels 358
11.8 Smart and Connected LED Lighting 361
11.9 LED Performance Metrics 363
Bibliography 365

A FOURIER TRANSFORM 368

A.1 Definition, Properties, and Examples 368
A.2 Temporal and Spectral Widths 371
Bibliography 374

SYMBOLS AND UNITS 375

AUTHOR 387

INDEX 389

PREFACE

LED Lighting: Devices and Colorimetry

This book is designed to provide a thorough understanding of LED lighting, including the mathematics and fundamental physical principles that underlie this technology. The book is entirely self-contained.

LED Lighting is meant to serve as:

- An introductory textbook for students of optics, photonics, electrical engineering, illumination engineering, and applied physics, at the undergraduate or first-year graduate level.
- A monograph for illumination engineers, lighting designers, and patent lawyers.
- A text for continuing professional development programs offered by colleges, professional societies, and industry.
- A text for practicing engineers in the industrial workplace, and for self-study.

The reader is assumed to have a background in optics, engineering, or applied physics, including elementary course work on waves, modern physics, and quantum mechanics.

LED Lighting is built on three venerable technical fields: optics, photonics, and vision science. *Optics* is the discipline that describes the propagation, diffraction, interference, and imaging of light, as well as its statistical and particulate properties. *Photonics*, an appellation that first came into use in the early 1990s, is an umbrella term for topics that rely on the interaction of light and matter, including semiconductor devices such as light-emitting diodes (LEDs). *Vision science* comprises the scientific study of optometry, photometry, colorimetry, color vision, visual neuroscience, and visual perception.

LED Lighting is the new arrival on the block. The incandescent filament lamp has been the workhorse of artificial lighting since its invention in the late 1800s. Despite its energy inefficiency, incandescent lighting maintained its primacy for more than 100 years because of its low cost, familiarity, and superior color rendering quality. Ultimately, however, an avalanche of advances in LED technology, along with persistent reports highlighting the merits of LED lighting, made their way into the public consciousness and incandescent lighting finally ceded its preeminent position in the early 2000s. Lighting accounts for $\approx 20\%$ percent of electricity use globally. In 2023, the public sale of incandescent lamps was permanently enjoined in the U.S. and the Conference of the Parties to the Minamata Convention agreed to phase out all general-purpose, mercury-containing fluorescent lighting by 2027, worldwide.

Semiconductor LEDs are now universally used in automotive lighting, aerospace and military lighting, entertainment lighting, human- and plant-centric lighting, and especially in residential, architectural, and street lighting. Compact and versatile LED sources with high luminous flux and efficacy can be expressly designed to provide light of any color, including white, with excellent color rendering quality. These sources offer numerous desirable features, including long operational life, slow failure, low cost, low energy consumption, broad choice of colors, dynamic operation, and smart-networking capabilities. In addition to multiquantum-well LEDs (MQWLEDs and μLEDs), we also consider newer types of sources that offer promise for lighting applications, such as quantum-dot devices (QLEDs & WQLEDs), organic light-emitting devices (OLEDs, SMOLEDs, PLEDs, & WOLEDs), and perovskite devices (PeLEDs & PeWLEDs).

Contents

The book consists of four parts, as indicated in the Table below: *Fundamentals*, *Devices*, *Colorimetry*, and *Lighting*. Part I (Chapters 1–5) is devoted to optics, the characterization of thermal and incandescent light, and the essentials of semiconductor physics. Part II (Chapters 6–7) deals with semiconductor photonics and semiconductor devices, particularly light-emitting diodes. Part III (Chapters 8–9) is directed toward human vision, the perception of color, and colorimetry. Part IV (Chapters 10–11) is dedicated to lighting.

Part I	Part II	Part III	Part IV
Fundamentals	Devices	Colorimetry	Lighting
Chapters 1–5	Chapters 6–7	Chapters 8–9	Chapters 10–11

More specifically, Chapters 1, 2, and 3 describe the behavior of light in terms of rays, waves, and photons, respectively. Each of these approaches is best suited to a particular set of applications, as is illustrated by tracing the transmission of light through common optical components such as mirrors, lenses, and optical fibers. Chapter 4 is dedicated to deriving and explaining the properties of blackbody radiation and detailing the characteristics of thermal and incandescent light. Chapter 5 presents the fundamentals of semiconductor physics and materials, outlines the operation of semiconductor junctions and heterojunctions, and introduces quantum and multiquantum wells, quantum dots, organic semiconductors, and perovskite semiconductors.

Chapter 6 considers the interaction of photons with semiconductors, and explains how light is generated via injection electroluminescence. Chapter 7 is devoted to describing semiconductor materials and device structures, along with the operation and properties of LEDs, QLEDs, OLEDs, and PeLEDs.

Chapter 8 discusses the essentials of visual photoreceptor operation, the transmission of information through the pathways of the visual system, and the role of trichromacy in color vision. It also delineates radiometric and photometric quantities as well as efficacy and efficiency measures. Chapter 9 provides an overview of colorimetry that encompasses color matching, color mixing, color appearance, and color spaces. This chapter also elucidates the significance of commonly used LED lighting metrics such as the chromaticity diagram, color temperature (CT), correlated color temperature (CCT), and color rendering index (CRI).

Chapter 10 focuses on photoluminescence, phosphors, and the use of discrete phosphor-conversion light-emitting diodes (PCLEDs) for generating cool- and warm-white light. It also discusses the features of LED filaments and chip-on-board (COB) LEDs. Chapter 11 chronicles the history of LED lighting and reviews its merits. It details the characteristics and properties of color-mixing LEDs, hybrid devices, retrofit LED lamps, LED luminaires, and OLED light panels. Finally, it introduces smart lighting and connected lighting, and concludes with a comparison of the performance metrics for traditional and LED sources of light.

Downloading LED Lighting

LED Lighting, published by Google Books, is a PDF e-book. A compressed version of the PDF file can be downloaded gratis at `people.bu.edu/teich/pdfs/LED-Lighting.pdf` A high-resolution version of the PDF file can be downloaded from Google Books at `books.google.com/books?id=sdX3EAAAQBAJ&newbks=0&hl=en&source=newbks_fb` for a nominal fee.

Presentation

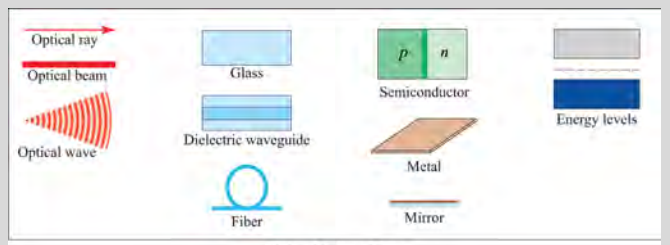
Interactive Features. The following interactive features are incorporated in the book:

- Hyperlinked table of contents at the beginning of the text.
- Hyperlinked table of contents at the beginning of each chapter.
- Hyperlinked table of contents as an optional sidebar.
- Hyperlinked section titles, equations, and figures.
- Hyperlinked index.

Equations, Examples, Bibliographies, and Appendix. Important equations are highlighted by boxes and labels to facilitate retrieval. Examples and derivations are included throughout the text to emphasize concepts and expand the material. Each chapter contains an extensive bibliography that includes a selection of relevant books, review articles, papers of special significance, and classic papers. The appendix summarizes the definition and properties of the Fourier transform.

Symbols, Notation, Units, and Conventions. We use the symbols, notation, units, and conventions commonly employed in the optics and photonics literature. To minimize confusion, symbols that have multiple meanings are delineated by the use of different fonts, where possible. A list of symbols, units, abbreviations, and acronyms follows the Appendix. We adhere to the International System of Units (SI units). This modern form of the metric system is based on the meter, kilogram, second, ampere, kelvin, candela, and mole, and is coupled with a collection of prefixes that indicate multiplication or division by various powers of ten. Still, the reader is cautioned that photonics in the service of different areas of science and engineering can make use of different conventions and symbols. In Chapter 2, for example, the complex wavefunction for a monochromatic plane wave is written in the form commonly used in engineering, which differs from that used in physics, as highlighted by an *in situ* footnote.

Color Coding of Illustrations. The color code used in illustrations is summarized in the chart presented below. Light rays, beams, and optical-field distributions are displayed in red. When optical fields are represented, white indicates negative values but when intensity is portrayed, white indicates zero. Glass, dielectric waveguides, and glass fibers are depicted in light blue; darker shades represent higher refractive indices. Semiconductors are cast in green, with various shades representing different levels of doping. Metal and mirrors are indicated in copper. Semiconductor energy-band diagrams are portrayed in blue and gray.



Back Cover

The royal blue and yellow of the back cover represent, respectively, the colors of the LED light and the phosphor used to generate metameric cool-white light for LED lighting. White LED retrofit lamps containing phosphor-conversion LEDs that operate in accordance with this principle have upended incandescent lighting.

Photo Credits for Chapter-Opening Pages

The photos displayed on the chapter-opening pages are in the public domain (additional information is provided where available): Fermat; Newton (1689, Portrait by Godfrey Kneller); Huygens (1671, Portrait by Caspar Netscher); Maxwell (Photograph by Fergus of Greenock); Planck (ca. 1878); Einstein (ca. 1904); Kelvin (Courtesy of the Kelvin Museum of the University of Glasgow); Boltzmann; Shockley, Bardeen, & Brattain (ca. 1964); Round (ca. 1920); Losev (ca. 1920); Keyes & Quist (1962, Courtesy of Robert J. Keyes and MIT Lincoln Laboratory); Young; von Helmholtz; Grassmann; Wright (Courtesy of the Colour Group of the UK); Guild (Courtesy of the Colour Group of the UK); Akasaki, Amano, & Nakamura (2014, Courtesy of the Embassy of Japan in Sweden); Holonyak (Courtesy of Nick Holonyak, Jr.); Craford (1996, Courtesy of the Grainger College of Engineering of the University of Illinois Urbana-Champaign); Teich (Courtesy of Boston University).

Acknowledgments

I am indebted to the legions of students and postdoctoral associates who, through a combination of vigilance and the desire to understand the material, engaged me in the classroom and in the laboratory, and taught me so much. I am also beholden to my colleagues in industry and in the academy, especially Bahaa Saleh, and to the many patent attorneys with whom I have had the pleasure of working over the years, who helped enlarge my perspective in untold ways.

I am most grateful for the financial support I received throughout my career from a multitude of U.S. Government agencies, including the National Science Foundation (NSF) and its Engineering Research Centers, particularly the Center for Telecommunications Research (CTR) at Columbia University and the Center for Subsurface Sensing and Imaging Systems (CenSSIS) at Boston University and Northeastern University; the Joint Services Electronics Program (JSEP) and the Columbia Radiation Laboratory (CRL) in the Department of Physics at Columbia University; the Defense Advanced Research Projects Agency (DARPA); the U.S. Office of Naval Research (ONR); and the U.S. Army Research Office (ARO). I also acknowledge with gratitude an Interdisciplinary Science Grant from the David and Lucile Packard Foundation.

Finally, I express my deep appreciation to the Columbia University School of Engineering and Applied Science (SEAS), to the Boston University College of Engineering (COE), and to the Boston University Photonics Center (BUPC), for generous and unwavering support over the course of my career.

MALVIN CARL TEICH

Boston, Massachusetts
February 20, 2024

Second Iteration

Minor errors in the first iteration have been corrected and a number of stylistic improvements have been made.

Boston, Massachusetts
April 17, 2024

CHAPTER

1

RAYS

1.1 RAY OPTICS 2

1.2 MIRRORS 5

1.3 PLANAR BOUNDARIES 8

1.4 SPHERICAL BOUNDARIES 12

1.5 LENSES 14

1.6 OPTICAL FIBERS 18

Pierre de Fermat (1601–1665) proposed a rule, now known as Fermat's Principle, whereby light rays travel along the path of minimum time relative to neighboring paths.

Sir Isaac Newton (1642–1727) set forth a corpuscular theory of optics in which light emissions comprise collections of corpuscles that propagate rectilinearly.

LED Lighting: Devices and Colorimetry. Malvin Carl Teich. Google Books. Published 2024.

Ray optics, also called **geometrical optics**, is the simplest theory of light. Although an approximate theory, it is nevertheless adequate for explaining most of our daily observations relating to the behavior of light. Light is treated as **rays** that travel in optical media in accordance with a set of geometrical rules that govern their locations and directions. Ray optics is suitable for describing the collection, guiding, and control of light, including that emitted by light-emitting diodes (LEDs). It is also useful for describing image formation, a process by means of which a collection of rays from a given point of an object is redirected by optical components to a corresponding point of an image.

The principles of ray optics are set forth in Sec. 1.1. They are then used in Secs. 1.2–1.6, without invoking any other assumptions regarding the nature of light, to determine the rules that govern the propagation of light rays at mirrors, planar boundaries, spherical boundaries, lenses, and fibers, respectively.

However, ray optics is silent on a number of features of light required for understanding its behavior in more subtle experiments, such as its wavelength, spectrum, color, diffraction, and interference. Accommodating those features requires the use of a more advanced theory, in which light is treated as waves (Chapter 2) or as photons (Chapter 3).

1.1 RAY OPTICS

Principles of Ray Optics

- Light travels in the form of rays. The rays are emitted by light sources and can be observed when they reach an observer or an optical detector.
- An optical medium is characterized by a quantity n called the **refractive index** ($n \geqslant 1$). This quantity is determined by $n = c_0/c$ where c_0 is the **speed of light** in free space and c is the speed of light in the medium. The time it takes light to travel a distance d is therefore $d/c = nd/c_0$; it is proportional to the product nd, which is called the **optical pathlength**.
- In an inhomogeneous medium, the refractive index $n(\mathbf{r})$ depends on the position $\mathbf{r} = (x, y, z)$. The optical pathlength along a given path between the two points A and B is then written as

$$\text{Optical pathlength} = \int_A^B n(\mathbf{r})\, ds, \tag{1.1-1}$$

  where ds is the differential element of length along the path.
- **Fermat's Principle** states that optical rays traveling between the points A and B follow a path such that the time of travel, which is proportional to the optical pathlength, is an extremum relative to neighboring paths. This is expressed mathematically as

$$\delta \int_A^B n(\mathbf{r})\, ds = 0\,, \tag{1.1-2}$$

Fermat's Principle

  where the symbol δ (which is read "the variation of") signifies that the optical pathlength is either minimized or maximized, or is a point of inflection. In most cases the optical pathlength minimized, in which case we have:

Light travels in the form of rays along the path of minimum time.

Propagation in a Homogeneous Medium

The refractive index is the same everywhere in a homogeneous medium, and therefore so too is the speed of light. The path of minimum time, as required by Fermat's principle, is then also the *path of minimum distance*, which is known as **Hero's principle** (Fig. 1.1-1).

In a homogeneous medium, light rays travels along paths of minimum distance, which are straight lines.

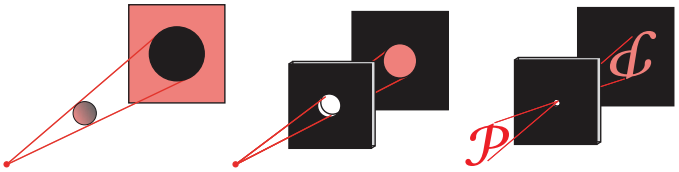


Figure 1.1-1 Hero's principle: light rays travel in straight lines in a homogeneous medium. Shadows are therefore perfect projections of stops.

Reflection from a Mirror

Mirrors are fabricated using dielectric or metallic films deposited on a substrate such as glass, or highly polished metallic surfaces. Light rays reflect from mirrors in accordance with the **law of reflection**:

The reflected ray lies in the plane of incidence and the angle of reflection equals the angle of incidence.

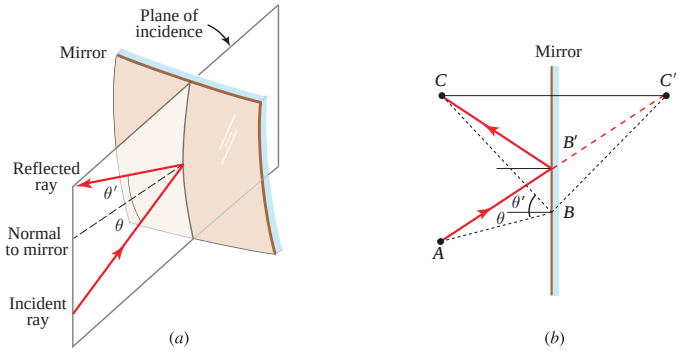


Figure 1.1-2 (a) Reflection at the surface of a curved mirror. (b) Geometrical construction to prove the law of reflection.

As depicted in Fig. 1.1-2(a), the plane of incidence is the plane formed by the incident ray and the normal to the mirror at the point of incidence. The angle of incidence θ and the angle of reflection θ' are specified. As shown in Fig. 1.1-2(b), the law of reflection is a simple consequence of Hero's principle: A ray traveling from point A to point C, via reflection at point B from a planar mirror of infinitesimal thickness, travels along a path of minimum distance, i.e., $\overline{AB} + \overline{BC}$ must be minimum. Now, if C' is a mirror image of C, then $\overline{BC} = \overline{BC'}$, so $\overline{AB} + \overline{BC'}$ must be minimum. This occurs when $\overline{ABC'}$ is a straight line, namely when B coincides with B', so that the angle of reflection equals the angle of incidence, i.e.,

$$\boxed{\theta' = \theta\,.} \qquad (1.1\text{-}3)$$

Law of Reflection

Reflection and Refraction at the Boundary Between Two Media

At the boundary between two media of refractive indices n_1 and n_2, an incident ray is split into a reflected ray and a refracted (or transmitted) ray, as shown in Fig. 1.1-3. The reflected ray obeys the law of reflection (1.1-3). The refracted ray obeys the law of refraction, known as **Snell's law**:

***The refracted ray lies in the plane of incidence and the angle of refraction θ_2 is related to the angle of incidence θ_1 by Snell's law*:**

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 . \tag{1.1-4}$$

Snell's Law

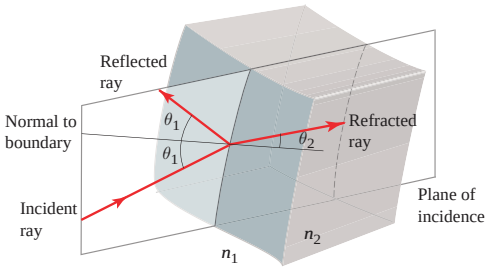


Figure 1.1-3 Reflection and refraction at the boundary between two media. The angle of refraction and the angle of incidence are related by Snell's law: $n_1 \sin\theta_1 = n_2 \sin\theta_2$.

□ **Proof of Snell's Law.** The proof of Snell's law is an exercise in the application of Fermat's principle. Referring to the construction in Fig. 1.1-4, the object is to minimize the optical pathlength $n_1\overline{AB} + n_2\overline{BC}$ between points A and C.

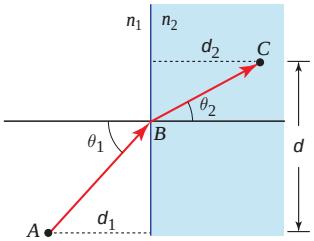


Figure 1.1-4 The optimization problem is to minimize the pathlength $n_1 d_1 \sec\theta_1 + n_2 d_2 \sec\theta_2$ with respect to the angles θ_1 and θ_2, subject to the condition $d_1 \tan\theta_1 + d_2 \tan\theta_2 = d$.

The pathlength is minimized when $\frac{\partial}{\partial\theta_1}[n_1 d_1 \sec\theta_1 + n_2 d_2 \sec\theta_2] = 0$, so that $n_1 d_1 \sec\theta_1 \tan\theta_1 + n_2 d_2 \sec\theta_2 \tan\theta_2 \frac{\partial\theta_2}{\partial\theta_1} = 0$. Also, $\frac{\partial}{\partial\theta_1}[d_1 \tan\theta_1 + d_2 \tan\theta_2] = 0$, which yields $d_1 \sec^2\theta_1 + d_2 \sec^2\theta_2 \frac{\partial\theta_2}{\partial\theta_1} = 0$, whereupon $\frac{\partial\theta_2}{\partial\theta_1} = -\frac{d_1 \sec^2\theta_1}{d_2 \sec^2\theta_2}$. Hence, $n_1 d_1 \sec\theta_1 \tan\theta_1 - n_2 \frac{d_1 \sec^2\theta_1 \tan\theta_2}{\sec\theta_2} = 0$, which provides $n_1 \tan\theta_1 = n_2 \sec\theta_1 \sin\theta_2$, from which we have $n_1 \sin\theta_1 = n_2 \sin\theta_2$, which is Snell's law. ■

Optical Energy

In isotropic media, i.e., media that behave in the same way in all directions, optical rays point in the direction of the flow of optical energy. Ray bundles can be constructed in which the density of rays is proportional to the density of light energy. When light is generated isotropically from a point source, for example, the energy associated with the rays in a given cone is proportional to the solid angle Ω of that cone. Rays may be traced through an optical system to determine the optical energy crossing a given area. However, ray optics cannot determine the proportion of optical energy reflected and refracted at the interface between media, nor can it assess optical loss.

Summary

In the remainder of this chapter, we apply the three simple rules set forth in Section 1.1 to a number of geometrical configurations consisting of mirrors and transparent optical components, without any need for further recourse to Fermat's principle:

1. *Rays travel in straight lines in a homogeneous medium.*
2. *The law of reflection* (1.1-3) *is satisfied*: $\theta' = \theta$.
3. *The law of refraction* (1.1-4) *is satisfied*: $n_1 \sin\theta_1 = n_2 \sin\theta_2$.

1.2 MIRRORS

Planar Mirrors

As illustrated in Fig. 1.2-1, a planar mirror reflects the rays originating from a point P_1 such that the reflected rays appear to originate from a point P_2 behind the mirror, called the image.

Paraboloidal Mirrors

The surface of a paraboloidal mirror is a reflective paraboloid of revolution, as depicted in Fig. 1.2-2. It has the useful property of focusing all rays arriving parallel to its axis to a single point, called the **focus** or **focal point**. The distance $\overline{PF} \equiv f$ is known as the **focal length**. Paraboloidal mirrors are often used as light-collecting elements in optical systems such as telescopes. They are also used in reverse to render parallel the rays from a point source of light, such as a light-emitting diode, placed at the focus. When used in this manner, the paraboloidal mirror serves as a **collimator**.

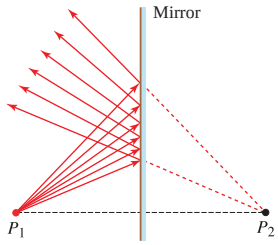


Figure 1.2-1 Reflection of light by a planar mirror.

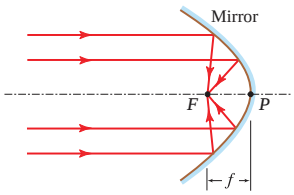


Figure 1.2-2 Focusing of light by a paraboloidal mirror.

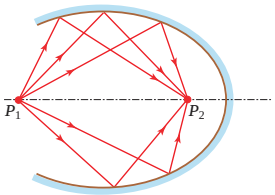


Figure 1.2-3 Reflection of light from an elliptical mirror.

Elliptical Mirrors

An elliptical mirror reflects all the rays emitted from one of its two foci and images them onto the other focus, e.g., from P_1 to P_2 as shown in Fig. 1.2-3. In accordance with Hero's principle, the distances traveled by the rays from P_1 to P_2 along any of the paths are equal.

Spherical Mirrors

In general, the spherical mirror has neither the focusing property of the paraboloidal mirror nor the imaging property of the elliptical mirror. As illustrated in Fig. 1.2-4, parallel incident rays reflected from the mirror meet the z axis at different locations.

The curve that is tangent to these rays, shown as dashed, defines an envelope called the **caustic**. By convention, the **radius of curvature** R is taken to be negative for concave mirrors and positive for convex mirrors.

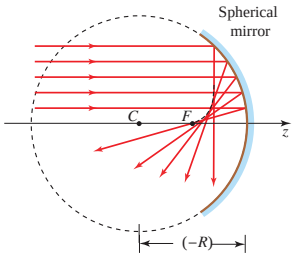


Figure 1.2-4 Reflection of parallel rays from a spherical mirror. The radius of curvature $(-R)$ is negative for concave mirrors.

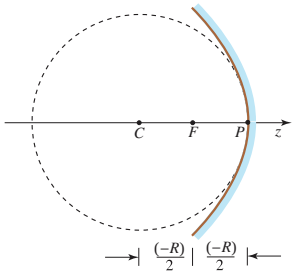


Figure 1.2-5 For paraxial rays, a spherical mirror of radius $(-R)$ approximates a paraboloidal mirror of focal length $(-R)/2$.

Paraxial Rays Reflected from a Spherical Mirror. Rays that make small angles with the axis of an optical component, such that $\sin\theta \approx \theta$ and $\tan\theta \approx \theta$, are called **paraxial rays**. The body of rules that results from considering only paraxial rays forms the field of **paraxial optics**, also called **first-order optics** or **Gaussian optics**. In the **paraxial approximation**, the spherical mirror does turn out to have the focusing property of a paraboloidal mirror *and* the imaging property of an elliptical mirror, as is demonstrated below.

As is understood from Fig. 1.2-5, at points near the axis, a parabola can be approximated by a circle whose radius matches the parabola's radius of curvature; hence, a spherical mirror of radius $(-R)$ acts as a paraboloidal mirror of focal length $f = (-R)/2$. The construction that explains the reflection of paraxial rays from a concave spherical mirror is provided in Fig. 1.2-6. All paraxial rays originating from each point on the axis of the mirror are reflected and focused onto a single corresponding point on the axis. This is readily verified by examining a ray that travels at an angle θ_1 (with respect to the z axis) from point P_1 at a distance z_1 to the left of a concave mirror of radius $(-R)$, and reflects to arrive at an angle $(-\theta_2)$ to meet the z axis at point P_2 at a distance z_2 to the left of the mirror. Considering the two triangles that include the vertex C, and using the fact that the three angles of a triangle sum to 180°, we obtain $\theta_1 = \theta_0 - \theta$ and $(-\theta_2) = \theta_0 + \theta$, which, when added, yield $(-\theta_2) + \theta_1 = 2\theta_0$.

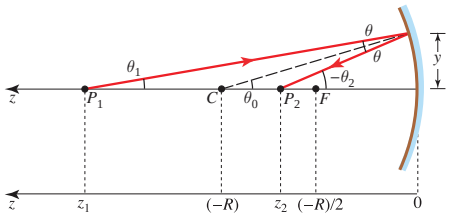


Figure 1.2-6 Construction displaying the reflection of paraxial rays from a concave spherical mirror ($R < 0$). A negative angle indicates a ray traveling downward with respect to the z axis.

Furthermore, if θ_0 is sufficiently small, such that $\tan\theta_0 \approx \theta_0$, we have $\theta_0 \approx y/(-R)$, whereupon

$$(-\theta_2) + \theta_1 \approx \frac{2y}{(-R)}\,, \tag{1.2-1}$$

where y is the height of the point above the z axis at which the reflection occurs. Similarly, if θ_1 and θ_2 are small, we have $\theta_1 \approx y/z_1$ and $(-\theta_2) = y/z_2$, so that (1.2-1) yields $y/z_1 + y/z_2 \approx 2y/(-R)$, or

$$\frac{1}{z_1} + \frac{1}{z_2} \approx \frac{2}{(-R)} . \tag{1.2-2}$$

Equation (1.2-2) remains valid whatever the value of y (i.e., for all values of θ_1) provided that the paraxial assumption holds. Hence, all paraxial rays originating from point P_1 arrive at P_2. It is understood from Fig. 1.2-6 that the distances z_1 and z_2 are measured in a coordinate system in which the z axis points to the left, and points of negative z lie to the right of the mirror.

Focusing by a Spherical Mirror. In accordance with (1.2-2), rays originating from a far point on the z axis ($z_1 = \infty$) are focused to a point F at the distance $z_2 = (-R)/2$. Hence, within the paraxial approximation, rays arriving parallel to the axis of the mirror (all of which come from infinity) are focused to a point at a distance f from the mirror known as its **focal length**:

$$\boxed{f = \frac{(-R)}{2} .} \tag{1.2-3}$$

Focal Length
Spherical Mirror

Combining (1.2-3) and (1.2-2) allows the latter to be written in the form

$$\boxed{\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f} ,} \tag{1.2-4}$$

Imaging Equation
(Paraxial Rays)

which is known as the **imaging equation**. All rays, both incident and reflected, must be paraxial for this equation to hold.

Image Formation by a Spherical Mirror. The application of the imaging equation (1.2-4) for rays that originate at a distance from the z axis is schematized in Figure 1.2-7. Within the paraxial approximation, rays originating from point $P_1 = (y_1, z_1)$ are reflected to point $P_2 = (y_2, z_2)$, where z_1 and z_2 satisfy (1.2-4) and also $y_2 = -y_1 z_2/z_1$. This indicates that rays from each point in the plane $z = z_1$ meet at a single corresponding point in the plane $z = z_2$, confirming that the mirror acts as an image-formation system with **magnification** $M = -z_2/z_1$. The derivations are furnished below using the construction provided in Fig. 1.2-8. We conclude that spherical mirrors are useful for both focusing and imaging, provided that all rays are paraxial.

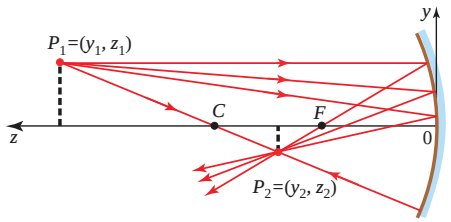


Figure 1.2-7 Image formation by a spherical mirror. The paths of four particular rays are illustrated. Negative magnification signifies an inverted image.

□ **Derivation of Relations for Image Formation by a Spherical Mirror.** A ray originating at $P_1 = (y_1, z_1)$ and traveling at angle θ_1 meets the mirror at height $y \approx y_1 + \theta_1 z_1$. Again, the three angles of a triangle sum to 180° so the angle of incidence at the mirror is given by $\phi = \psi - \theta_1 \approx y/(-R) - \theta_1$.

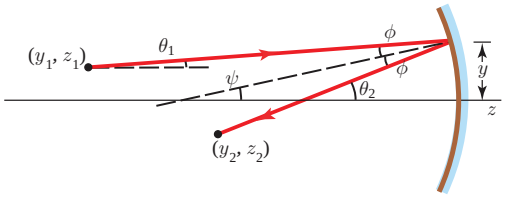


Figure 1.2-8 Construction for image formation at a spherical mirror, corresponding to Fig. 1.2-7.

The reflected ray makes angle θ_2 with the z axis where $\theta_2 = 2\phi + \theta_1 = 2[y/(-R) - \theta_1] + \theta_1 = 2y/(-R) - \theta_1 = 2(y_1 + \theta_1 z_1)/(-R) - \theta_1$. Substituting $f = (-R)/2$ leads to $\theta_2 = (y_1 + \theta_1 z_1)/f - \theta_1$. The height y_2 is determined from $[y + (-y_2)]/z_2 \approx \theta_2$. Combining these results yields $y_1 + \theta_1 z_1 - y_2 = z_2[(y_1 + \theta_1 z_1)/f - \theta_1]$ and $y_2 = y_1 - z_2 y_1/f + \theta_1(z_1 - z_1 z_2/f + z_2)$. If $z_1 - z_1 z_2/f + z_2 = 0$, or $1/z_1 + 1/z_2 = 1/f$, we arrive at $y_2 = y_1(1 - z_2/f)$, which is independent of θ_1. Hence, $z_2/f = 1 - y_2/y_1$ so that $y_2 = -y_1 z_2/z_1$. ■

Ray Tracing

The process of following rays as they undergo reflection and refraction at each surface of an optical system, as carried out above, is known as **ray tracing**. Many of the results derived in this chapter are formulated only for paraxial rays, and are therefore approximate. In practice, ray tracing is often implemented via software, which has the merit that it is not constrained by the paraxial approximation.

1.3 PLANAR BOUNDARIES

External and Internal Refraction

The relation between the angles of refraction and incidence, θ_2 and θ_1 respectively, at a planar boundary between two media of refractive indices n_1 and n_2 is governed by Snell's law (1.1-4). This relation is illustrated in Fig. 1.3-1 for two cases:

- *External Refraction* $(n_1 < n_2)$. When the ray is incident from the medium of lower refractive index, the refracted ray bends toward the normal and $\theta_2 < \theta_1$.
- *Internal Refraction* $(n_1 > n_2)$. When the ray is incident from the medium of higher refractive index, the refracted ray bends away from the normal and $\theta_2 > \theta_1$.

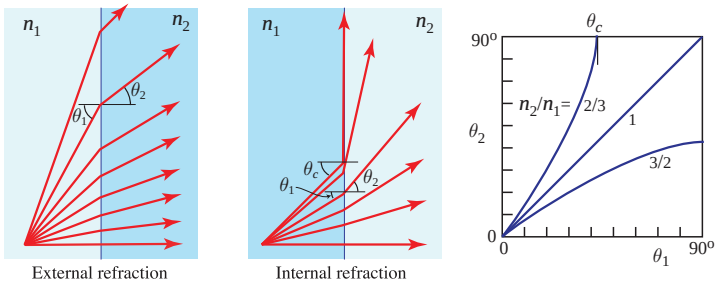


Figure 1.3-1 Relation between the angle of refraction θ_2 and the angle of incidence θ_1 at a planar boundary between two media of different refractive indices. Results are shown for both external refraction $(n_1 < n_2)$ and internal refraction $(n_1 > n_2)$. The medium of higher index is more deeply shaded. For paraxial rays, θ_2 is linearly proportional to θ_1.

The refracted rays bend in such a way as to minimize the optical pathlength, i.e., to increase pathlength in the lower-index medium at the expense of pathlength in the higher-index medium.

Paraxial Snell's Law

For both external and internal refraction, when the angles θ_1 and θ_2 are small, the rays are paraxial and we can use the approximation $\sin\theta \approx \theta$ to write Snell's law (1.1-4) in the form of its paraxial approximation:

$$n_1\theta_1 \approx n_2\theta_2\,. \tag{1.3-1}$$

Snell's Law (Paraxial Rays)

The relation between θ_2 and θ_1 is then approximately linear, $\theta_2 \approx (n_1/n_2)\,\theta_1$, as is observed in the right-hand panel of Fig. 1.3-1.

Total Internal Reflection

For internal refraction $(n_1 > n_2)$, where the angle of refraction is greater than the angle of incidence $(\theta_2 > \theta_1)$, as θ_1 increases, θ_2 ultimately reaches 90°. At that point, θ_1 is said to reach the **critical angle** θ_c, as depicted in Fig. 1.3-1. This occurs when $n_1 \sin\theta_c = n_2 \sin(\pi/2) = n_2$, so the critical angle is determined by

$$\theta_c = \sin^{-1}\frac{n_2}{n_1}\,. \tag{1.3-2}$$

Critical Angle

For $\theta_1 > \theta_c$, Snell's law (1.1-4) cannot be satisfied and refraction does not occur. Rather, the incident ray is totally reflected, as if the surface were a perfect mirror, as illustrated in Fig. 1.3-2(a). This phenomenon, called **total internal reflection (TIR)**, is the basis of operation of many optical devices and systems, such as reflecting prisms and optical fibers, as schematically shown in Figs. 1.3-2(b) and (c), respectively. Other examples of TIR will come to the fore subsequently. Electromagnetic optics reveals that all of the energy in TIR is carried by the reflected light so the process is highly efficient.

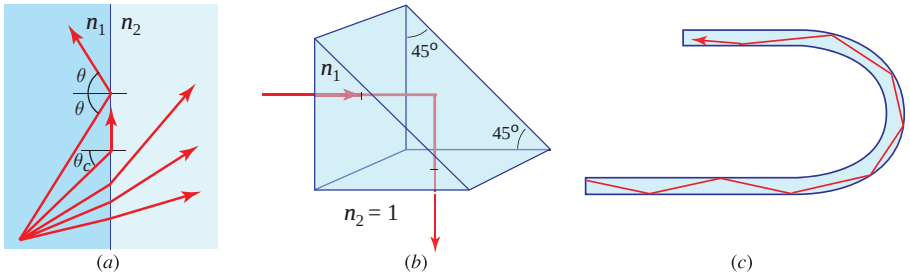


Figure 1.3-2 (a) Total internal reflection (TIR) takes place at a planar boundary between two media when $\theta_1 > \theta_c$. (b) Internal reflection in a reflecting prism: If $n_1 > \sqrt{2}$ and $n_2 = 1$ (air), then $\theta_c < 45^\circ$. For glass, $n_2 \approx 1.5 > \sqrt{2}$ so the ray is totally reflected when $\theta_1 = 45^\circ$. (c) In an optical fiber, rays are guided by total internal reflection from the internal surface of the fiber.

EXAMPLE 1.3-1. ***Light Trapped by Total Internal Reflection in an LED.*** Rays originating within a medium of high refractive index, such as an LED, can remain trapped within the medium, especially if its surfaces are parallel. This occurs because a certain proportion of the rays undergo multiple total internal reflections and never refract into air.

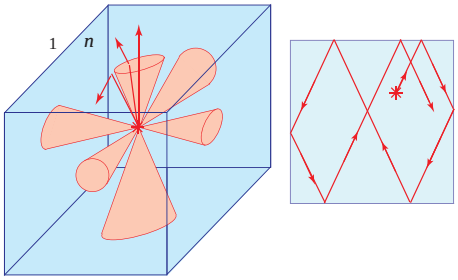
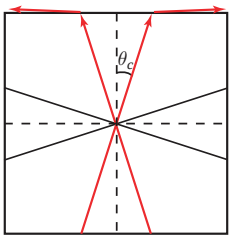


Figure 1.3-3 Trapping of light in a parallelepiped of high refractive index.

Consider a material of refractive index n cut in the shape of a parallelepiped (Fig. 1.3-3) surrounded by air ($n = 1$). Assuming that the light is generated uniformly and isotropically within the material, all of the rays within the cones of internal half angle $\theta_c = \sin^{-1}(1/n)$, as illustrated at the right, are refracted into air. All rays outside of the six cones corresponding to the surfaces of the parallelepipid are internally reflected. In materials for which $\theta_c < 45°$, the cones do not overlap, and those rays reflect endlessly outside the cones and remain trapped within the material. For GaAs, $n = 3.6$ so that the critical angle $\theta_c = 16.1° < 45°$.



If we assume that the optical power associated with the rays in a given cone is proportional to its solid angle, the fraction of the optical power that can be extracted is determined by the critical angle at the surfaces of the structure. For uniform and isotropic emission within a parallelepiped of refractive index n surrounded by air ($n = 1$), and for $\theta_c < 45°$ or $n > \sqrt{2}$ so that the cones do not overlap, this result is established by calculating the area of the spherical cap atop each of these cones, which is $A = \int_0^{\theta_c} 2\pi r \sin\theta \, r \, d\theta = 2\pi r^2(1 - \cos\theta_c)$. Since the area of the entire sphere is $4\pi r^2$, the fraction of the emitted light lying within the solid angle subtended by a single cone is $\Omega = A/4\pi r^2 = \frac{1}{2}(1 - \cos\theta_c)$. Hence, for the six faces of the parallelepiped, the ratio of the extracted light to the total light is $3(1 - \cos\theta_c) = 3(1 - \sqrt{1 - 1/n^2})$. For GaAs, this fraction in LEDs is 11.8% Techniques used to mitigate the difficulty of extracting light from materials of high refractive index are discussed in Sec. 7.1.

Prisms

A **prism** of apex angle α and refractive index n, such as that portrayed in Fig. 1.3-4, bends a ray arriving at an angle of incidence θ by the **deflection angle**

$$\theta_d = \theta - \alpha + \sin^{-1}\left[\sqrt{n^2 - \sin^2\theta}\ \sin\alpha - \sin\theta\cos\alpha\right]. \tag{1.3-3}$$

This equation, which is valid for arbitrary values of α, θ, and n, is arrived at by using Snell's law twice, at the two refracting surfaces of the prism. The behavior of the deflection angle as a function of the angle of incidence is graphically illustrated in Fig. 1.3-4. When α is small (**thin prism**), and θ is also small (paraxial approximation), (1.3-3) may be approximated by

$$\theta_d \approx (n - 1)\alpha. \tag{1.3-4}$$

Deflection Angle
(Thin Prism, Paraxial Rays)

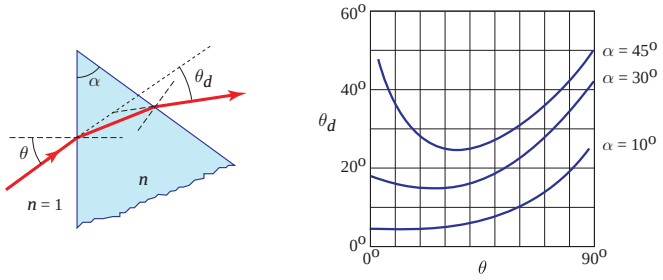
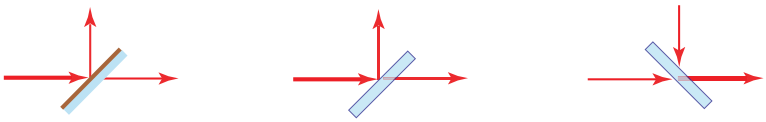


Figure 1.3-4 (a) Ray deflection by a prism. (b) Plot of (1.3-3) for the deflection angle θ_d imparted by a prism ($n = 1.5$), as a function of the angle of incidence θ, with the apex angle α as a parameter. When both α and θ are small, the angle of deflection can be approximated by $\theta_d \approx (n-1)\alpha$, which is approximately independent of θ, as is apparent from the $\alpha = 10°$ curve. For $\theta = 0°$ and $\alpha = 45°$, total internal reflection is operative so a deflection angle does not exist [see Fig. 1.3-2(b)].

Beamsplitters

A beamsplitter, or partially reflective mirror, is an optical component that splits an incident ray into a reflected ray and a transmitted ray, as illustrated in Fig. 1.3-5. Beamsplitters are often constructed by depositing a thin, semitransparent dielectric or metallic film on a glass or plastic substrate. A thin, bare glass plate, such as a microscope slide [Fig. 1.3-5(b)], can also serve as a beamsplitter although the fraction of light reflected is small (the relative proportion of light transmitted and reflected is established by the Fresnel equations of electromagnetic optics). Beamsplitters are also frequently used to combine two light rays into one, as shown schematically in Fig. 1.3-5(c), in which case they are called beam combiners.



(a) Partially reflective mirror (b) Thin glass plate (c) Beam combiner

Figure 1.3-5 Beamsplitters and beam combiners.

Beam Directors

Simple optical components can also be used to direct rays in particular directions. The devices illustrated in Fig. 1.3-6 redirect parallel incident rays into rays tilted at fixed angles with respect to each other. The **biprism** depicted in Fig. 1.3-6(a) is the juxtaposition of a prism and an identical inverted prism. The **Fresnel biprism** portrayed in Fig. 1.3-6(b) is formed from rows of adjacently placed tiny prisms. This device is equivalent to the biprism but is thinner and lighter, although beam quality can be limited by diffraction at the discontinuities. The cone-shaped optic depicted in Fig. 1.3-6(c), known as an **axicon**, converts incident rays into a collection of circularly symmetric rays directed toward its central axis in the form of a cone. It has the same cross section as the biprism, namely an isosceles triangle.

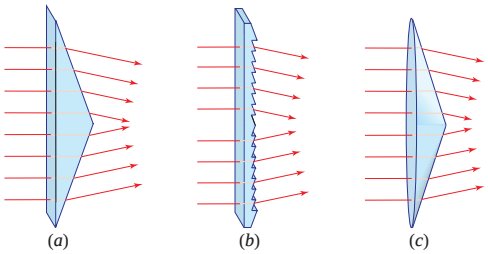


Figure 1.3-6 (a) Biprism. (b) Fresnel biprism. (c) Plano-convex axicon.

1.4 SPHERICAL BOUNDARIES

Having examined the refraction of rays at planar boundaries in Sec. 1.3, we turn now to the refraction of rays at spherical boundaries. In particular, we examine the refraction at a spherical boundary of radius R between two media of refractive indices n_1 and n_2. The results are obtained by applying Snell's law, which relates the angles of refraction and incidence relative to the normal to the surface, which is defined by the radius vector from the center C. These angles are to be distinguished from the angles θ_1 and θ_2 defined relative to the z axis. In analogy with the spherical mirror, convention dictates that for a ray entering from the left, the radius of curvature R is positive for a convex boundary and negative for a concave boundary.

Paraxial Rays Incident on a Spherical Boundary. We consider only paraxial rays that make small angles with respect to the axis of the system so that $\sin\theta \approx \theta$ and $\tan\theta \approx \theta$. Our calculations are therefore accommodated by the paraxial version of Snell's law provided in (1.3-1), i.e., $n_1\theta_1 \approx n_2\theta_2$, which leads to the following relations:

- As depicted in Fig. 1.4-1(a), a ray that makes an angle θ_1 with the z axis meets the boundary at a point y above the z axis and changes direction when it refracts at the boundary to make an angle $(-\theta_2)$ with the z axis. In accordance with the derivation provided below, the angle θ_2 turns out to be

$$\theta_2 \approx \frac{n_1}{n_2}\theta_1 - \frac{n_2 - n_1}{n_2}\frac{y}{R}. \tag{1.4-1}$$

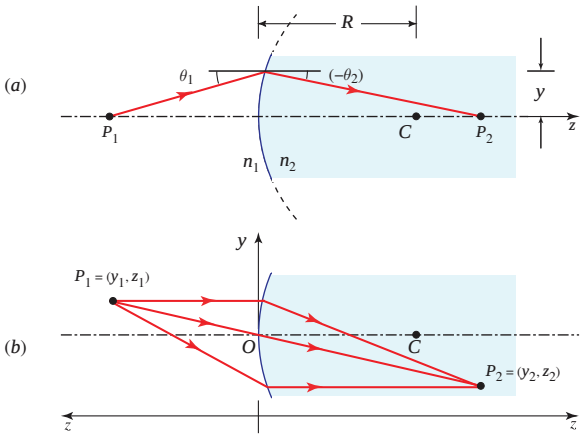


Figure 1.4-1 Refraction at a spherical boundary. By convention, P_1 and P_2 are measured in coordinate systems that point to the left and right, respectively (if P_2 were to lie to the left of the boundary, then z_2 would be negative). Convention also dictates that for a ray entering from the left, the radius of curvature R is positive for a convex boundary and negative for a concave boundary. A negative angle indicates a ray traveling downward with respect to the z axis and negative magnification signifies an inverted image.

- As is evident in Fig. 1.4-1(b), all paraxial rays originating from a point $P_1 = (y_1, z_1)$ in the $z = z_1$ plane meet at a point $P_2 = (y_2, z_2)$ in the $z = z_2$ plane, where

$$\frac{n_1}{z_1} + \frac{n_2}{z_2} \approx \frac{n_2 - n_1}{R} \tag{1.4-2}$$

and

$$y_2 = -\frac{n_1}{n_2}\frac{z_2}{z_1} y_1 \,. \tag{1.4-3}$$

The $z = z_1$ and $z = z_2$ planes are said to be conjugate: Every point in the $z = z_1$ plane has a corresponding point (image) in the $z = z_2$ plane, with **magnification** $-(n_1/n_2)(z_2/z_1)$.

The similarities between these properties and those of the spherical mirror discussed in Sec. 1.2 are evident. The results provided aboveare derived below by making use of the construction provided in Fig. 1.4-2. It is important to keep in mind that rays at large angles do not obey these paraxial laws and the deviations results in image distortion called **aberration**.

□ **Derivation of Relations for Image Formation by a Spherical Boundary.** We first provide a derivation of (1.4-1) and then show that paraxial rays originating from P_1 pass through P_2 when (1.4-2) and (1.4-3) are satisfied.

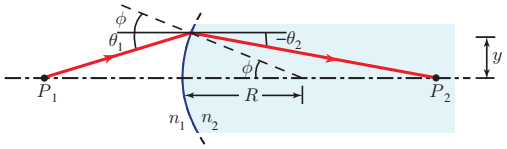


Figure 1.4-2 Construction for refraction at a convex spherical boundary ($R > 0$), corresponding to Fig. 1.4-1.

- Snell's law dictates that $n_1 \sin(\theta_1 + \phi) = n_2 \sin[\phi - (-\theta_2)]$. Hence, the paraxial version of Snell's law provides $n_1(\theta_1 + \phi) \approx n_2(\phi + \theta_2)$, which gives rise to $\theta_2 \approx (n_1/n_2)\theta_1 + [(n_1 - n_2)/n_2]\phi$. Because $\phi \approx y/R$, we obtain $\theta_2 \approx (n_1/n_2)\theta_1 - [(n_2 - n_1)/n_2](y/R)$, thereby reproducing (1.4-1).
- Substituting $\theta_1 \approx y/z_1$ and $(-\theta_2) \approx y/z_2$ into (1.4-1) gives rise to $-y/z_2 \approx (n_1/n_2)y/z_1 - [(n_2 - n_1)/n_2](y/R)$, from which (1.4-2) follows.
- Referring to Fig. 1.4-1(b) and considering the ray passing through the origin O, the angles of incidence and refraction are given by y_1/z_1 and $-y_2/z_2$, respectively, so that the paraxial version of Snell's law leads to (1.4-3). Rays at other angles are also directed from P_1 to P_2, as is readily demonstrated using arguments similar to those employed in connection with Fig. 1.2-8.

■

EXAMPLE 1.4-1. ***Collimator for Light Emitted by an Light-Emitting Diode.*** The light emitted by a light-emitting diode (LED) is often collimated by making use of an optic whose surface takes the form of a paraboloid of revolution, as depicted in Fig. 1.4-3. The LED is placed at the focus of the paraboloid by inserting its hemispherical dome (darker blue) into a recess formed in the narrow end of the optic. Rays emanating from the sides of the LED dome impinge on the paraboloidal boundary at angles of incidence greater than the critical angle θ_c and are thus reflected out of the device via total internal reflection. Rays emanating from the central portion of the LED dome are refracted out of the device at the spherical boundary. Optical systems that combine reflection and refraction are known as **catadioptric systems**.

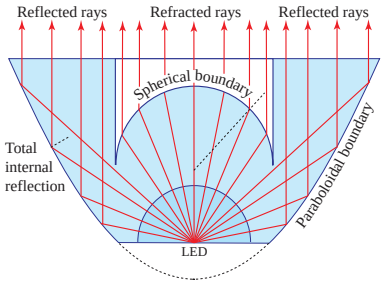


Figure 1.4-3 Cross section of a **collimator** for LED light. LED collimators come in many configurations but most make use of both total internal reflection and refraction to provide rays of light that are approximately parallel at the exit. Such devices are often fabricated from molded acrylic or polycarbonate plastic, which have refractive indices similar to that of glass ($n \approx 1.5$). The diameter of the narrow end of the optic shown is ≈ 1 cm.

Aspheric Optics

An equation for a convex aspherical (nonspherical) surface between media of refractive indices n_1 and n_2 can be determined such that all rays (not only those that are paraxial) originating at an axial point P_1 at a distance z_1 to the left of the surface are imaged onto an axial point P_2 at a distance z_2 to its right, as illustrated in Fig. 1.4-4. The imaging is then **aberration-free**.

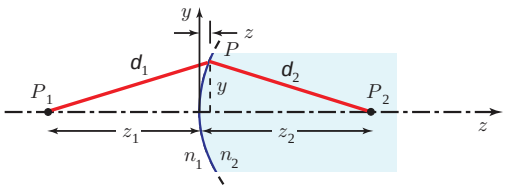


Figure 1.4-4 Construction for an aberration-free imaging surface that corresponds to Fig. 1.4-1(*a*).

In accordance with Fermat's principle, the optical path length associated with the rays in Fig. 1.4-4 then satisfies $n_1 d_1 + n_2 d_2 = \text{constant} = n_1 z_1 + n_2 z_2$. This constitutes an equation that defines the aspherical surface, which is written in Cartesian coordinates as $n_1\sqrt{(z+z_1)^2+y^2} + n_2\sqrt{(z_2-z)^2+y^2} = n_1 z_1 + n_2 z_2$. Given z_1 and z_2, as well as n_1 and n_2, this equation relates y to z and therefore defines the surface. The use of aspheric optics often circumvents the necessity of using complex multicomponent spherical optics and therefore serves to simplify an optical system.

1.5 LENSES

Spherical Lenses

A **spherical lens** is a transparent material bounded by two spherical surfaces. It is defined by the radii of curvature of its two surfaces, R_1 and R_2, its thickness Δ, and the refractive index n of the material from which it is fabricated, as exhibited in Fig. 1.5-1. Alternative appellations are **biconvex lens** and **double-convex lens**.

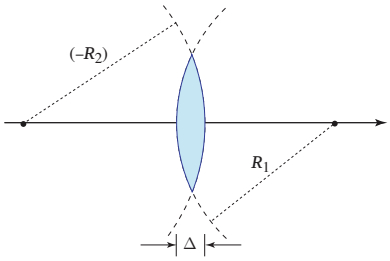


Figure 1.5-1 A glass spherical lens, also known as a biconvex or double-convex lens, can be viewed as a combination of two spherical boundaries, air-to-glass (at left) and glass-to-air (at right). As seen by a ray entering from the left, the air-to-glass boundary is convex (with radius of curvature R_1) while the glass-to-air boundary is concave (with radius of curvature $-R_2$).

As sketched in Fig. 1.5-2(a), a ray crossing the left surface at height y and angle θ_1 with respect to the z axis is traced by applying (1.4-1) for a spherical boundary at that surface to obtain the inclination angle θ of the refracted ray. That ray is then extended until it meets the right surface, whereupon (1.4-1) is used once again, with θ replacing θ_1, to obtain the inclination angle θ_2 of the ray after refraction from the right surface. For a lens of arbitrary thickness Δ, the results are generally quite complex.

Thin Spherical Lens. For a **thin lens**, however, the ray emerges from the lens at roughly the same height y at which it entered, and the results simplify considerably. Under that assumption, the following relations apply (derivations follow):

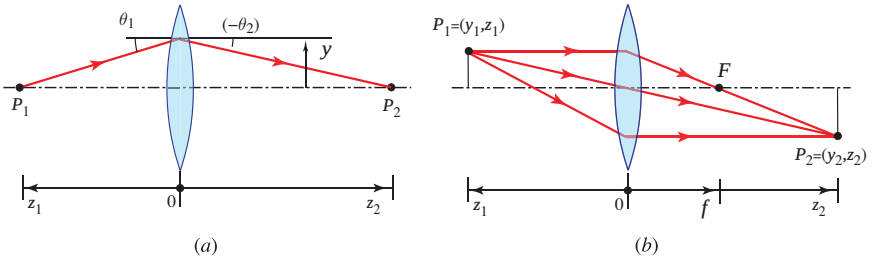


Figure 1.5-2 (a) Ray bending by a thin spherical lens. (b) Image formation by a thin spherical lens.

- The angles of the refracted and incident rays in Fig. 1.5-2(a) are related by

$$\theta_2 = \theta_1 - \frac{y}{f}, \tag{1.5-1}$$

where the **focal length** f is given by

$$\boxed{\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right).} \tag{1.5-2}$$

Focal Length
Thin Spherical Lens

- All rays originating from a point $P_1 = (y_1, z_1)$ meet at a point $P_2 = (y_2, z_2)$, as portrayed in Fig. 1.5-2(b), where

$$\boxed{\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f}} \tag{1.5-3}$$

Imaging Equation
(Paraxial Rays)

and

$$\boxed{y_2 = -\frac{z_2}{z_1} y_1.} \tag{1.5-4}$$

Magnification

These results are identical to those obtained for the spherical mirror, as provided in (1.2-4) and in the derivation surrounding Fig. 1.2-8.

The equations provided above indicate that each point in the $z = z_1$ plane is imaged onto a corresponding point in the $z = z_2$ plane. The magnification $-z_2/z_1$ is unity when $z_1 = z_2 = 2f$, and the focal length f completely determines the effect of the lens on paraxial rays. As indicated earlier, P_1 and P_2 are measured in coordinate systems pointing to the left and to the right, respectively. For the biconvex lens shown in Fig. 1.5-1, R_1 is positive and R_2 is negative, for reasons described in the caption to that figure, so that the two terms of (1.5-2) add to provide a positive focal length.

□ **Derivation of Relations for Image Formation by a Thin Spherical Lens.** Equations (1.5-1) and (1.5-3) for the thin spherical lens may be obtained from (1.4-1) and the definition of the focal length provided in (1.5-2). A ray at angle θ_1 and at height y refracts at the left surface in accordance with (1.4-1) and its angle is altered to $\theta = \theta_1/n - [(n-1)/nR_1]y$, where R_1 is the radius of the left surface. At the right surface, the angle is altered to $\theta_2 = n\theta - [(1-n)/R_2]y$, where R_2 is the radius of the right surface. The ray height does not change since the lens is thin. Combining these two equations leads to $\theta_2 = n\{\theta_1/n - [(n-1)/nR_1]y\} - [(1-n)/R_2]y = \theta_1 - (n-1)y(1/R_1 - 1/R_2)$. We now invoke the relation $\theta_2 = \theta_1 - (y/f)$, in accordance with (1.5-1). If $\theta_1 = 0$, then $\theta_2 = (-y/f)$ and $z_2 \approx (y/-\theta_2) = f$, where f is the focal length of the lens. In general, $\theta_1 \approx y/z_1$ and $-\theta_2 \approx y/z_2$. Finally, using (1.5-1), we obtain $-y/z_2 = y/z_1 - y/f$, from which (1.5-3) follows. Equation (1.5-4) can be derived by using an approach similar to that employed in connection with Fig. 1.2-8. ■

Aberrations. Again, it is emphasized that the foregoing results are applicable only for paraxial rays; the presence of nonparaxial rays results in aberrations, as illustrated in Fig. 1.5-3.

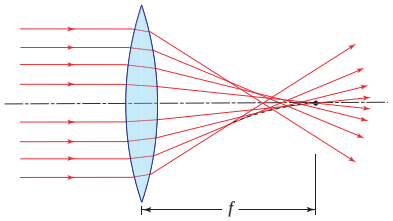


Figure 1.5-3 Nonparaxial rays do not meet at the paraxial focus. The dashed envelope of the refracted rays is the caustic.

Convex and Concave Lenses. Lenses are transparent optical components that bend rays in a manner prescribed by the shapes of their surfaces. Lenses ground or molded from a single piece of material (glass and plastic are favored in the visible region) are called **simple lenses**, whereas those that comprise multiple simple lenses, usually juxtaposed along a common axis, are known as **compound lenses**.

The surface of a lens can be convex or concave, depending on whether it projects out of, or recedes into, the body of the lens, respectively, or it can be planar, indicating that it has a flat surface. A **cylindrical lens** is curved in only one direction; if the axis of the cylinder is aligned with the x axis of the coordinate system, it has a focal length f for rays in the y–z plane, but has no focusing power for rays in the x–z plane. A lens in which one surface is convex and the other concave is called a **meniscus lens**, which is often used in spectacles. A lens in which one or both surfaces have a shape that is neither spherical nor cylindrical is known as an **aspheric lens**. Most commonly encountered lenses are spherical although aspheric lenses are widely used.

Several different types of lenses are illustrated in Fig. 1.5-4. Biconvex and plano-convex lenses result in **ray convergence** and are useful for image formation, as depicted

in Fig. 1.5-2(*b*). Biconcave and plano-concave lenses lead to **ray divergence** and are used in projection and focal-length expansion.

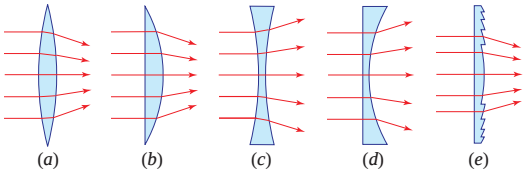


Figure 1.5-4 Lenses: (*a*) Biconvex. (*b*) Plano-convex. (*c*) Biconcave. (*d*) Plano-concave. (*e*) Fresnel. The Fresnel lens displayed in (*e*) is the counterpart of the plano-convex lens shown in (*b*).

Fresnel Lenses

Imaging Fresnel Lenses.

A **Fresnel lens** is constructed by removing various constant-thickness portions of material from a conventional lens, which are nonrefracting and therefore superfluous. For example, the Fresnel-lens illustrated in Fig. 1.5-4(*e*), which is the equivalent of the plano-convex conventional lens depicted in Fig. 1.5-4(*b*), consists of a set of concentric surfaces, each with a curvature identical to that of the plano-convex lens at the same radius. While such a lens can produce a sharp image, its imaging capability generally falls short of that of a conventional lens, principally because of diffraction at the discontinuities. However, the Fresnel design allows for the construction of thin, light, and inexpensive plastic lenses that have short focal lengths and a broad variety of sizes ranging from micrometers to meters. Fresnel lenses can be converging, diverging, or cylindrical.

Nonimaging Fresnel Lenses.

Nonimaging Fresnel lenses, which are more economical to fabricate, are widely used in lighting applications, as depicted in Example 1.5-1. They can be made of glass or plastic in a broad range of sizes. Large glass nonimaging Fresnel lenses were originally developed by Fresnel for use in lighthouses in the early nineteenth century (oil lamps served as the source of light). Today, they are widely used as lenses for directing the light emitted by LEDs, and in automobile headlights and taillights. Nonimaging Fresnel lenses are often economically fabricated by replacing the curved spherical ring segments by flat surfaces tilted at the average angle of the ring's surface, so the tilt is absent at the center of the lens and steepest at its edges. Such a device can thus be pictured as a flat substrate crowned with a collection of raised concentric rings, each with the cross section of a right triangle whose hypotenuse has a length that decreases with increasing radius. Nonimaging Fresnel lenses can also be fabricated in the form of large plastic sheets that are as thin as 1 mm; these are widely used in light-gathering applications such as solar-power collection.

EXAMPLE 1.5-1. ***LED Optics Using Nonimaging Fresnel Lenses.*** LED optics that make use of molded-plastic, nonimaging Fresnel lenses of different types offer light beams with various characteristics. Several examples are illustrated.

Figure 1.5-5 (*a*) A narrow spot of light is provided by a collimating optic such as that shown in Fig. 1.4-3, covered with a flat transparent plate. (*b*) A medium or wide spot is produced by incorporating a diverging Fresnel lens in the optic. (*c*) An elliptical spot of light is created by using a Fresnel prism or Fresnel biprism [Fig. 1.3-6(*b*)]. (*d*) An optic with a flat, roughened surface scatters the light and offers softer illumination.

Graded-Index Lenses

A graded-index (GRIN) material has a refractive index that varies with position in accordance with a continuous function $n(\mathbf{r})$. Such materials are usually fabricated by adding impurities (dopants) with controlled concentrations to the underlying material. Since a GRIN medium is inhomogeneous, the optical rays follow curved trajectories instead of straight lines, and Hero's principle does not apply. However, Fermat's principle remains applicable and $n(\mathbf{r})$ can be chosen in such a way that the GRIN plate has the same effect on light rays as a conventional optical component, such as a lens or prism. **Graded-index lenses** will be discussed in Sec. 2.5.

1.6 OPTICAL FIBERS

Light Guiding

Light may be guided from one location to another by making use of a set of lenses or mirrors, as schematically illustrated in Figs. 1.6-1(a) and (b), respectively. However, because refractive elements such as lenses are usually partially reflective, and reflective elements such as mirrors are usually partially absorptive, the cumulative loss of optical power mounts rapidly as the number of guiding elements is increased. Although it is possible to fabricate components in which these effects are minimized (e.g., antireflection-coated lenses), the assembly of such components into an integrated system is both cumbersome and costly.

Fortunately, there is an attractive alternative. Light may be guided from one location to another via total internal reflection at the dielectric interface between two media of different refractive indices, as portrayed in Fig. 1.6-1(c). Rays are then repeatedly reflected via total internal reflection, a process that is devoid of refraction and absorption. Optical fibers are ideal for guiding light over long distances in this manner, with minimal loss of optical power.

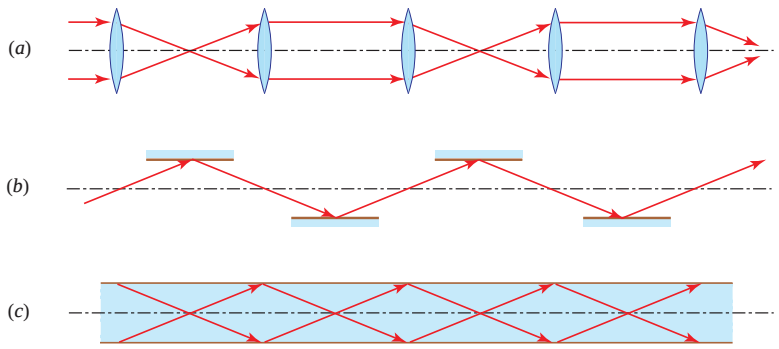


Figure 1.6-1 Guiding light via: (a) lenses, (b) mirrors, (c) total internal reflection.

Total Internal Reflection

An optical fiber is a light conduit constructed from two concentric transparent cylinders, usually glass or plastic. As portrayed in Fig. 1.6-2, it consists of a central **core**Fibers, optical!core in which the light is guided, embedded in an outer **cladding**. The core is a material of refractive index n_1 while the cladding is a material of slightly smaller refractive index, $n_2 < n_1$.

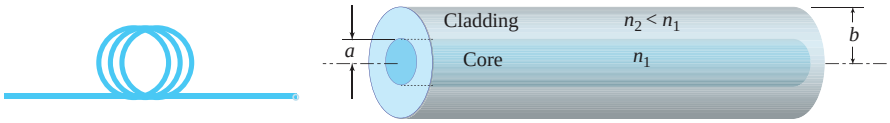


Figure 1.6-2 An optical fiber is a cylindrical dielectric waveguide with a core of refractive index n_1 and a cladding of slightly lower refractive index, $n_2 < n_1$. In conventional optical fibers, known as **step-index fibers**, the refractive indices in the core and in the cladding are independent of position. Examples of standard core-to-cladding diameter ratios (in units of μm/μm) are $2a/2b = 8/125$, $50/125$, $62.5/125$, $85/125$, and $100/140$.

As illustrated in Fig. 1.6-3, light rays traveling in the fiber core undergo total internal reflection at the core–cladding boundary if their angle of incidence $\bar{\theta}$ is greater than the critical angle specified in (1.3-2), i.e., if $\bar{\theta} > \theta_c = \sin^{-1}(n_2/n_1)$. Hence, rays that make an angle $\theta = 90^\circ - \bar{\theta}$ with respect to the optical axis of the fiber will be confined to the fiber core if $\theta < \bar{\theta}_c$, where $\bar{\theta}_c = 90^\circ - \theta_c = \cos^{-1}(n_2/n_1)$, and will be guided without refraction into the cladding and without loss. Rays at greater inclinations to the fiber axis will lose a portion of their power into the cladding at each reflection and are not guided.

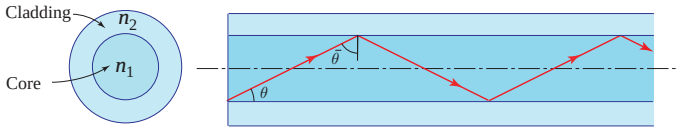


Figure 1.6-3 End view and cross section of an optical fiber that includes the fiber axis. Light rays can be guided by multiple total internal reflections. The angle θ is measured with respect to the axis of the optical fiber so its complement $\bar{\theta} = 90^\circ - \theta$ represents the angle of incidence at the dielectric interface at the core–cladding boundary.

Ray Paths

Rays whose paths are confined to planes that pass through the fiber axis, such as that displayed in Fig. 1.6-3, are known as meridional rays. All other rays are called skewed rays.

Meridional Rays. As illustrated in Fig. 1.6-4, rays confined to meridional planes that pass through the fiber axis have a particularly simple guiding condition. These rays intersect the fiber axis and repeatedly reflect in a given plane without any change in the angle of incidence. Meridional rays are guided if $\theta < \bar{\theta}_c = \cos^{-1}(n_2/n_1)$. Since $n_1 \approx n_2$, $\bar{\theta}_c$ is usually small and the guided rays are approximately paraxial.

Figure 1.6-4 The trajectory of a meridional ray lies in a plane that passes through the fiber axis. The ray is guided if $\theta < \bar{\theta}_c = \cos^{-1}(n_2/n_1)$.

Skewed Rays. An arbitrary ray in the fiber is identified by its plane of incidence, which is the plane parallel to the fiber axis through which the ray passes, and by the angle θ it makes with that axis, as illustrated in Fig. 1.6-5. The plane of incidence intersects the core–cladding cylindrical boundary at an angle ϕ with respect to the normal to the boundary and the plane of incidence lies at a distance R from the fiber axis. The ray is identified by its angle θ with respect to the fiber axis and by the angle ϕ of its plane. When $\phi \neq 0$ ($R \neq 0$) the ray is said to be skewed (for meridional rays, $\phi = 0$ and $R = 0$). A skewed ray reflects repeatedly into planes that make the same angle ϕ with the core–cladding boundary and follows a helical trajectory confined within a cylindrical shell of inner and outer radii R and a, respectively, as illustrated in Fig. 1.6-5. The projection of the trajectory onto the transverse (x–y) plane is a regular polygon that is not necessarily closed. In common with a meridional ray, the condition for a skewed ray to always undergo total internal reflection is that its angle with respect to the z axis be smaller than the complementary critical angle, i.e., $\theta < \overline{\theta}_c$.

Figure 1.6-5 A skewed ray lies in a plane offset from the fiber axis by a distance R. The ray is identified by the angles θ and ϕ. It follows a helical trajectory confined within a cylindrical shell whose inner and outer radii are R and a, respectively. The projection of the ray on the transverse plane is a regular polygon that is not necessarily closed.

Acceptance Angle and Numerical Aperture

A ray incident from air into an optical fiber becomes a guided ray if, upon refraction into the core, it makes an angle θ with respect to the fiber axis that is smaller than the complementary critical angle $\overline{\theta}_c$. As is understood from Fig. 1.6-6(a), applying Snell's law (1.1-4) at the air–core boundary for an acceptance angle θ_a in air corresponding to $\overline{\theta}_c$ in the core yields $1 \cdot \sin\theta_a = n_1 \sin\overline{\theta}_c$, since the refractive index of air is 1. This in turn leads to $\sin\theta_a = n_1\sqrt{1-\cos^2\overline{\theta}_c} = n_1\sqrt{1-(n_2/n_1)^2} = \sqrt{n_1^2 - n_2^2}$, since $\overline{\theta}_c = \cos^{-1}(n_2/n_1)$. The **acceptance angle** of a fiber in air,

$$\theta_a = \sin^{-1}\left(\sqrt{n_1^2 - n_2^2}\right) = \sin^{-1}(\mathrm{NA}), \tag{1.6-1}$$

Acceptance Angle
Optical Fiber

therefore specifies the cone of external rays that are guided by the fiber. Rays incident at angles greater than θ_a are refracted into the fiber but are guided only for short distances since they fail to undergo total internal reflection.

The **numerical aperture** (NA) of the fiber is defined as

$$\mathrm{NA} \equiv \sin\theta_a = \sqrt{n_1^2 - n_2^2} \approx n_1\sqrt{2\Delta}\,, \tag{1.6-2}$$

Numerical Aperture
Optical Fiber

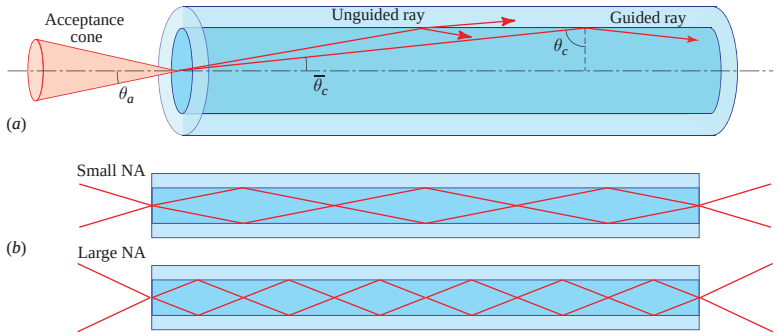


Figure 1.6-6 (a) Acceptance angle θ_a of a fiber. Rays within the acceptance cone are guided by total internal reflection once inside the core. The numerical aperture (NA) of the fiber, defined by $\text{NA} = \sin\theta_a$, assumes a value between 0 and 1. The angles θ_a and $\overline{\theta}_c$ are typically quite small but are exaggerated in the diagram for clarity. (b) The light-gathering capacity of a small NA fiber lies below that of a large NA fiber.

where the fractional refractive-index change is given by $\Delta \equiv (n_1^2 - n_2^2)/2n_1^2 \approx (n_1 - n_2)/n_1 \ll 1$ since $n_1 + n_2 \approx 2n_1$. As illustrated in Fig. 1.6-6(b), the numerical aperture, which assumes a value between zero and unity, characterizes the light-gathering capacity of the fiber. When guided rays arrive at the terminus of a fiber, they are refracted into a cone of angle θ_a that forms a mirror image of their entrance cone. Hence, the numerical aperture is a crucial design parameter for coupling light into and out of optical fibers. Tiny spherical glass balls are sometimes used as lenses to effect this coupling.

EXAMPLE 1.6-1. ***Acceptance Angle and Numerical Aperture for a Silica-Glass Fiber.*** In accordance with (1.6-2), a silica-glass fiber with $n_1 = 1.475$ and $n_2 = 1.460$ has a numerical aperture NA = 0.21 and an acceptance angle $\theta_a = 12.1°$. The refractive index n_1 for silica glass ranges from 1.44 to 1.46, depending on the wavelength, so that Δ typically lies between 0.001 and 0.02. Silica glass, also known as fused silica, is amorphous silicon dioxide (SiO_2). It is widely used in fiber optics because of its excellent optical and mechanical properties, and the fact that its refractive index can be readily modified by doping (e.g., with GeO_2). The loss per unit length of a silica-glass fiber at the wavelength of its maximum transparency is ≈ 0.15 dB/km ($\approx 3.4\%$), an exceptionally low value. Today, optical fibers are fabricated from many materials and take many forms, including photonic-crystal, specialty, multimaterial, and multifunctional versions. They play central roles in many areas of optics and photonics, particularly in sensing, security, transportation, defense, and biomedicine.

EXAMPLE 1.6-2. ***Acceptance Angle and Numerical Aperture for an Uncladded Fiber.*** For a silica-glass fiber with $n_1 = 1.46$ and $\Delta = (n_1 - n_2)/n_1 = 0.01$, according to (1.6-2) the complementary critical angle $\overline{\theta}_c = \cos^{-1}(n_2/n_1) = 8.1°$ and the acceptance angle $\theta_a = 11.9°$, corresponding to a numerical aperture $\text{NA} = 0.206$. By comparison, a fiber with silica-glass core ($n_1 = 1.46$) and a cladding with a substantially smaller refractive index, $n_2 = 1.064$, has $\overline{\theta}_c = 43.2°$, $\theta_a = 90°$, and $\text{NA} = 1$. Rays incident from all directions are then guided since they fall within a cone of angle $\overline{\theta}_c = 43.2°$ inside the core. Likewise, for a totally uncladded fiber ($n_2 = 1$), we have $\overline{\theta}_c = 46.8°$, and rays incident from air at any angle are again refracted into guided rays, which provides maximum light-gathering capacity. However, uncladded fibers are generally not used as optical waveguides for fiber-optic communications applications because they support a large number of modes.

EXAMPLE 1.6-3. ***Coupling Efficiency for an Optical Fiber.*** It is shown that the power collected by an optical fiber from a source of optical power P_0, whose power per unit solid angle distributed as $I(\theta) = (P_0 \cos\theta)/\pi$ where θ is the angle with respect to the axis of a fiber, is given by $P_{\text{col}} = (\text{NA})^2 P_0$, where NA is the numerical aperture of the fiber. The power collected by the fiber is determined by integrating the optical power distribution over the solid angle of the fiber's acceptance cone (angle θ_a). The collected power is thus given by $P_{\text{col}} = \frac{1}{\pi} P_0 \int_0^{2\pi} \int_0^{\theta_a} \cos\theta' \sin\theta' \, d\theta' \, d\phi =$

$2P_0 \int_0^{\theta_a} \cos\theta' \sin\theta' \, d\theta' = (P_0/2) \sin^2\theta|_0^{\theta_a} = (P_0/2)[1 - \cos 2\theta_a]$. Since (1.6-1) dictates that $\theta_a = \sin^{-1}(\mathrm{NA})$, the collected power can be written as $P_{\mathrm{col}} = (P_0/2)[1 - \cos(2\sin^{-1}(\mathrm{NA}))] = P_0\,(\mathrm{NA})^2$. The coupling efficiency is therefore given by $P_{\mathrm{col}}/P_0 = (\mathrm{NA})^2$, where NA is the numerical aperture of the fiber. The effects of loss and Fresnel reflection are not accommodated by ray optics.

EXAMPLE 1.6-4. ***Numerical Aperture for a Butt-Coupled Optical Fiber.*** If a planar light-emitting diode of refractive index n_s is bonded to an optical fiber whose cross-sectional area is larger than the LED emission area, we show that the numerical aperture is determined from the relation $n_s\mathrm{NA} = \sqrt{n_1^2 - n_2^2}$. Our point of departure is the formula for the numerical aperture of an optical fiber in air, provided in (1.6-2). If the fiber is butt-coupled to a medium of refractive index n_s instead of to air, Snell's law in the derivation must be applied at the medium–core boundary rather than at an air–core boundary. Hence, the acceptance angle θ_s corresponding to the complementary critical angle $\bar{\theta}_c$ in the core follows from the use of $n_s \sin\theta_s = n_1 \sin\bar{\theta}_c$, which provides $\sin\theta_s = (n_1/n_s)\sqrt{1 - \cos^2\bar{\theta}_c} = (n_1/n_s)\sqrt{1 - (n_2/n_1)^2} = (1/n_s)\sqrt{n_1^2 - n_2^2}$. The numerical aperture of the butt-coupled fiber is therefore determined from the expression $n_s\mathrm{NA} = \sqrt{n_1^2 - n_2^2}$.

BIBLIOGRAPHY

General Optics

A. M. Gretarsson, *A Practical Guide to Laboratory Optics*, Cambridge University Press, 2021.

B. D. Guenther, *Modern Optics Simplified*, Oxford University Press, 2020.

B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, 3rd ed. 2019, Chaps. 1–13.

B. D. Guenther and D. G. Steel, eds. *Encyclopedia of Modern Optics*, Elsevier, 2nd ed. 2018.

F. L. Pedrotti, L. M. Pedrotti, and L. S. Pedrotti, *Introduction to Optics*, Cambridge University Press, 3rd ed. 2018.

T.-C. Poon and T. Kim, *Engineering Optics with* MATLAB, World Scientific, 2nd ed. 2018.

E. Hecht, *Optics*, Pearson, 5th ed. 2016.

C. A. DiMarzio, *Optics for Engineers*, CRC Press/Taylor & Francis, 2012.

M. Mansuripur, *Classical Optics and Its Applications*, Cambridge University Press, 2nd ed. 2009.

A. Walther, *The Ray and Wave Theory of Lenses*, Cambridge University Press, 1995, paperback ed. 2006.

A. Siciliano, *Optics: Problems and Solutions*, World Scientific, 2006.

J. R. Meyer-Arendt, *Introduction to Classical and Modern Optics*, Prentice Hall, 1972, 4th ed. 1995.

F. A. Jenkins and H. E. White, *Fundamentals of Optics*, McGraw–Hill, 1937, 4th revised ed. 1991.

Ray Optics

S. Schwartz, *Geometrical and Visual Optics*, McGraw–Hill, 3rd ed. 2019.

P. D. Lin, *Advanced Geometrical Optics*, Springer, 2017.

V. N. Mahajan, *Fundamentals of Geometrical Optics*, SPIE Optical Engineering Press, 2014.

E. Dereniak and T. D. Dereniak, *Geometrical and Trigonometric Optics*, Cambridge University Press, 2008.

Yu. A. Kravtsov, *Geometrical Optics in Engineering Physics*, Alpha Science, 2005.

J. E. Greivenkamp, *Field Guide to Geometrical Optics*, SPIE Optical Engineering Press, 2004.

M. Katz, *Introduction to Geometrical Optics*, World Scientific, 2002.

R. Ditteon, *Modern Geometrical Optics*, Wiley, 1998.

F. Colombini and N. Lerner, eds., *Geometrical Optics and Related Topics*, Birkhäuser, 1997.

P. Mouroulis and J. Macdonald, *Geometrical Optics and Optical Design*, Oxford University Press, 1997.

G. A. Fry, *Geometrical Optics*, Chilton, 1969, reprinted 1981.

W. T. Welford and R. Winston, *The Optics of Nonimaging Concentrators*, Academic Press, 1978.

O. N. Stavroudis, *The Optics of Rays, Wavefronts, and Caustics*, Academic Press, 1972.

H.-G. Zimmer, *Geometrical Optics*, Springer, 1970.

A. Nussbaum, *Geometric Optics: An Introduction*, Addison–Wesley, 1968.

Fiber Optics

R. Hui and M. O'Sullivan, *Fiber-Optic Measurement Techniques*, Academic/Elsevier, 2nd ed. 2023.

G. Keiser, *Fiber Optic Communications*, Springer, 2021.

F. Mitschke, *Fiber Optics: Physics and Technology*, Springer, 2nd ed. 2016.

J. Hecht, *Understanding Fiber Optics*, Laser Light Press, 5th ed. 2015.

Y. Koike, *Fundamentals of Plastic Optical Fibers*, Wiley–VCH, 2015.

C. K. Kao, Sand from Centuries Past: Send Future Voices Fast (Nobel Lecture in Physics, 2009), in L. Brink, ed., *Nobel Lectures in Physics 2006–2010*, World Scientific, pp. 253–264, 2014.

R. Paschotta, *Field Guide to Optical Fiber Technology*, SPIE Optical Engineering Press, 2010.

M. G. Kuzyk, *Polymer Fiber Optics: Materials, Physics, and Applications*, CRC Press/Taylor & Francis, 2007.

J. Hecht, *City of Light: The Story of Fiber Optics*, Oxford University Press, Revised ed. 2004.

C. K. Kao, *Optical Fiber Systems: Technology, Design, and Applications*, McGraw–Hill, 1982.

Optical System Design

S. Rolt, *Optical Engineering Science*, Wiley, 2019.

D. Popmintchev and T. Popmintchev, *Introduction to Design of Optical Systems*, Kindle, 2018.

D. Malacara-Hernández and Z. Malacara-Hernández, *Handbook of Optical Design*, CRC Press/Taylor & Francis, 3rd ed. 2013.

J. Sasián, *Introduction to Aberrations in Optical Imaging Systems*, Cambridge University Press, 2013.

K. J. Kasunic, *Optical Systems Engineering*, McGraw–Hill, 2011.

R. E. Fischer, B. Tadic-Galeb, and P. R. Yoder, *Optical System Design*, McGraw–Hill, 2nd ed. 2008.

W. J. Smith, *Modern Optical Engineering*, McGraw–Hill, 1966, 4th ed. 2008.

A. Nussbaum, *Optical System Design*, Prentice Hall, 1998.

D. C. O'Shea, *Elements of Modern Optical Design*, Wiley, 1985.

R. Kingslake, *Optical System Design*, Academic Press, 1983.

Popular and Historical

R. J. Weiss, *A Brief History of Light and Those that Lit the Way*, World Scientific, 1996.

A. R. Hall, *All was Light: An Introduction to Newton's Opticks*, Clarendon Press/Oxford University Press, 1993.

R. Kingslake, *A History of the Photographic Lens*, Academic Press, 1989.

M. I. Sobel, *Light*, University of Chicago Press, 1987.

A. I. Sabra, *Theories of Light from Descartes to Newton*, Cambridge University Press, 1981.

V. Ronchi, *The Nature of Light: An Historical Survey*, Harvard University Press, 1970.

W. H. Bragg, *Universe of Light*, Dover, paperback ed. 1959.

I. Newton, *Opticks: or A Treatise of the Reflections, Refractions, Inflections & Colours of Light*, Samuel Smith and Banjamin Walford, Printers to the Royal Society, 1st ed. 1704; 4th ed. 1730, Dover, reissued 1979.

CHAPTER

2

WAVES

2.1 SCALAR WAVES 26

2.2 MONOCHROMATIC SCALAR WAVES 27

2.3 ELEMENTARY SCALAR WAVES 30

2.4 FREQUENCY AND WAVELENGTH 34

2.5 OPTICAL COMPONENTS 35

2.6 ELECTROMAGNETIC WAVES 44

2.7 RANDOM WAVES 51

Christiaan Huygens (1629–1695) proposed a theory of light whereby it propagates via the emission of spherical waves at every point along the wavefront, a construct now known as the Huygens–Fresnel principle.

James Clerk Maxwell (1831–1879) advanced the notion that light is an electromagnetic wave phenomenon. He formulated an extraordinarily important fundamental set of equations that bears his name.

LED Lighting: Devices and Colorimetry. Malvin Carl Teich. Google Books. Published 2024.

Ray optics, despite its simplicity, is eminently successful for describing the collection, guiding, and control of light, as well as for image formation (Chapter 1). However, it does not have the capability of addressing phenomena that rely on the wavelength, spectrum, phase, or color of light. Accommodating those phenomena requires a more advanced, and more complex, theory of light.

Within the confines of **classical optics**, light is more accurately described as an electromagnetic wave phenomenon that obeys the same laws as other forms of electromagnetic radiation, such as radiowaves, microwaves, and X-rays. In this conception of light, called **electromagnetic optics**, light propagates in the form of waves whose electric- and magnetic-field vectors are mutually coupled. This chapter is devoted to exploring the increased reach offered by electromagnetic optics. We also discuss a simplified version of this theory, known as **scalar wave optics**, that relies on the propagation of a single scalar wave rather than on two coupled vector waves.

This approximate theory is far simpler than electromagnetic optics, yet it is capable of explaining a substantial subset of wave phenomena. Scalar wave optics properly describes diffraction and interference, for example, although these phenomena are not considered in any detail in this text since they are not central to the functioning of light-emitting diodes. It is also useful for representing light waves that vary randomly in time, such as those emitted by astronomical bodies such as the sun and stars, as well as by hot objects and light-emitting diodes.

As portrayed in Fig. 2.0-1, electromagnetic optics encompasses scalar wave optics, which in turn encompasses ray optics.

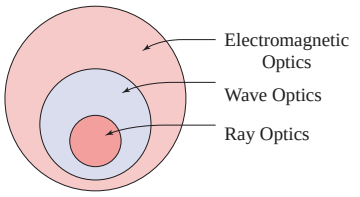


Figure 2.0-1 Electromagnetic optics is a *vector* theory of light comprising coupled electric and magnetic fields that vary in time and space. Wave optics is an approximation to electromagnetic optics and relies on a wavefunction that is a *scalar* function of time and space. Ray optics is the limit of wave optics when the wavelength is very short.

There are, however, optical sources and phenomena that are characteristically quantum mechanical in nature and require **nonclassical optics** for their representation. Accommodating such features requires a quantum version of electromagnetic theory called **quantum optics**, in which the electric- and magnetic-field vectors of electromagnetic theory are promoted to operators in a Hilbert space that satisfy established operator equations and commutation relations. A simplification of quantum optics, known as **photon optics**, in which light propagates in the form of photon streams, provides a suitable approximation for describing many of these quantum effects. Photon optics can be effectively used to augment electromagnetic optics with a number of simple relationships that pertain to the corpuscularity, localization, and fluctuations of quantum fields and energy. The theory of photon optics is useful for elucidating various features of both nonclassical and classical light, as will be discussed in Chapter 3.

Still, when light waves propagate around and through objects whose dimensions are much greater than the wavelength of the light, such as prisms and lenses, neither the corpuscular nor wave nature of light is discernible without careful observation. The particle, wave, and ray approaches then lead to similar outcomes and the propagation of light can be adequately described by rays that obey the set of geometrical rules prescribed by **ray optics**, as presented in Chapter 1.

2.1 SCALAR WAVES

Principles of Scalar Wave Optics

- Light propagates in the form of waves. An optical wave is mathematically described by a real function of position $\mathbf{r} = (x, y, z)$ and time t, denoted $u(\mathbf{r}, t)$ and known as the **wavefunction**. It satisfies a partial differential equation called the **wave equation**,

$$\nabla^2 u - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = 0\,, \tag{2.1-1}$$

Wave Equation in a Medium

where ∇^2 represents the Laplacian operator, which, in Cartesian coordinates is expressed as $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$. Any function that satisfies (2.1-1) represents a possible optical wave.

- In free space, light waves travel at the constant speed c_0. A homogeneous transparent medium such as glass is characterized by a single constant, its refractive index n ($\geqslant 1$). In a medium of refractive index n, light waves travel at a reduced speed,

$$c = \frac{c_0}{n}\,. \tag{2.1-2}$$

Speed of Light in a Medium

The speed of light in free space is $c_0 \approx 3.0 \times 10^8$ m/s = 30 cm/ns = 0.3 mm/ps = 0.3 μm/fs = 0.3 nm/as.

Light propagates in the form of scalar waves that obey the wave equation and travel at the speed of light c .

Superposition

Because the wave equation is linear, the **principle of superposition** applies: if $u_1(\mathbf{r}, t)$ and $u_2(\mathbf{r}, t)$ represent possible optical waves, then $u(\mathbf{r}, t) = u_1(\mathbf{r}, t) + u_2(\mathbf{r}, t)$ also represents a possible optical wave.

Intensity, Power, and Energy

The optical **intensity** $I(\mathbf{r}, t)$, defined as the optical power per unit area (units of W/m^2), is proportional to the average of the squared wavefunction:

$$I(\mathbf{r}, t) = 2\langle u^2(\mathbf{r}, t)\rangle\,. \tag{2.1-3}$$

Optical Intensity

The operation $\langle\cdot\rangle$ denotes averaging over a time interval much longer than the time of an optical cycle, but much shorter than any other time of interest (such as the duration of a pulse of light). The duration of an optical cycle is short: 2×10^{-15} s $= 2$ fs for light of wavelength 600 nm, for example. The quantity $I(\mathbf{r}, t)$ is also called the **irradiance**, a designation widely used in radiometry. There is some arbitrariness in the definition of the wavefunction and its relation to the intensity. For example, (2.1-3) could have been

written without the factor of 2, and concomitantly scaling the wavefunction by a factor of $\sqrt{2}$, in which case the intensity would remain the same. Incorporating the factor of 2 in (2.1-3) proves convenient, however, as will become apparent in the sequel.

Equation (2.1-3) connects the wavefunction $u(\mathbf{r}, t)$ with a physically measurable quantity — the optical intensity. However, the physical significance of the wavefunction itself must await a discussion of electromagnetic waves in Sec. 2.6 since it is associated with the vector field components of electromagnetic optics. The underlying physical origin of the refractive index, as well as the laws that govern its change at the boundary between two different media, are also specified by the principles of electromagnetic optics.

The optical **power** $P(t)$ (units of W) flowing into an area A normal to the direction of propagation of light is the intensity integrated over that area,

$$P(t) = \int_A I(\mathbf{r}, t)\, dA\,. \tag{2.1-4}$$
Optical Power

The optical **energy** E (units of J) collected over a given time interval T is the integral of the optical power over that time interval,

$$E = \int_0^T P(t)\, dt = \int_0^T \int_A I(\mathbf{r}, t)\, dA\, dt\,. \tag{2.1-5}$$
Optical Energy

Graded-Index Media

The wave equation is also approximately applicable for media with refractive indices that are position dependent, but vary slowly over distances of the order of a wavelength. The medium is then said to be locally homogeneous. For such media, the refractive index n in (2.1-2) and the speed of light c in (2.1-1) are replaced by the appropriate position-dependent functions $n(\mathbf{r})$ and $c(\mathbf{r})$, respectively.

2.2 MONOCHROMATIC SCALAR WAVES

We now consider the mathematical representation for a monochromatic scalar wave and chronicle the emergence of the Helmholtz equation from the wave equation. The optical intensity and wavefronts for monochromatic scalar waves are defined.

Wavefunction

A monochromatic wave is represented by a wavefunction $u(\mathbf{r}, t)$ whose time dependence is harmonic, i.e., varies sinusoidally or cosinusoidally at a fixed frequency ν,

$$u(\mathbf{r}, t) = \mathrm{a}(\mathbf{r}) \cos[2\pi\nu t + \varphi(\mathbf{r})]\,. \tag{2.2-1}$$

This wave is illustrated in Fig. 2.2-1(a), where

$$\begin{aligned}
\mathrm{a}(\mathbf{r}) &= \text{amplitude}\\
\varphi(\mathbf{r}) &= \text{phase}\\
\nu &= \text{frequency (Hz or cycles/s)}\\
\omega &= 2\pi\nu = \text{angular frequency (radians/s or s}^{-1}\text{)}\\
T &= 1/\nu = 2\pi/\omega = \text{period (s)}\,.
\end{aligned}$$

Both the amplitude and phase of the wave are generally dependent on position, but the wavefunction is a harmonic function of time with frequency ν at all positions.

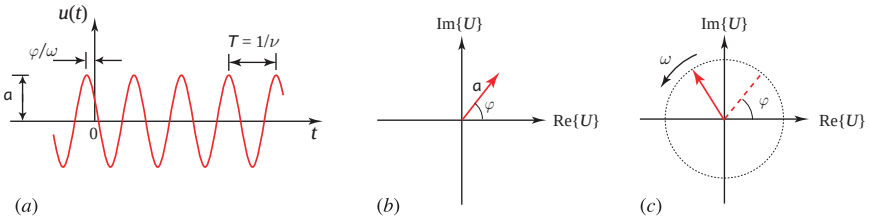


Figure 2.2-1 Representations of a monochromatic wave at a fixed position $\mathbf{r}$: (a) The wavefunction $u(t)$ is a harmonic function of time. (b) The complex amplitude $U = \mathrm{a}\exp(j\varphi)$ is a fixed phasor. (c) The complex wavefunction $U(t) = U\exp(j2\pi\nu t)$ is a phasor that rotates with angular velocity $\omega = 2\pi\nu$ radians/s. Optical waves, which traditionally include the infrared, visible, and ultraviolet regions of the electromagnetic spectrum, have frequencies that stretch from 1×10^{12} to 3×10^{16} Hz, as illustrated in Figs. 2.4-1 and 2.6-1.

Monochromatic waves described by (2.2-1), in which the dependence of the wavefunction on time *and* position is perfectly periodic and predictable, are said to be **coherent** or **deterministic**.

Complex Wavefunction

The real wavefunction $u(\mathbf{r},t)$ set forth in (2.2-1) is conveniently represented in terms of a complex function,

$$U(\mathbf{r},t) = \mathrm{a}(\mathbf{r})\exp[j\varphi(\mathbf{r})]\exp(j2\pi\nu t)\,, \tag{2.2-2}$$

so that

$$u(\mathbf{r},t) = \mathrm{Re}\{U(\mathbf{r},t)\} = \tfrac{1}{2}[U(\mathbf{r},t) + U^*(\mathbf{r},t)]\,, \tag{2.2-3}$$

where the symbol * signifies complex conjugation. The function $U(\mathbf{r},t)$, known as the **complex wavefunction**, provides a complete description of the wave, as does the wavefunction $u(\mathbf{r},t)$, which is simply its real part. Like the wavefunction, the complex wavefunction also satisfies the wave equation,

$$\nabla^2 U - \frac{1}{c^2}\frac{\partial^2 U}{\partial t^2} = 0\,, \tag{2.2-4}$$

Wave Equation (Complex Wavefunction)

and the two functions satisfy the same boundary conditions.

Complex Amplitude

Equation (2.2-2) may be rewritten in the form

$$U(\mathbf{r},t) = U(\mathbf{r})\exp(j2\pi\nu t)\,, \tag{2.2-5}$$

in which the time-independent factor $U(\mathbf{r}) = \mathrm{a}(\mathbf{r})\exp[j\varphi(\mathbf{r})]$ is known as the **complex amplitude** of the wave. The wavefunction $u(\mathbf{r},t)$ is therefore related to the complex

amplitude by

$$u(\mathbf{r},t) = \mathrm{Re}\{U(\mathbf{r})\exp(j2\pi\nu t)\} = \tfrac{1}{2}[U(\mathbf{r})\exp(j2\pi\nu t) + U^*(\mathbf{r})\exp(-j2\pi\nu t)]\,. \tag{2.2-6}$$

At a given position $\mathbf{r}$, the complex amplitude $U(\mathbf{r})$ is a deterministic complex variable [as depicted in Fig. 2.2-1(b)] whose magnitude $|U(\mathbf{r})| = a(\mathbf{r})$ is the amplitude of the wave and whose argument $\arg\{U(\mathbf{r})\} = \varphi(\mathbf{r})$ is its phase. Hence, the complex wavefunction $U(\mathbf{r},t)$, schematized in Fig. 2.2-1(c), depicts a phasor rotating with angular velocity $\omega = 2\pi\nu$ radians/s. Its initial value at $t = 0$ is the complex amplitude $U(\mathbf{r})$.

Helmholtz Equation

Substituting $U(\mathbf{r},t) = U(\mathbf{r})\exp(j2\pi\nu t)$ from (2.2-5) into the wave equation (2.2-4) leads to a differential equation for the complex amplitude $U(\mathbf{r})$:

$$\boxed{\nabla^2 U + k^2 U = 0\,,} \tag{2.2-7}$$

Helmholtz Equation

as formulated by Helmholtz (p. 234), and which has come to be known as the **Helmholtz equation**, where

$$\boxed{k = \frac{2\pi\nu}{c} = \frac{\omega}{c}} \tag{2.2-8}$$

Wavenumber

is referred to as the **wavenumber**. Different solutions obtain from different boundary conditions.

Optical Intensity

The optical intensity is determined by inserting (2.2-1) into (2.1-3):

$$\begin{aligned} 2u^2(\mathbf{r},t) &= 2a^2(\mathbf{r})\cos^2[2\pi\nu t + \varphi(\mathbf{r})] \\ &= |U(\mathbf{r})|^2\{1 + \cos(2[2\pi\nu t + \varphi(\mathbf{r})])\}\,. \end{aligned} \tag{2.2-9}$$

Averaging (2.2-9) over a time longer than an optical period, $1/\nu$, causes the cosinusoidal term in (2.2-9) to vanish, which results in

$$\boxed{I(\mathbf{r}) = |U(\mathbf{r})|^2\,,} \tag{2.2-10}$$

Optical Intensity

a quantity that does not vary in time.

The optical intensity of a monochromatic wave is the absolute square of its complex amplitude.

Wavefronts

Wavefronts are defined as surfaces of equal phase: $\varphi(\mathbf{r}) =$ constant. Because of the periodic nature of phase, the constants are often taken to be multiples of 2π so that $\varphi(\mathbf{r}) = 2\pi q$, where q is an integer. The wavefront normal at position $\mathbf{r}$ is parallel to the

gradient vector $\nabla\varphi(\mathbf{r})$. The components of this vector in a Cartesian coordinate system are $\partial\varphi/\partial x$, $\partial\varphi/\partial y$, and $\partial\varphi/\partial z$; the direction of this vector reveals where the rate of change of the phase is maximum.

Summary: Monochromatic Scalar Waves

- A monochromatic scalar wave of frequency ν is described by a complex wavefunction $U(\mathbf{r},t) = U(\mathbf{r})\exp(j2\pi\nu t)$ that satisfies the wave equation.
- The complex amplitude $U(\mathbf{r})$ satisfies the Helmholtz equation; its magnitude $|U(\mathbf{r})|$ and argument $\arg\{U(\mathbf{r})\}$ are the amplitude and phase of the wave, respectively. The optical intensity is $I(\mathbf{r}) = |U(\mathbf{r})|^2$. The wavefronts are the surfaces of constant phase, $\varphi(\mathbf{r}) = \arg\{U(\mathbf{r})\} = 2\pi q$, where q is integer.
- The wavefunction $u(\mathbf{r},t)$ is the real part of the complex wavefunction, i.e., $u(\mathbf{r},t) = \text{Re}\{U(\mathbf{r},t)\}$. The wavefunction also satisfies the wave equation.

2.3 ELEMENTARY SCALAR WAVES

We now proceed to examine two simple solutions of the Helmholtz equation in a homogeneous medium: the plane wave and the spherical wave. The paraboloidal wave, a useful approximation to the spherical wave, is also introduced. This is followed by a discussion of general paraxial waves, whose wavefront normals make small angles with the axis of the optical system. The complex envelopes of paraxial waves, such as the paraboloidal wave, obey an equation known as the paraxial Helmholtz equation.

Plane Wave

We begin by studying the behavior of a plane wave, which has a complex amplitude

$$U(\mathbf{r}) = A\exp(-j\mathbf{k}\cdot\mathbf{r}) = A\exp\left[-j(k_x x + k_y y + k_z z)\right], \tag{2.3-1}$$

where A is a complex constant called the **complex envelope** and represents the strength of the wave, and $\mathbf{k} = (k_x, k_y, k_z)$ is known as the **wavevector**.[†] Substituting (2.3-1) into the Helmholtz equation (2.2-7) yields the relation $k_x^2 + k_y^2 + k_z^2 = k^2$, so that the magnitude of the wavevector $\mathbf{k}$ is the wavenumber k.

Since the phase of the wave is $\arg\{U(\mathbf{r})\} = \arg\{A\} - \mathbf{k}\cdot\mathbf{r}$, the surfaces of constant phase (wavefronts) obey $\mathbf{k}\cdot\mathbf{r} = k_x x + k_y y + k_z z = 2\pi q + \arg\{A\}$ with q integer. This is an equation that describes parallel planes perpendicular to the wavevector $\mathbf{k}$, which is the basis of the appellation "plane wave." Consecutive planes are separated by the distance $\lambda = 2\pi/k$, so that

$$\boxed{\lambda = \frac{c}{\nu},} \tag{2.3-2}$$

Wavelength

[†] The complex wavefunction for a monochromatic plane wave is written in the form commonly used in electrical engineering: $U(\mathbf{r},t) = A\exp[j(\omega t - \mathbf{k}\cdot\mathbf{r})]$. In the physics literature, however, this wave is usually written as $U(\mathbf{r},t) = A\exp[-i(\omega t - \mathbf{k}\cdot\mathbf{r})]$; correspondence is attained by simply replacing i with $-j$, where $i = j = \sqrt{-1}$. This choice has no bearing on the final result, as is evidenced by observing that the wavefunction $u(\mathbf{r},t)$ in (2.3-3) takes the form of a cosine function, for which $\cos(x) = \cos(-x)$.

where λ is called the **wavelength**. The plane wave has a constant intensity $I(\mathbf{r}) = |A|^2$ everywhere in space so that it carries infinite power. This wave is clearly an idealization since it exists everywhere and at all times.

If the direction of the wavevector $\mathbf{k}$ is taken to lie along the z axis, then $U(\mathbf{r}) = A\exp(-jkz)$ and the corresponding wavefunction associated with (2.2-6) is

$$u(\mathbf{r},t) = |A|\cos\left[2\pi\nu t - kz + \arg\{A\}\right] = |A|\cos\left[2\pi\nu(t - z/c) + \arg\{A\}\right]. \tag{2.3-3}$$

This wavefunction is periodic in time with period $1/\nu$, and periodic in space with period $2\pi/k$, which is equal to the wavelength λ, as illustrated in Fig. 2.3-1. Since the phase of the complex wavefunction, $\arg\{U(\mathbf{r},t)\} = 2\pi\nu(t - z/c) + \arg\{A\}$, varies with time and position as a function of the variable $t - z/c$ (Fig. 2.3-1), the quantity c is called the **phase velocity** of the wave.

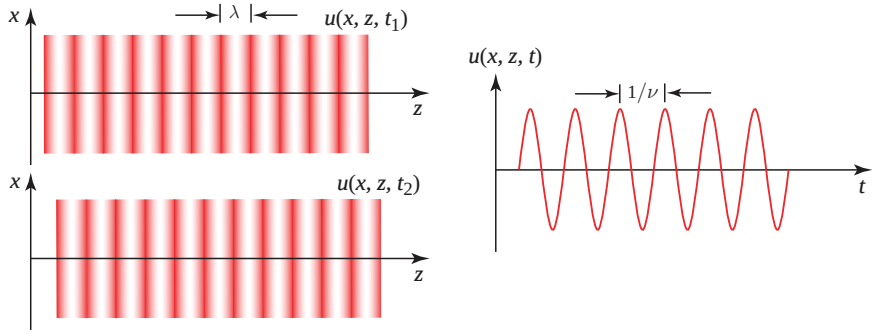


Figure 2.3-1 The wavefunction of a plane wave traveling in the z direction, schematically drawn as a graded red pattern, is a periodic function of z with spatial period λ; and a periodic function of t with temporal period $1/\nu$. The wavefronts (surfaces of constant phase) comprise a set of parallel planes normal to the z axis.

In a medium of refractive index n, the wave has phase velocity $c = c_0/n$ and a wavelength $\lambda = c/\nu = c_0/n\nu$, so that $\lambda = \lambda_0/n$, where $\lambda_0 = c_0/\nu$ is the wavelength in free space. Hence, for a given frequency ν, the wavelength in the medium is reduced relative to that in free space by the factor n. Consequently, the wavenumber $k = 2\pi/\lambda$ is increased relative to that in free space $(k_0 = 2\pi/\lambda_0)$ by the factor n.

***As a monochromatic wave propagates through media of different refractive indices, its frequency remains the same, but its velocity, wavelength, and wavenumber are modified as follows*:**

$$c = \frac{c_0}{n}, \qquad \lambda = \frac{\lambda_0}{n}, \qquad k = nk_0. \tag{2.3-4}$$

Velocity, Wavelength, and Wavenumber of a Monochromatic Wave

Spherical Wave

Another simple solution of the Helmholtz equation, this time in spherical coordinates, is the spherical wave complex amplitude

$$U(\mathbf{r}) = \frac{A_0}{r}\exp(-jkr), \tag{2.3-5}$$

where r is the distance from the origin, $k = 2\pi\nu/c = \omega/c$ is the wavenumber, and A_0 is a constant. The intensity $I(\mathbf{r}) = |A_0|^2/r^2$ is seen to be inversely proportional to the square of the distance. Taking $\arg\{A_0\} = 0$ for simplicity, the wavefronts are the surfaces $kr = 2\pi q$ or $r = q\lambda$, where q is an integer. Hence, the wavefronts are concentric spheres, separated by the radial distance $\lambda = 2\pi/k$, and advance radially at the phase velocity c, as portrayed in Fig. 2.3-2. A wave with complex amplitude $U(\mathbf{r}) =$

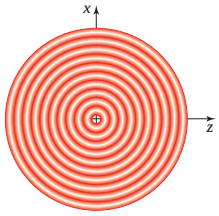


Figure 2.3-2 Cross section of the wavefunction of a spherical wave. The associated wavefronts are a set of concentric spheres.

$(A_0/r)\exp(+jkr)$ is a spherical wave traveling in an inward direction (toward the origin) instead of in an outward direction (away from the origin). A spherical wave originating at the position $\mathbf{r}_0$ has a complex amplitude $U(\mathbf{r}) = (A_0/|\mathbf{r}-\mathbf{r}_0|)\exp(-jk\,|\mathbf{r}-\mathbf{r}_0|)$, and has wavefronts that are spheres centered about $\mathbf{r}_0$.

Paraboloidal Wave

We now consider an approximation for a spherical wave originating at $\mathbf{r} = 0$, at points $\mathbf{r} = (x, y, z)$ that are sufficiently close to the z axis but sufficiently far from the origin, that $\sqrt{x^2+y^2} \ll z$. Were these positions the endpoints of rays beginning at the origin, this would be the paraxial approximation of ray optics. Denoting $\theta^2 = (x^2+y^2)/z^2 \ll 1$, we make use of the following approximation based on a Taylor-series expansion:

$$\begin{aligned} r = \sqrt{x^2+y^2+z^2} &= z\sqrt{1+\theta^2} = z\left(1+\theta^2/2-\theta^4/8+\cdots\right) \\ &\approx z\left(1+\theta^2/2\right) = z+(x^2+y^2)/2z\,. \end{aligned} \tag{2.3-6}$$

Substituting this approximation, $r \approx z+(x^2+y^2)/2z$, into the phase of $U(\mathbf{r})$ in (2.3-5), along with the less accurate but satisfactory approximation $r \approx z$ for the magnitude (which is less sensitive to error than the phase), we arrive at the **Fresnel approximation** of a spherical wave:

$$U(\mathbf{r}) \approx \frac{A_0}{z}\exp(-jkz)\exp\left[-jk\frac{x^2+y^2}{2z}\right]. \tag{2.3-7}$$

Fresnel Approximation of a Spherical Wave

This approximation plays an important role in simplifying the theory of wave transmission through optical components.

The complex amplitude in (2.3-7) may be viewed as representing a plane wave $A_0\exp(-jkz)$ modulated by the factor $(1/z)\exp[-jk(x^2+y^2)/2z]$, with associated phase $k(x^2+y^2)/2z$. This phase factor serves to bend the planar wavefronts of the underlying plane wave into paraboloidal surfaces since the equation for a paraboloid of revolution is $(x^2+y^2)/z =$ constant, as sketched in Fig. 2.3-3. In this region the spherical wave is well approximated by a **paraboloidal wave**. When z becomes very large, the paraboloidal phase factor in (2.3-7) approaches zero so the overall phase of the wave becomes kz. Since the magnitude A_0/z varies slowly with z, the spherical wave eventually approaches the plane wave $\exp(-jkz)$, as illustrated in Fig. 2.3-3.

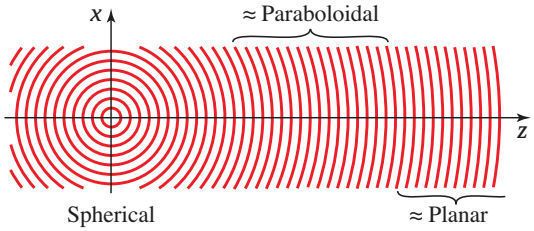


Figure 2.3-3 At points sufficiently far from the origin, but near the z axis, a spherical wave may be approximated by a paraboloidal wave. For points very far from the origin, the spherical wave approaches a plane wave.

Paraxial Waves

A wave is said to be paraxial if its wavefront normals are paraxial rays. One way of constructing a paraxial wave is to start with a plane wave $A\exp(-jkz)$, regard it as a "carrier" wave, and modify or "modulate" its complex envelope A to render it a slowly varying function of position, $A(\mathbf{r})$, whereby the complex amplitude of the modulated wave becomes

$$U(\mathbf{r}) = A(\mathbf{r})\exp(-jkz)\,. \tag{2.3-8}$$

The variation of the envelope $A(\mathbf{r})$, and its derivative with respect to position z, must be slow within the distance of a wavelength $\lambda = 2\pi/k$ so that the wave approximately maintains its underlying plane-wave nature.

The wavefunction of a paraxial wave, $u(\mathbf{r},t) = |A(\mathbf{r})|\,\cos[2\pi\nu t - kz + \arg\{A(\mathbf{r})\}]$, is sketched in Fig. 2.3-4(a) as a function of z at $t = 0$ and $x = y = 0$. It is a sinusoidal function of z with amplitude $|A(0,0,z)|$ and phase $\arg\{A(0,0,z)\}$, both of which vary slowly with z. Since the phase $\arg\{A(x,y,z)\}$ changes little within the distance of a wavelength, the planar wavefronts $kz = 2\pi q$ of the carrier plane wave bend only slightly, so that their normals form paraxial rays, as displayed in Fig. 2.3-4(b).

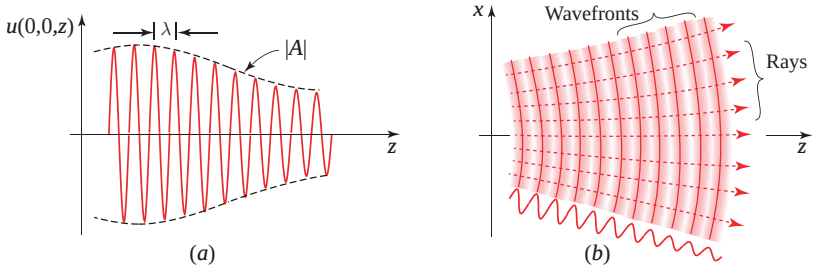


Figure 2.3-4 (a) Wavefunction of a paraxial wave as a function of the axial distance z at $t = 0$ and $x = y = 0$. (b) Sketch of the wavefronts and wavefront normals ("rays") of a paraxial wave in the x–z plane.

The complex envelope $A(\mathbf{r})$ of a paraxial wave, such as the paraboloidal wave, obeys the **paraxial Helmholtz equation**,

$$\boxed{\nabla_T^2 A - j\,2k\frac{\partial A}{\partial z} = 0\,,} \tag{2.3-9}$$

Paraxial Helmholtz Equation

where $\nabla_T^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the transverse Laplacian operator.

□ **Verification that the Paraboloidal Wave Satisfies the Paraxial Helmholtz Equation (2.3-9).** The paraboloidal wave described by $A = (A_0/z)\exp[-jk(x^2 + y^2)/2z]$ has partial derivatives with respect to x given by $\partial A/\partial x = -jxAk/z$ and $\partial^2 A/\partial x^2 = -j(k/z)(x\,\partial A/\partial x + A) = -jk(-jx^2Ak/z + A)/z = -jAk/z - (k/z)^2x^2A$. Similarly, the partial derivatives with respect to y are $\partial^2 A/\partial y^2 = -jAk/z - (k/z)^2y^2A$. Taken together, these results lead to $\nabla_T^2 A = -j2Ak/z - (k/z)^2(x^2 + y^2)A$. Finally, the partial derivative with respect to z can be written as $\partial A/\partial z = -(A_0/z^2)\exp[-jk(x^2 + y^2)/2z] + (A_0/z)[jk(x^2 + y^2)/2z^2]\exp[-jk(x^2 + y^2)/2z] = -A/z + (jk/2z^2)(x^2 + y^2)A$. Substituting this collection of partial derivatives into (2.3-9) confirms that the paraxial Helmholtz equation is satisfied. ■

2.4 FREQUENCY AND WAVELENGTH

As illustrated in Fig. 2.4-1, the range of optical wavelengths in free space encompasses three principal sub-regions: **infrared** (0.760 to 300 μm), **visible** (390 to 760 nm), and **ultraviolet** (10 to 390 nm). The corresponding range of optical frequencies stretches from 1 THz in the far infrared to 30 PHz in the extreme ultraviolet. The infrared, visible, and ultraviolet regions all fall under the rubric "optical" since they make use of similar types of components (e.g., mirrors and lenses). The terahertz (THz) region occupies frequencies that stretch from 0.3 to 3 THz, corresponding to wavelengths extending from 1 mm to 100 μm, so that the THz region partially overlaps the far-infrared band.

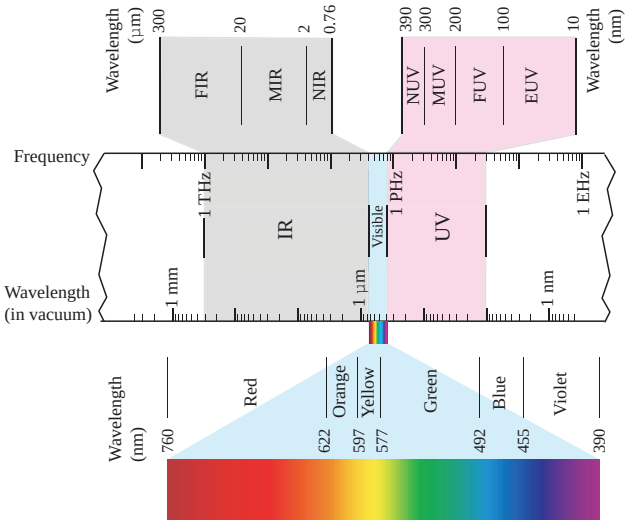


Figure 2.4-1 Free-space optical wavelengths and frequencies. The infrared (IR) region of the spectrum comprises the near-infrared (NIR), mid-infrared (MIR), and far-infrared (FIR) bands. The medium-wave infrared (MWIR) and long-wave infrared (LWIR) subbands both lie within the MIR band; radiation in these regions can penetrate the atmosphere. The ultraviolet (UV) region comprises the near-ultraviolet (NUV), mid-ultraviolet (MUV) or deep-ultraviolet (DUV), far-ultraviolet (FUV), and extreme-ultraviolet (EUV or XUV) bands. The vacuum ultraviolet (VUV) consists of the FUV and EUV bands. The ultraviolet region is also divided into the UVA, UVB, and UVC bands, designations that have chemical and biological significance. Figure 2.6-1 displays the optical region in the context of the broad spectrum of electromagnetic waves that stretches from VLF (very low frequency) waves to γ-rays.

2.5 OPTICAL COMPONENTS

We now turn to an investigation of the effects of various optical components on optical waves (which we often take to be paraxial). In particular, we consider planar mirrors, planar boundaries, transparent plates of arbitrary thickness and refractive index, prisms, and lenses. The results all turn out to be in substantial agreement with those obtained using paraxial ray optics in Chapter 1. We also consider the effects of diffraction gratings on optical waves, a topic that lies beyond the reach of ray optics.

Planar Mirrors

Consider a plane wave of wavevector $\mathbf{k}_1$ incident on a planar mirror located in the $z = 0$ plane in free space. As illustrated in Fig. 2.5-1, a reflected plane wave of wavevector $\mathbf{k}_2$ is created. The angles of incidence and reflection are denoted θ_1 and θ_2, respectively.

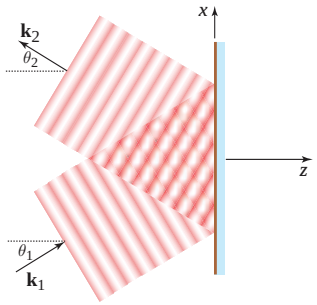


Figure 2.5-1 Reflection of a plane wave from a planar mirror. Phase matching at the surface of the mirror demands that the angles of incidence and reflection be equal. The law of reflection of optical rays thus applies to the wavevectors of plane waves.

The sum of the two waves satisfies the Helmholtz equation if the wavenumber is the same, i.e., if $k_1 = k_2 = k_0$. Certain boundary conditions must be satisfied at the surface of the mirror; since these conditions are the same at all points (x, y), the phases of the two waves must match, i.e.,

$$\mathbf{k}_1 \cdot \mathbf{r} = \mathbf{k}_2 \cdot \mathbf{r} \qquad \text{for all } \mathbf{r} = (x, y, 0)\,. \tag{2.5-1}$$

This phase-matching condition may also be regarded as matching of the tangential components of the two wavevectors in the plane of the mirror. Substituting $\mathbf{r} = (x, y, 0)$, $\mathbf{k}_1 = (k_0 \sin\theta_1, 0, k_0 \cos\theta_1)$, and $\mathbf{k}_2 = (k_0 \sin\theta_2, 0, -k_0 \cos\theta_2)$ into (2.5-1), leads to $k_0 x \sin\theta_1 = k_0 x \sin\theta_2$, from which we obtain $\theta_1 = \theta_2$, thereby confirming that the angles of incidence and reflection must be equal. Hence, the law of reflection for optical rays is applicable to the wavevectors of plane waves.

Planar Boundaries

We now consider a plane wave of wavevector $\mathbf{k}_1$ incident on a planar boundary between two homogeneous media of refractive indices n_1 and n_2. The boundary lies in the $z = 0$ plane. As illustrated in Fig. 2.5-2, a refracted plane wave of wavevector $\mathbf{k}_2$ emerges, as does a reflected plane wave of wavevector $\mathbf{k}_3$.

The combination of the three waves satisfies the Helmholtz equation everywhere if each of the waves has the appropriate wavenumber in the medium in which it propagates, i.e., $k_1 = k_3 = n_1 k_0$ and $k_2 = n_2 k_0$. Since the boundary conditions are invariant to x and y, the phases of the three waves must match, i.e.,

$$\mathbf{k}_1 \cdot \mathbf{r} = \mathbf{k}_2 \cdot \mathbf{r} = \mathbf{k}_3 \cdot \mathbf{r} \qquad \text{for all } \mathbf{r} = (x, y, 0)\,. \tag{2.5-2}$$

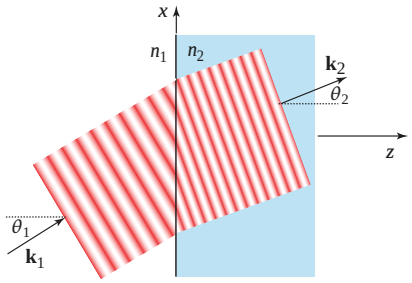


Figure 2.5-2 Refraction of a plane wave at a dielectric boundary. The wavefronts are matched at the boundary so that the distance between wavefronts for the incident wave, $\lambda_1/\sin\theta_1 = \lambda_0/n_1\sin\theta_1$, equals that for the refracted wave, $\lambda_2/\sin\theta_2 = \lambda_0/n_2\sin\theta_2$, from which Snell's law follows.

This phase-matching condition is tantamount to matching the tangential components of the three wavevectors at the boundary plane, as indicated in Sec. 2.6. Since $\mathbf{k}_1 = (n_1k_0\sin\theta_1, 0, n_1k_0\cos\theta_1)$, $\mathbf{k}_3 = (n_1k_0\sin\theta_3, 0, -n_1k_0\cos\theta_3)$, and $\mathbf{k}_2 = (n_2k_0\sin\theta_2, 0, n_2k_0\cos\theta_2)$, where θ_1, θ_2, and θ_3 are the angles of incidence, refraction, and reflection, respectively, it follows from (2.5-2) that $\theta_1 = \theta_3$ and $n_1\sin\theta_1 = n_2\sin\theta_2$, which is Snell's law. Determining the amplitudes and powers of the reflected and refracted waves requires electromagnetic optics, however, since the boundary conditions are not completely specified in wave optics (see Sec. 2.6).

The laws of reflection and refraction of optical rays apply to the wavevectors of plane waves.

Transparent Plates

We turn now to the transmission of optical waves through transparent plates that have arbitrary thickness or refractive-index distributions. Our treatment focuses on the phase shifts and the associated wavefront bending imparted by these components. The expressions developed serve as templates in upcoming sections for establishing the effects imposed on waves by common optical components such as prisms, lenses, and diffraction gratings; these particular components impart phase shifts that bend the wavefronts linearly, quadratically, and periodically, respectively. We do not consider surface reflection and material absorption at this juncture, since these features cannot be accommodated by scalar wave theory.

Transparent Plate of Fixed Thickness and Fixed Refractive Index. Consider first the transmission of a plane wave through a transparent plate of refractive index n and thickness d surrounded by free space. The surfaces of the plate are taken to be at the $z = 0$ and $z = d$ planes. The wave is assumed to be traveling in the z direction and normally incident on the plate, as illustrated in Fig. 2.5-3. External and internal reflections are ignored so the complex amplitude of the wave $U(x, y, z)$ is assumed to be continuous at the boundaries. The ratio $\mathrm{t}(x, y) = U(x, y, d)/U(x, y, 0)$ therefore represents the **complex amplitude transmittance** of the plate, which permits $U(x, y, d)$ to be determined for arbitrary $U(x, y, 0)$ at the input.

Once inside the plate, the wave continues to propagate as a plane wave, but with wavenumber nk_0, so that $U(x, y, z) \propto \exp(-jnk_0z)$. Hence, the complex amplitude transmittance of the plate is given by $U(x, y, d)/U(x, y, 0)$, which yields

$$\mathrm{t}(x, y) = \exp(-jnk_0d)\,, \tag{2.5-3}$$

Transmittance
Transparent Plate

where nd is the optical pathlength. The plate is seen to introduce a phase shift $nk_0d = 2\pi(d/\lambda)$. In this special case, with the plane wave normally incident on the plate, the

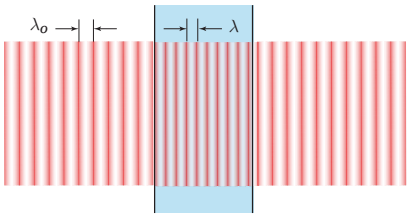


Figure 2.5-3 Transmission of a plane wave at normal incidence through a transparent plate of fixed thickness d and fixed refractive index n. The wavelength inside the material is λ and the plate introduces a phase shift $nk_0d = 2\pi(d/\lambda)$.

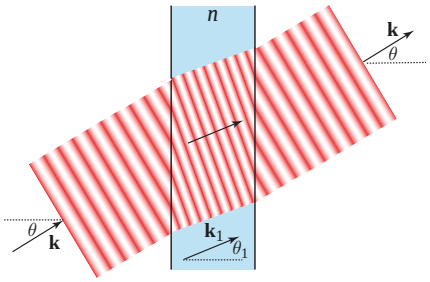


Figure 2.5-4 Transmission of a plane wave arriving at an oblique angle through a transparent plate of fixed thickness and fixed refractive index. The direction of the emerging wave is parallel to that of the incident wave.

power reflectance predicted by electromagnetic optics takes a particularly simple form [see (2.6-23) and Example 2.6-1].

If the wavevector $\mathbf{k}$ of the incident plane wave instead makes an angle θ with respect to the z axis, as portrayed in Fig. 2.5-4, the refracted and transmitted waves are also plane waves, with wavevectors $\mathbf{k}_1$ and $\mathbf{k}$, and angles θ_1 and θ, respectively, where θ_1 and θ are related by Snell's law: $1 \cdot \sin\theta = n \sin\theta_1$. The complex amplitude $U(x, y, z)$ inside the plate is proportional to $\exp(-j\mathbf{k}_1 \cdot \mathbf{r}) = \exp[-jnk_0(z\cos\theta_1 + x\sin\theta_1)]$, so that the complex amplitude transmittance of the plate $U(x, y, d)/U(x, y, 0)$ is given by

$$\mathrm{t}(x, y) = \exp\left(-jnk_0 d\cos\theta_1\right). \tag{2.5-4}$$

If the angle of incidence θ is small (i.e., if the incident wave is paraxial), then the paraxial Snell's law yields $\theta_1 \approx \theta/n$, which is also small, whereupon use of the approximation $\cos\theta_1 \approx 1 - \frac{1}{2}\theta_1^2$ gives rise to $\mathrm{t}(x, y) \approx \exp(-jnk_0 d)\exp(jk_0\theta^2 d/2n)$. If the plate is *sufficiently thin*, and the angle θ is *sufficiently small* such that $k_0\theta^2 d/2n \ll 2\pi$ [or $(d/\lambda_0)\theta^2/2n \ll 1$], then the transmittance of the plate may be roughly approximated by (2.5-3), an expression that is independent of the angle of incidence θ.

Transparent Plate of Fixed Thickness and Varying Refractive Index. Since the thickness and refractive index of the transparent plate appear as the product nd in (2.5-3), a prescribed phase shift may be imparted in two different, but equivalent, ways: 1) by controlling the variation in the thickness of the material with transverse distance from the optical axis (the technique used in fabricating conventional optical components); or 2) by controlling the refractive index of the material with transverse distance from the optical axis (the technique used in fabricating graded-index optical components).

In the latter case, (2.5-3) dictates that the complex amplitude transmittance of a thin transparent planar plate of fixed thickness d_0 and graded refractive index $n(x, y)$ be written as

$$\boxed{\mathrm{t}(x, y) = \exp\left[-jn(x, y)k_0 d_0\right].} \tag{2.5-5}$$

Transmittance
Graded-Index Plate

This equation reveals that the action of any constant-index thin conventional optical component can be mimicked by selecting an appropriate corresponding variation of $n(x, y)$ with x and y, as will be illustrated subsequently for a graded-index lens.

Transparent Plate of Varying Thickness and Fixed Refractive Index. Finally, we consider the situation germane to conventional optical components, in which the thickness of a material of fixed refractive index is sculpted to a specific shape. We restrict our attention to the amplitude transmittance for an arbitrary paraxial wave incident on a thin transparent plate whose thickness $d(x, y)$ varies smoothly as a function of x and y. The plate lies between the planes $z = 0$ and $z = d_0$, which are regarded as planar boundaries that encase an arbitrary optical component, as displayed in Fig. 2.5-5.

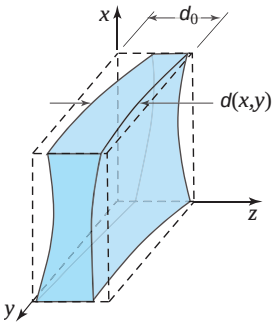


Figure 2.5-5 A transparent plate of arbitrarily varying thickness $d(x, y)$.

In the vicinity of the position $(x, y, 0)$, the incident paraxial wave may be regarded locally as a plane wave traveling along a direction that makes a small angle with the z axis. It crosses a thin plate of material of thickness $d(x, y)$, which is surrounded by thin layers of air whose overall thickness is $d_0 - d(x, y)$, as illustrated in Fig. 2.5-5. In accordance with the approximate relation provided in (2.5-3), the local transmittance is then the product of the transmittances of the thin layer of air of thickness $d_0 - d(x, y)$ and the thin layer of material of thickness $d(x, y)$, which leads to $\mathrm{t}(x, y) \approx \exp[-jnk_0 d(x, y)] \exp[-jk_0(d_0 - d(x, y))]$, and thence to

$$\mathrm{t}(x, y) \approx h_0 \exp[-j(n-1)k_0 d(x, y)]\,, \tag{2.5-6}$$

Transmittance Variable-Thickness Plate

where $h_0 = \exp(-jk_0 d_0)$ is a constant phase factor. This relation is valid in the paraxial approximation, where all angles θ are small, and when the thickness d_0 is sufficiently small such that $(d_0/\lambda_0)\theta^2/2n \ll 1$. This latter condition, which was derived earlier in connection with the transmission of an oblique plane wave through a transparent plate of fixed thickness and fixed refractive index, ensured that the transmittance was approximately independent of the angle of incidence. In the present case, it ensures that (2.5-6) is applicable for paraxial waves.

Prisms

The general expression (2.5-6) for the complex amplitude transmittance of a thin transparent plate of variable thickness is applied to a thin inverted prism of thickness d_0 and small apex angle $\alpha \ll 1$, as portrayed in Fig. 2.5-6. The dependence of d on x is determined by the apex angle α via $\tan\alpha = d/x$. Since $\alpha \ll 1$, $\tan\alpha \approx \alpha$ and $d(x) \approx \alpha x$. The prism is assumed to extend in the y direction, so $d(x, y)$ is independent of y. Hence, $\mathrm{t}(x, y) \approx h_0 \exp[-j(n-1)\alpha k_0 x]$, where $h_0 = \exp(-jk_0 d_0)$. The linear change of phase with increasing x causes the wavevector to acquire a tilt toward the x axis, in accordance with the deflection angle $\theta_d \approx (n-1)\alpha$. The ray-based calculation for the deflection angle reported in (1.3-4) is seen to be in accord with this result.

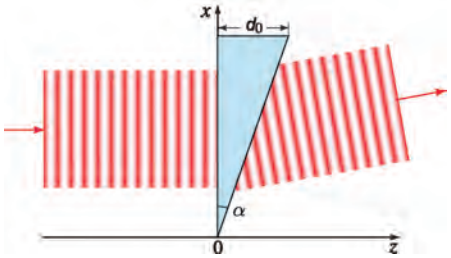


Figure 2.5-6 Transmission of a plane wave through a thin prism. The wave, which is incident in the z direction, tilts toward the x axis after passing through the prism.

EXAMPLE 2.5-1. ***Transmission Through a Biprism and an Axicon.***

- The biprism depicted in Fig. 1.3-6(a) comprises an inverted prism, such as that illustrated in Fig. 2.5-6, juxtaposed with an identical uninverted prism. Taking its thickness to be d_0 and its edge angle $\alpha \ll 1$, the results for the simple prism provided above generalize to $\mathrm{t}(x,y) = h_0\{\exp[-j\,(n-1)\alpha k_0 x] + \exp[+j\,(n-1)\alpha k_0 x]\} = 2h_0 \cos\,[(n-1)\alpha k_0 x]$, with $h_0 = \exp(-jk_0 d_0)$. The biprism thus converts an incident plane wave into a pair of waves that are tilted with respect to each other. The Fresnel biprism portrayed in Fig. 1.3-6(b) behaves in the same way.
- The cone-shaped axicon displayed in Fig. 1.3-6(c) is constructed by rotating the prism cross section depicted in Fig. 2.5-6 about a horizontal axis located at its top edge, from $\phi = -\pi$ to π. The cross section of this device is an isosceles triangle of height d_0 and edge angle $\alpha \ll 1$. Using polar coordinates and integrating over ϕ provides $\mathrm{t}(x,y) = h_0 \int_{-\pi}^{\pi} \exp[-j\,(n-1)\,\alpha(k_0 \cos\phi)\,x - j\,(n-1)\,\alpha(k_0 \sin\phi)\,y]\,d\phi = h_0 \int_{-\pi}^{\pi} \exp[-j\,(n-1)\,\alpha k_0 \sqrt{x^2+y^2}\,\sin(\phi+\theta)]\,d\phi$. Since the integration is over 2π, the integral is independent of θ. Given that $\int_{-\pi}^{\pi} \exp(-ju\sin\phi)\,d\phi = 2\pi J_0(u)$, where $J_0(u)$ is the Bessel function of the first kind and zeroth order, the amplitude transmittance may be rewritten as $\mathrm{t}(x,y) = 2\pi h_0 J_0[(n-1)\,\alpha k_0 \sqrt{x^2+y^2}\,]$. The axicon thus converts an incident plane wave into an infinite number of plane waves, all directed toward its central axis in the form of a cone of half angle $(n-1)\alpha$.

Lenses

We now examine the transmission of optical waves through lenses. Again, our principal emphasis is on the phase shift introduced by these components and on the associated wavefront bending, which we examine via a number of examples. Reflection at the surfaces of these components and absorption in the material are ignored.

Thin Plano-Convex Lens. We once again invoke the general expression (2.5-6) for the complex amplitude transmittance of a thin transparent plate of variable thickness, this time for the plano-convex thin lens displayed in Fig. 2.5-7.

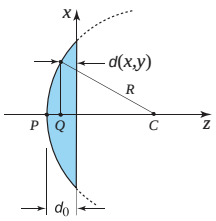


Figure 2.5-7 A thin plano-convex lens imparts a phase proportional to $x^2 + y^2$ to an incident plane wave, thereby transforming it into a paraboloidal wave centered at a distance f from the lens (see Fig. 2.5-9).

Since this lens is the cap of a sphere of radius R, the thickness at the point (x,y) is

$d(x,y) = d_0 - \overline{PQ} = d_0 - (R - \overline{QC})$, or

$$d(x,y) = d_0 - \left[R - \sqrt{R^2 - (x^2+y^2)}\right]. \tag{2.5-7}$$

This expression may be simplified by considering only points for which x and y are sufficiently small in comparison with R so that $x^2 + y^2 \ll R^2$. In that case

$$\sqrt{R^2 - (x^2+y^2)} = R\sqrt{1 - (x^2+y^2)/R^2} \approx R\left(1 - (x^2+y^2)/2R^2\right), \tag{2.5-8}$$

where we have used the same Taylor-series expansion that led to the Fresnel approximation of a spherical wave in (2.3-7). Using this approximation in (2.5-7) then provides

$$d(x,y) \approx d_0 - (x^2+y^2)/2R. \tag{2.5-9}$$

Finally, substitution into (2.5-6) yields

$$\boxed{\mathrm{t}(x,y) \approx h_0 \exp\left[jk_0(x^2+y^2)/2f\right],} \tag{2.5-10}$$

Transmittance
Thin Lens

where

$$f = R/(n-1) \tag{2.5-11}$$

is the focal length of the lens (see Sec. 1.4) and $h_0 = \exp(-jnk_0 d_0)$ is another constant phase factor that is generally of no significance.

EXAMPLE 2.5-2. ***Complex Amplitude Transmittance for a Thin Spherical Lens.***
The complex amplitude transmittance of a thin spherical lens (also called a biconvex lens or a double-convex lens), such as that displayed in Fig. 2.5-8, is readily determined by calculating the amplitude transmittance for a cascade of two plano-convex lenses with focal lengths $f_1 = R_1/(n-1)$ and $f_2 = -R_2/(n-1)$.

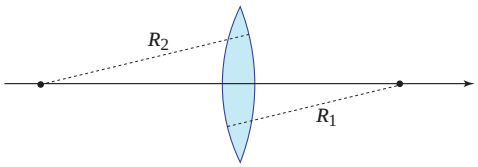


Figure 2.5-8 A thin spherical (biconvex) lens. By convention, the radius of a convex (concave) surface is positive (negative), so R_1 is positive and R_2 is negative.

Forming a product from (2.5-10) leads to

$$\mathrm{t}(x,y) = \mathrm{t}_1(x,y)\,\mathrm{t}_2(x,y) \approx h_{01}\exp\left[jk_0\left(x^2+y^2\right)/2f_1\right]\cdot h_{02}\exp\left[jk_0\left(x^2+y^2\right)/2f_2\right], \tag{2.5-12}$$

where h_{01} and h_{02} are constants and therefore so too is $h_0 = h_{01}h_{02}$. Combining exponentials we arrive at

$$\mathrm{t}(x,y) \approx h_0 \exp\left[jk_0\left(x^2+y^2\right)/2f\right], \tag{2.5-13}$$

with

$$\boxed{\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right).} \tag{2.5-14}$$

Focal Length
Thin Spherical Lens

Equation (2.5-13) mimics (2.5-10) except that the focal length specified in (2.5-14) involves both R_1 and R_2 for the thin spherical lens. The expression for the focal length derived in the context of ray optics for this component, provided in (1.5-2), is identical to (2.5-14).

EXAMPLE 2.5-3. ***Focusing of a Plane Wave by a Thin Lens.***

- Consider a plane wave $U_1(x,y) = \exp(-jk_0 z)$ traveling in a direction parallel to the axis of a thin lens of focal length f and transmittance $\mathrm{t}(x,y) = h_0 \exp\left[jk_0\left(x^2+y^2\right)/2f\right]$. The transmitted wave is described by $U_2(x,y) = U_1(x,y)\cdot \mathrm{t}(x,y) = h_0 \exp\left\{-jk_0\left[z-\left(x^2+y^2\right)/2f\right]\right\}$. The wavefronts of this wave are paraboloids of revolution defined by $z-\left(x^2+y^2\right)/2f =$ constant, with radius of curvature $-f$. The plane wave entering the lens is therefore converted into a paraboloidal wave, which is the Fresnel approximation of a spherical wave, centered at a point at a distance f from the lens, as illustrated in Fig. 2.5-9.

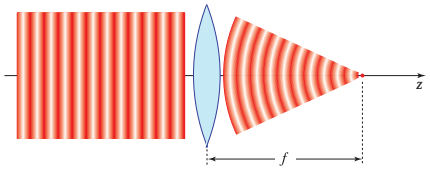


Figure 2.5-9 A thin lens transforms a plane wave into a paraboloidal wave.

- If the incident wave is instead traveling at a small angle θ with respect to the z axis, we have $U_1(x,y) \approx \exp\left[-jk_0(z+\theta x)\right]$, whereupon

$$\begin{aligned} U_2(x,y) = U_1(x,y)\cdot \mathrm{t}(x,y) &\approx h_0 \exp\left\{-jk_0\left[z+\theta x-\left(x^2+y^2\right)/2f\right]\right\} \\ &= h_0 \exp\left\{-jk_0\left[z-\left(x^2-2\theta f x+y^2\right)/2f\right]\right\} \\ &\approx h_0 \exp\left\{-jk_0\left[z-\left([x-\theta f]^2+y^2\right)/2f\right]\right\}. \end{aligned} \quad (2.5\text{-}15)$$

Equation (2.5-15) represents a paraboloidal wave centered about the point $(\theta f, 0, f)$, as illustrated in Fig. 2.5-10.

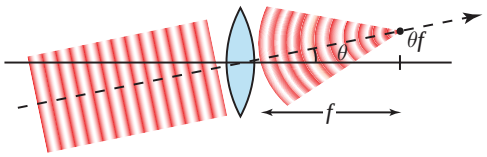


Figure 2.5-10 An incident plane wave traveling through a thin lens at a small angle θ with respect to the z axis becomes a paraboloidal wave centered about the point $(\theta f, 0, f)$ after passage through the lens.

EXAMPLE 2.5-4. ***Imaging Property of a Thin Lens.*** We now consider a paraboloidal wave centered at P_1 ($z = -z_1$) entering a thin lens of focal length f ($z = 0$), as portrayed in Fig. 2.5-11. Since the incident wave is described by $U_1(x,y) \approx \exp\left[-jk_0\left(x^2+y^2\right)/2z_1\right]$ and the lens is

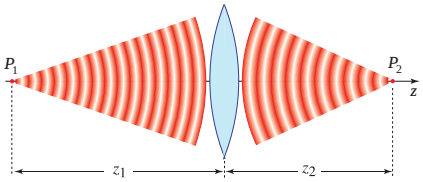


Figure 2.5-11 A thin lens transforms a paraboloidal wave into another paraboloidal wave. The two waves are centered at distances that satisfy the imaging equation.

characterized by the amplitude transmittance $\mathrm{t}(x,y) \approx h_0 \exp\left[jk_0(x^2+y^2)/2f\right]$, the wave leaving the lens satisfies

$$\begin{aligned} U_2(x,y) &\approx \exp\left[-jk_0\left(x^2+y^2\right)/2z_1\right]\cdot \exp\left[jk_0\left(x^2+y^2\right)/2f\right] \\ &= \exp\left[jk_0\left(x^2+y^2\right)/2z_2\right], \end{aligned} \quad (2.5\text{-}16)$$

where $^1/_{z_2} = {}^1/_f - {}^1/_{z_1}$ or $^1/_{z_1} + {}^1/_{z_2} = {}^1/_f$. We conclude that the transmitted wave is a paraboloidal wave centered at $z = z_2$ and that the distances indicated in Fig. 2.5-11 obey the imaging equation,

$$\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f}. \qquad (2.5\text{-}17)$$

Imaging Equation

The imaging equation derived for a thin spherical lens in the context of ray optics, (1.5-3), is identical to (2.5-17).

Graded-Index Lens. The refractive index of a thin plate of uniform thickness can be graded in such a way that it acts as a lens, as schematized in Fig. 2.5-12.

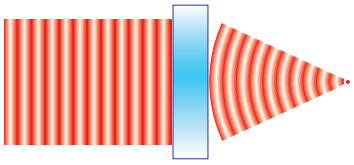


Figure 2.5-12 A graded-index plate can act as a lens.

In accordance with (2.5-5), the transmittance of a graded-index plate of uniform thickness d_0 and arbitrary grading profile $n(x, y)$ is given by $\mathrm{t}(x, y) = \exp[-jn(x, y)k_0 d_0]$. If the refractive index is quadratically graded as $n(x, y) = n_0[1 - \alpha^2(x^2 + y^2)/2]$, subject to $\alpha d_0 \ll 1$, we have $\mathrm{t}(x, y) = h_0 \exp[jn_0\alpha^2 k_0 d_0(x^2 + y^2)/2]$, where $h_0 = \exp(-jn_0 k_0 d_0)$ is a constant phase factor. We therefore arrive at $\mathrm{t}(x, y) = h_0 \exp[jk_0(x^2 + y^2)/2f]$, with $1/2f = n_0\alpha^2 d_0/2$, which is the expression for the amplitude transmittance of a lens of focal length $f = 1/n_0\alpha^2 d_0$. Similar results can be arrived at using ray optics, as mentioned in Sec. 1.5.

Diffraction Gratings

A **diffraction grating** is an optical component that imposes a periodic modulation on the phase or amplitude of an incident wave. It can be fabricated from a transparent plate whose thickness or refractive index is made to vary periodically.

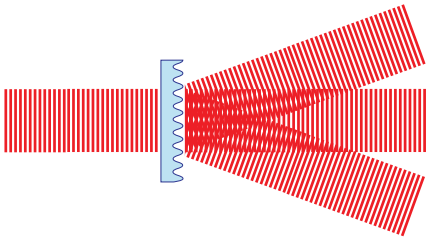


Figure 2.5-13 A thin transparent plate with periodically varying thickness serves as a diffraction grating. It splits an incident plane wave into multiple plane waves traveling in different directions.

We demonstrate the effect of a diffraction grating on an incident plane wave in Fig. 2.5-13. The grating, placed at the $z = 0$ plane, comprises a thin transparent plate whose thickness varies periodically, with period Λ, in the x direction. The plane wave, of wavelength λ, travels at an angle θ_i with respect to the z axis. The diffraction grating converts the incident plane wave into a collection of plane waves, at angles θ_q with respect to the z axis, in accordance with

$$\sin\theta_q = \sin\theta_i + q\frac{\lambda}{\Lambda}\,. \tag{2.5-18}$$

(2.5-18) Grating Equation

This result is also applicable for a reflection diffraction grating, which can be made from a periodically ruled thin film of aluminum evaporated onto a glass substrate.

□ **Proof of the Grating Equation Provided in (2.5-18).**

- Consider a thin transparent plate whose thickness varies as a harmonic function in the x direction, as schematically illustrated in Fig. 2.5-13. In accordance with (2.5-6), the transmittance of a plate of uniform refractive index and varying thickness can be written as $\mathrm{t}(x,y) \approx h_0 \exp[-j(n-1)k_0 d(x,y)]$, where $h_0 = \exp(-jk_0 d_0)$ is a constant phase factor. If the thickness varies as $d(x,y) = \frac{1}{2}d_0[1+\cos(2\pi x/\Lambda)]$, where Λ is the spatial period of the thickness variations, the amplitude transmittance becomes $\mathrm{t}(x) \approx \exp(-jk_0 d_0)\exp[-j(n-1)k_0 d(x)] = h_1 \exp[-j(n-1)(k_0 d_0/2)\cos(2\pi x/\Lambda)]$, where $h_1 = \exp[-j(n+1)(k_0 d_0/2)]$.
- Now consider a plane wave $U_1(x) \propto \exp(-jk_0 x\sin\theta_i)$, traveling at an angle θ_i with respect to the z axis, that is incident on the grating. The transmitted wave, $U_2(x) = \mathrm{t}(x)\cdot U_1(x)$, is determined by recognizing that $\mathrm{t}(x)$ is a periodic function of x with period Λ and can therefore be expanded in a Fourier series as $\mathrm{t}(x) = \sum_q C_q \exp(-jq2\pi x/\Lambda)$, where the C_q are the Fourier coefficients, and $q = 0, \pm1, \pm2, \ldots$ specifies the diffraction order. The amplitude of the component of the transmitted wave that travels at the angle θ_q may therefore be written as $\exp(-jk_0 x\sin\theta_q) = \exp(-jk_0 x\sin\theta_i)\exp(-jq2\pi x/\Lambda) = \exp[-jk_0 x(\sin\theta_i + q2\pi/k_0\Lambda)]$, which leads directly to $\sin\theta_q = \sin\theta_i + q\lambda/\Lambda$, since $2\pi/k_0 = \lambda$. The transmitted wave therefore comprises a collection of plane waves that travel at the angles θ_q specified by (2.5-18) and sketched in Fig. 2.5-13 for $\theta_i \approx 0$.

■

When all angles are small, and when the period of the thickness variation Λ is much greater than the wavelength λ, use of the approximation $\sin\theta \approx \theta$ in (2.5-18) leads to the paraxial approximation for the grating equation,

$$\theta_q \approx \theta_i + q\frac{\lambda}{\Lambda}\,. \tag{2.5-19}$$

(2.5-19) Grating Equation (Paraxial Approximation)

Diffraction Grating as a Spectrum Analyzer. Diffraction gratings are extensively used as filters and spectrum analyzers, particularly in spectroscopy. Since the angles θ_q depend on the wavelength λ (and therefore on the frequency ν), an incident polychromatic wave is separated by the grating into its spectral components, as sketched in Fig. 2.5-14.

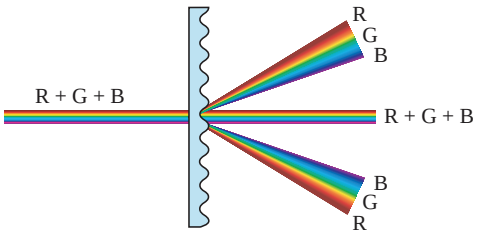


Figure 2.5-14 A diffraction grating directs two waves of different wavelengths, λ_1 and λ_2, into two different directions, θ_1 and θ_2. It therefore serves as a spectrum analyzer or a spectrometer. The letters R, G, and B signify red, green, and blue, respectively.

2.6 ELECTROMAGNETIC WAVES

Principles of Electromagnetic Optics

- Light propagates in the form of electromagnetic waves, which are described by two coupled vector fields that are functions of position and time: the **electric-field vector** $\mathcal{E}(\mathbf{r},t)$ and the **magnetic-field vector** $\mathcal{H}(\mathbf{r},t)$. The description of light in a dielectric medium therefore entails six scalar functions of position $\mathbf{r} = (x,y,z)$ and time t. Each of these components, denoted $u(\mathbf{r},t)$, satisfies the **wave equation**,

$$\nabla^2 u - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = 0\,, \tag{2.6-1}$$

Wave Equation in a Medium

where ∇^2 represents the Laplacian operator, which, in Cartesian coordinates, is expressed as $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$. Any function that satisfies (2.6-1) represents a possible electromagnetic wave.

- In free space, electromagnetic waves travel at a constant speed c_0. In a homogeneous transparent medium of refractive index n ($\geqslant 1$), light waves travel at a reduced **speed of light**,

$$c = \frac{c_0}{n} = \frac{1}{\sqrt{\epsilon\mu}}\,, \tag{2.6-2}$$

Speed of Light in a Medium

where ϵ and μ are the electric permittivity and magnetic permeability of the medium, respectively.

Light propagates in the form of electromagnetic waves that obey the wave equation and travel at the speed of light c .

Maxwell's Equations in a Dielectric Medium

- The wave equation (2.6-1) follows from Maxwell's equations. In a linear, nondispersive, homogeneous, isotropic, and source-free dielectric medium, the electric field $\mathcal{E}(\mathbf{r},t)$ and the magnetic field $\mathcal{H}(\mathbf{r},t)$ obey a set of coupled partial differential equations known as **Maxwell's equations**:

$$\nabla \times \mathcal{H} = \epsilon\frac{\partial \mathcal{E}}{\partial t} \tag{2.6-3}$$

$$\nabla \times \mathcal{E} = -\mu\frac{\partial \mathcal{H}}{\partial t} \tag{2.6-4}$$

$$\nabla \cdot \mathcal{E} = 0 \tag{2.6-5}$$

$$\nabla \cdot \mathcal{H} = 0\,. \tag{2.6-6}$$

Maxwell's Equations in a Medium

The vector operators $\nabla\cdot$ and $\nabla\times$ are the divergence and curl, respectively. In Cartesian coordinates, $\nabla\cdot\boldsymbol{\mathcal{E}} = \partial\mathcal{E}_x/\partial x + \partial\mathcal{E}_y/\partial y + \partial\mathcal{E}_z/\partial z$ is a scalar while $\nabla\times\boldsymbol{\mathcal{E}}$ is a vector with components $(\partial\mathcal{E}_z/\partial y - \partial\mathcal{E}_y/\partial z)$, $(\partial\mathcal{E}_x/\partial z - \partial\mathcal{E}_z/\partial x)$, and $(\partial\mathcal{E}_y/\partial x - \partial\mathcal{E}_x/\partial y)$. Maxwell's original formulation in 1865 comprised 20 simultaneous equations with 20 variables, which were condensed into their present form by Oliver Heaviside in 1885.

The wave equation (2.6-1) is readily derived from Maxwell's equations by applying the curl operation $\nabla\times$ to (2.6-4), employing the vector identity $\nabla\times(\nabla\times\boldsymbol{\mathcal{E}}) = \nabla(\nabla\cdot\boldsymbol{\mathcal{E}}) - \nabla^2\boldsymbol{\mathcal{E}}$, and then using (2.6-3) and (2.6-5) to demonstrate that each component of $\boldsymbol{\mathcal{E}}$ satisfies the wave equation. A similar procedure is followed for $\boldsymbol{\mathcal{H}}$. A necessary condition required for $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{H}}$ to satisfy Maxwell's equations is that each of the six interrelated components, $(\mathcal{E}_x, \mathcal{E}_y, \mathcal{E}_z)$ and $(\mathcal{H}_x, \mathcal{H}_y, \mathcal{H}_z)$, satisfy the wave equation.

Speed of Light

In free space, the **electric permittivity** is $\epsilon = \epsilon_0 \approx 8.8542\times10^{-12}$ F/m and the **magnetic permeability** is $\mu = \mu_0 = 1.2566\times10^{-6}$ H/m. In accordance with (2.6-2), the speed of light in free space (or air) is thus $c_0 = 1/\sqrt{\epsilon_0\mu_0} \approx 3.0\times10^8$ m/s = 30 cm/ns = 0.3 mm/ps = 0.3 μm/fs = 0.3 nm/as. The **refractive index** n, defined as the ratio of the speed of light in free space to that in a medium in (2.6-2), is therefore described by

$$n = \frac{c_0}{c} = \sqrt{\frac{\epsilon}{\epsilon_0}\frac{\mu}{\mu_0}}\,. \qquad \text{(2.6-7)}$$

Refractive Index

For nonmagnetic media, $\mu = \mu_0$, whereupon

$$n = \sqrt{\epsilon/\epsilon_0}\,, \qquad \text{(2.6-8)}$$

in which case the refractive index is simply the square root of the relative permittivity.

Superposition

Because Maxwell's equations and the wave equation are linear, the **principle of superposition** applies: if two separate sets of electric and magnetic fields are solutions to these equations, their sum is also a solution.

Boundary Conditions

At the boundary between two dielectric media, the tangential components of the electric field $\boldsymbol{\mathcal{E}}$ and of the magnetic field $\boldsymbol{\mathcal{H}}$ must be continuous, as suggested in Sec. 2.5.

Intensity, Power, and Energy

The flow of electromagnetic **power** is governed by the vector

$$\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{E}}\times\boldsymbol{\mathcal{H}}, \qquad \text{(2.6-9)}$$

which is known as the **Poynting vector**. The direction of power flow is along the direction of the Poynting vector, i.e., orthogonal to both $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{H}}$. The electromagnetic **intensity** $I(\mathbf{r}, t)$ (power flow across a unit area normal to the vector $\boldsymbol{\mathcal{S}}$) is the magnitude of the Poynting vector $\langle\boldsymbol{\mathcal{S}}\rangle$ averaged over a time inverval long in comparison with an optical cycle, but short in comparison with other times of interest. The wave-optics equivalent is provided in (2.1-3). The Poynting theorem, which is based on Maxwell's equations (2.6-3) and (2.6-4), takes the form of a continuity equation, $\nabla\cdot\boldsymbol{\mathcal{S}} = -\partial\mathcal{W}/\partial t$, and the **energy density** $\mathcal{W}$ stored in the medium can be expressed as

$$\mathcal{W} = \tfrac{1}{2}\epsilon\,\boldsymbol{\mathcal{E}}^2 + \tfrac{1}{2}\mu\boldsymbol{\mathcal{H}}^2\,. \qquad \text{(2.6-10)}$$

Domain of Electromagnetic Phenomena

The reach of electromagnetic theory, displayed in Fig. 2.6-1, stretches from VLF (very low frequency) waves to γ-rays. Optical frequencies occupy a band of the electromag-

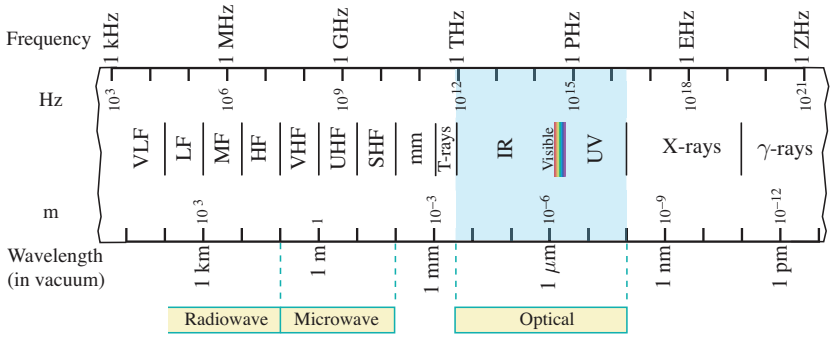


Figure 2.6-1 The reach of electromagnetic theory extends from VLF (very low frequencies and very long wavelengths) to gamma rays (very high frequencies and very short wavelengths). The optical region (shaded) is displayed in detail in Fig. 2.4-1.

netic spectrum that extends from the infrared (IR) through the visible to the ultraviolet (UV). The range of wavelengths generally considered to lie in the optical region thus extends from 300 μm to 10 nm, as is shown in greater detail in Fig. 2.4-1. Because these wavelengths are substantially shorter than those of microwaves or radiowaves, the techniques involved in their generation, transmission, and detection have traditionally had their own unique character. However, the march toward miniaturization in recent decades has blurred such differences, and it is now commonplace to encounter wavelength- and subwavelength-size cavities, antennas, waveguides, and other structures that resemble their longer wavelength counterparts.

Monochromatic Electromagnetic Waves

For the special case of monochromatic electromagnetic waves in an optical medium, the amplitude and phase generally depend on position, but all electric- and magnetic-field components are harmonic functions of time with a common frequency ν and a corresponding angular frequency $\omega = 2\pi\nu$, at all positions. Adopting the complex representation used in Sec. 2.2, the six real field components may be expressed as

$$\mathcal{E}(\mathbf{r},t) = \text{Re}\{\mathbf{E}(\mathbf{r})\,\exp(j\omega t)\} \quad \text{and} \quad \mathcal{H}(\mathbf{r},t) = \text{Re}\{\mathbf{H}(\mathbf{r})\exp(j\omega t)\}, \qquad (2.6\text{-}11)$$

where $\mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r})$ represent electric- and magnetic-field complex-amplitude vectors, respectively.

Inserting (2.6-11) into Maxwell's equations (2.6-3)–(2.6-6) for a linear, nondispersive, homogeneous, and isotropic medium, and noting that $(\partial/\partial t)\,e^{j\omega t} = j\omega\,e^{j\omega t}$ for monochromatic waves of angular frequency ω, we arrive at a set of equations obeyed by the field complex-amplitude vectors:

$$\nabla \times \mathbf{H} = j\omega\epsilon\mathbf{E} \qquad (2.6\text{-}12)$$

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H} \qquad (2.6\text{-}13)$$

$$\nabla \cdot \mathbf{E} = 0 \qquad (2.6\text{-}14)$$

$$\nabla \cdot \mathbf{H} = 0\,. \qquad (2.6\text{-}15)$$

Maxwell's Equations in a Medium (Monochromatic Light)

Also, substituting the electric and magnetic fields $\mathcal{E}$ and $\mathcal{H}$ given in (2.6-11) into the wave equation (2.6-1) yields the Helmholtz equation

$$\boxed{\nabla^2 U + k^2 U = 0,} \quad k = nk_0 = \omega\sqrt{\epsilon\mu}\,, \tag{2.6-16}$$

Helmholtz Equation

where the scalar function $U = U(\mathbf{r})$ represents the complex amplitude of any of the three components (E_x, E_y, E_z) of $\mathbf{E}$ or the three components (H_x, H_y, H_z) of $\mathbf{H}$; and where $n = \sqrt{(\epsilon/\epsilon_0)(\mu/\mu_0)}$, $k_0 = \omega/c_0$, and $c = c_0/n$. The Helmholtz equation for scalar waves, which was cast in terms of the complex amplitude $U(\mathbf{r})$ of the real wavefunction $u(\mathbf{r}, t)$ as provided in (2.2-7), is identical in form to (2.6-16).

Intensity and Power. As indicated in the discussion surrounding (2.6-9), the flow of electromagnetic power is governed by the time average of the Poynting vector $\mathcal{S} = \mathcal{E} \times \mathcal{H}$. Casting this expression in terms of complex amplitudes for monochromatic waves yields

$$\begin{aligned}\mathcal{S} &= \text{Re}\{\mathbf{E}e^{j\omega t}\} \times \text{Re}\{\mathbf{H}e^{j\omega t}\} = \tfrac{1}{2}\left(\mathbf{E}e^{j\omega t} + \mathbf{E}^* e^{-j\omega t}\right) \times \tfrac{1}{2}\left(\mathbf{H}e^{j\omega t} + \mathbf{H}^* e^{-j\omega t}\right) \\ &= \tfrac{1}{4}\left(\mathbf{E} \times \mathbf{H}^* + \mathbf{E}^* \times \mathbf{H} + e^{j2\omega t}\,\mathbf{E} \times \mathbf{H} + e^{-j2\omega t}\,\mathbf{E}^* \times \mathbf{H}^*\right). \end{aligned} \tag{2.6-17}$$

The terms containing the factors $e^{j2\omega t}$ and $e^{-j2\omega t}$ oscillate at optical frequencies and are therefore washed out by the averaging process, which is slow in comparison with an optical cycle. We therefore arrive at

$$\langle \mathcal{S} \rangle = \tfrac{1}{4}(\mathbf{E} \times \mathbf{H}^* + \mathbf{E}^* \times \mathbf{H}) = \tfrac{1}{2}(\mathbf{S} + \mathbf{S}^*) = \text{Re}\{\mathbf{S}\}, \tag{2.6-18}$$

where the vector

$$\mathbf{S} = \tfrac{1}{2}\mathbf{E} \times \mathbf{H}^* \tag{2.6-19}$$

may be regarded as a complex Poynting vector. The optical intensity is the magnitude of the vector $\text{Re}\{\mathbf{S}\}$ per unit area normal to the vector $\mathbf{S}$.

Inhomogeneous Media. In an inhomogeneous, nonmagnetic medium, Maxwell's equations (2.6-12)–(2.6-15) remain applicable, but the electric permittivity of the medium becomes position dependent, i.e., $\epsilon = \epsilon(\mathbf{r})$. For locally homogeneous media in which $\epsilon(\mathbf{r})$ varies slowly with respect to the wavelength, the Helmholtz equation provided in (2.6-16) remains approximately valid, subject to the substitutions $k = n(\mathbf{r})k_0$ and $n(\mathbf{r}) = \sqrt{\epsilon(\mathbf{r})/\epsilon_0}$.

Elementary Electromagnetic Waves

We now examine the spatial features of plane and dipole electromagnetic waves, which are analogous to the plane and spherical scalar waves considered in Sec. 2.3. Again, we restrict our consideration to linear, nondispersive, homogeneous, isotropic, and source-free media.

Electromagnetic Plane Wave. The transverse electromagnetic (TEM) wave is characterized by the plane-wave magnetic- and electric-field complex-amplitude vectors

$$\mathbf{H}(\mathbf{r}) = \mathbf{H}_0 \exp(-j\mathbf{k} \cdot \mathbf{r}) \quad \text{and} \quad \mathbf{E}(\mathbf{r}) = \mathbf{E}_0 \exp(-j\mathbf{k} \cdot \mathbf{r})\,, \tag{2.6-20}$$

respectively, where the complex envelopes $\mathbf{H}_0$ and $\mathbf{E}_0$ are constant vectors, and $\mathbf{k}$ is the wavevector. All six components of $\mathbf{H}(\mathbf{r})$ and $\mathbf{E}(\mathbf{r})$ satisfy the Helmholtz equation

(2.6-16) provided that the magnitude of $\mathbf{k}$ is $k = nk_0$, where n is the refractive index of the medium. To satisfy Maxwell's equations (2.6-12)–(2.6-15), it can be shown that $\mathbf{E}$, $\mathbf{H}$, and $\mathbf{k}$ must form a mutually orthogonal trio, as portrayed in Fig. 2.6-2. Since $\mathbf{E}$ and $\mathbf{H}$ lie in a plane normal to the direction of propagation $\mathbf{k}$, the wave is called a **transverse electromagnetic (TEM)** wave.

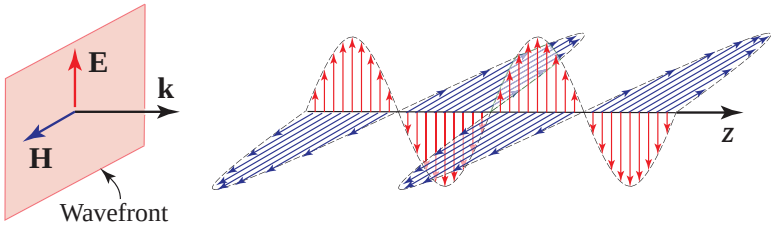


Figure 2.6-2 The transverse electromagnetic (TEM) plane wave. The vectors $\mathbf{E}$, $\mathbf{H}$, and $\mathbf{k}$ are mutually orthogonal. The wavefronts (surfaces of constant phase) are normal to the wavevector $\mathbf{k}$.

The complex Poynting vector $\mathbf{S} = \frac{1}{2}\mathbf{E} \times \mathbf{H}^*$ specified in (2.6-19) is parallel to the wavevector $\mathbf{k}$, so that the optical power flows along a direction normal to the wavefronts. The optical intensity I of the wave is given by

$$I = |E_0|^2/2\eta\,, \qquad \text{(2.6-21)}$$

Optical Intensity

where the **impedance** η of the medium is

$$\eta = E_0/H_0 = \sqrt{\mu/\epsilon}\,. \qquad \text{(2.6-22)}$$

The impedance of free space is $\eta_0 = \sqrt{\mu_0/\epsilon_0} \approx 377\ \Omega$. The intensity of a monochromatic TEM wave is proportional to the absolute square of the complex envelope of the electric field, as provided in (2.6-21). Still, the intensity of a monochromatic scalar wave behaves as $I = |U|^2$, as specified in (2.2-10), so it exhibits analogous behavior. A paraxial electromagnetic wave can be approximated by a TEM plane wave.

Electromagnetic Dipole Wave. The wave generated by an oscillating electric dipole has features that resemble the scalar spherical wave discussed in Sec. 2.3. The dipole wave is constructed using a spherical coordinate system; the details can be found in a textbook on electromagnetic optics. The complex-amplitude vectors and spherical wavefront associated with this wave are illustrated in Fig. 2.6-3.

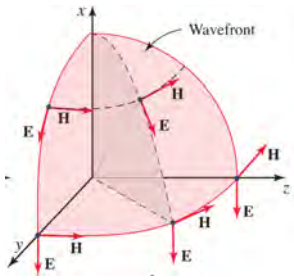


Figure 2.6-3 The electromagnetic wave radiated by an oscillating dipole. At distances from the origin large in comparison with a wavelength ($r \gg \lambda/2\pi$), the wavefronts are spherical. The electric- and magnetic-field vectors are orthogonal to each another and to the radial direction $\hat{\mathbf{r}}$. The electric field points in the polar direction and the magnetic field points in the azimuthal direction.

In analogy with the scalar spherical wave illustrated in Fig. 2.3-2, at distances far from the origin ($r \gg \lambda/2\pi$ or $kr = 2\pi r/\lambda \gg 1$), the wavefronts are spherical and the wave can be approximated by a paraboloidal wave (and ultimately by a TEM plane), as sketched in Fig. 2.3-3.

Electromagnetic Waves in Optical Fibers

An electromagnetic-wave treatment of the propagation of light in optical fibers augments that offered by ray optics, which is based solely on total internal reflection (Sec. 1.6). The use of Maxwell's equations, together with the boundary conditions imposed by the cylindrical dielectric core and cladding, enable the electric and magnetic fields of guided waves to be determined.

Optical fibers are classified as step-index or graded-index (GRIN), and as multimode (MMF) or single-mode (SMF), as illustrated in Fig. 2.6-4. Step-index fibers, which are the most common, have a constant refractive index in the core and a slightly lower constant refractive index in the cladding. GRIN fibers, where the refractive index of the fiber core is graded from a maximum value at its center to a minimum value at the core–cladding boundary, are used in some specialized applications. An optical fiber with a large core diameter is labeled multimode because it can support multiple optical modes, whereas an optical fiber with a sufficiently small core diameter supports only a single mode. Multimode and single-mode fibers exhibit distinct propagation constants, characteristic transverse field distributions, and pairs of independent polarization states.

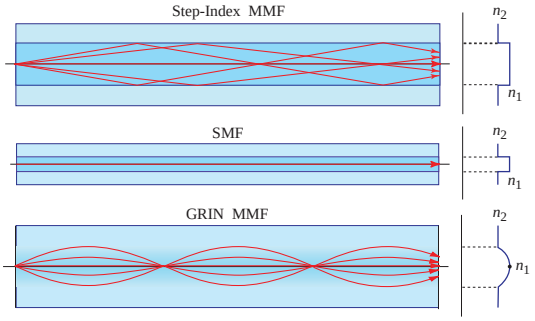


Figure 2.6-4 Geometry, refractive-index profiles, and typical ray traces for a step-index multimode fiber (MMF), a single-mode fiber (SMF), and a graded-index (GRIN) multimode fiber.

Power Reflectance at the Boundary Between Dielectric Media

The amplitude reflectance and transmittance of a monochromatic plane wave at the boundary between two lossless dielectric media of different refractive indices, determined by the **Fresnel equations**, depends strongly on the angle of incidence and the polarization of the incident wave. Although the Fresnel equations are complex, the **power (or intensity) reflectance** $\mathcal{R}$ assumes a simple form, applicable for both external and internal reflection, for a normally incident TEM wave:

$$\mathcal{R} = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2. \tag{2.6-23}$$

At oblique angles of incidence, $\mathcal{R}$ can be far greater or far smaller than the value dictated by (2.6-23).

Under certain circumstances, the power reflectance also exhibits special behavior when the wave is not normally incident on the boundary. One example is total internal reflection, which occurs when the angle of incidence exceeds the critical angle θ_c, as discussed in Sec. 1.3 in the context of ray optics. A diametrically opposite example is the total transmission of optical power, without reflection, of a transverse-magnetic (TM) polarized wave when the angle of incidence equals the **Brewster angle** θ_B.

EXAMPLE 2.6-1. ***Power Reflectance at the Boundary Between Air and Glass.***

- At the boundary between air ($n = 1$) and glass ($n = 1.5$) at normal incidence, (2.6-23) yields $\mathcal{R} = 0.04$, so 4% of the incident optical power is reflected. Since the **power (or intensity) transmittance** $\mathcal{T} = 1 - \mathcal{R}$ for lossless media, 96% of the power is transmitted.

- The power reflectance at normal incidence from a transparent, lossless plate with two flat surfaces is given by $\mathcal{R}(1+\mathcal{T}^2)$ since the power reflected from the far surface involves a double passage of the light through the plate. For glass, $\mathcal{R}(1+\mathcal{T}^2) = 0.04[1+(0.96)^2] \approx 0.077$, indicating that $\approx 7.7\%$ of the incident light power is reflected.
- At the boundary between air ($n = 1$) and GaAs ($n = 3.6$) at normal incidence, $\mathcal{R} \approx 0.32$ so that 32% of the light is reflected from a single surface.

Nonlinear, Dispersive, Inhomogeneous, Anisotropic, Conductive Media

Analogous, but more complex, versions of Maxwell's equations (2.6-3)–(2.6-6) and the wave equation (2.6-1) are available when one or more of the properties of linearity, nondispersiveness, homogeneity, and isotropy are not satisfied, or when the medium is not source-free. For dielectric media, this is achieved by incorporating two auxiliary vector fields into Maxwell's equations, the electric flux density $\mathcal{D}(\mathbf{r},t)$ (which in turn depends on $\mathcal{E}$ and the polarization density of the medium $\mathcal{P}$) and the magnetic flux density $\mathcal{B}(\mathbf{r},t)$ (which in turn depends on $\mathcal{H}$ and the magnetization density of the medium $\mathcal{M}$). For conductive media, such as metals or semiconductors, the current-density vector $\mathcal{J}$ must be added to the mix.

Relation of Scalar and Electromagnetic Waves

Scalar wave optics has the following connections with electromagnetic optics:

- Scalar wave optics forms a suitable approximation to electromagnetic optics when the vector nature of electromagnetic waves is not of importance in the problem under consideration.
- The wave equation (2.1-1) at the heart of wave optics is a scalar version of the wave equation of electromagnetic optics (2.6-1), which follows from Maxwell's equations.
- The speed of light postulated in wave optics is established by the medium's electric permittivity ϵ and magnetic permeability μ in electromagnetic optics, as is evident in (2.6-2).
- The scalar wavefunction $u(\mathbf{r},t)$ set forth in Sec. 2.1 represents the six components of the electric- and magnetic-field vectors of electromagnetic optics.
- The Helmholtz equation (2.2-7) of wave optics is a single-component version of the Helmholtz equation (2.6-16) of electromagnetic optics.
- The intensity of a paraxial scalar wave is proportional to the absolute square of the complex wavefunction; the intensity of a paraxial electromagnetic wave, approximated by a TEM plane wave, is proportional to the absolute square of the complex electric-field envelope.
- Scalar wave optics and electromagnetic optics both accommodate phenomena that involve the phase of the wavefunction, such as diffraction and interference.

The following topics are accommodated by electromagnetic optics but are largely inaccessible to scalar wave optics:

- The behavior of light in media where polarization is a central feature, such as anisotropic, optically active, magneto-optic, and liquid-crystal media, as well as photonic crystals and metamaterials.
- The behavior of light in devices whose operation relies on a quantitative reckoning of the proportion of light reflected and refracted at boundaries, such as optical waveguides, optical fibers, and optical resonators.
- The behavior of light in nonlinear, dispersive, scattering, and/or conductive media.

2.7 RANDOM WAVES

In the earlier sections of this Chapter, the light was considered to be deterministic. The monochromatic wavefunction $u(\mathbf{r},t) = \text{Re}\{U(\mathbf{r})\exp(j2\pi\nu t)\}$ set forth in (2.2-1), for example, assumed a time dependence that was perfectly periodic and predictable, as portrayed in Fig. 2.7-1(a). The amplitude $U(\mathbf{r})$ was also taken to be a deterministic complex function of position. This was also the case for all electric- and magnetic-field components associated with monochromatic electromagnetic waves, as expressed in (2.6-11).

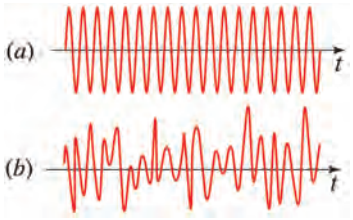


Figure 2.7-1 Sketch of the time dependence of the wavefunction $u(\mathbf{r},t)$ for: (a) a deterministic, coherent, monochromatic wave; (b) a random wave.

For random light, in contrast, the dependence of the wavefunction $u(\mathbf{r},t)$ on time (and position) is not totally predictable, as illustrated in Fig. 2.7-1(b). Random optical waves are a result of the statistical fluctuations inherent in many sources of light. Such fluctuations can arise when the emissions comprise superpositions of large collections of independent radiators with different frequencies and phases. Natural light, such as that radiated by the sun and stars for example, varies randomly in time, as does the thermal light radiated by a hot object. So too does the light generated in the junction region of a light-emitting diode, as a consequence of the recombination of large numbers of electron–hole pairs at random times. Although the discussion in this section is directed toward temporal randomness, it should be mentioned that random variations in the optical wavefront arise when deterministic light is passed through a spatially random medium, such as a ground-glass diffuser or a turbulent fluid, or when it is scattered by a rough surface.

Random waves are characterized by making use of statistical averaging to define various (nonrandom) measures associated with the waves. Because a random wavefunction $u(\mathbf{r},$ satisfies certain laws, such as the wave equation and associated boundary conditions, so too do its statistical averages. In this section, we examine various properties of random light, including its optical intensity, temporal coherence function, degree of temporal coherence, coherence time, power spectral density, and spectral width. Scalar wave theory suffices for describing these properties, although electromagnetic theory is required to address the polarization properties of random light. Even so, random light is effectively coherent when its coherence time is much larger than any time-delay differences encountered in the optical system, as will be discussed in the sequel.

The area of optics concerned with the study of random light, including the salient statistical averages and the laws that govern them, along with the measures that determine whether the light is classified as coherent, incoherent, or partially coherent, is called **optical coherence theory** or **statistical optics**.

Optical Intensity

An arbitrary optical wave is described by a wavefunction $u(\mathbf{r},t) = \text{Re}\{U(\mathbf{r},t)\}$, where $U(\mathbf{r},t)$ is the complex wavefunction. For example, $U(\mathbf{r},t)$ may take the form $U(\mathbf{r})\exp(j2\pi\nu t)$ for monochromatic light, or it may comprise a sum of such functions with many different values of ν for polychromatic light. As discussed in Sec. 2.2, the

intensity $I(\mathbf{r}, t)$ of a coherent (deterministic) wave is related to the absolute square of the complex wavefunction $U(\mathbf{r}, t)$ via

$$I(\mathbf{r}, t) = |U(\mathbf{r}, t)|^2 . \tag{2.7-1}$$

For monochromatic deterministic light, the intensity is independent of time, whereas for pulsed deterministic light, it is time varying.

For random light, on the other hand, the functions $u(\mathbf{r}, t)$ and $U(\mathbf{r}, t)$, as well as the intensity $|U(\mathbf{r}, t)|^2$, are random functions of time and position, which causes us to rely on statistical averaging. The **average intensity** is defined as

$$\boxed{I(\mathbf{r}, t) = \langle |U(\mathbf{r}, t)|^2 \rangle,} \tag{2.7-2}$$

Average Intensity
(Ensemble Average)

where the symbol $\langle \cdot \rangle$ denotes an ensemble average over many realizations of the random function within the brackets. Despite the fact that a random wave repeatedly generated under identical conditions yields a different wavefunction on each trial, the average intensity at each time and position, established by (2.7-2), is deterministic. We call $I(\mathbf{r}, t)$ the intensity of the light (with the modifier "average" implied), when there is no ambiguity in meaning. The unaveraged quantity $|U(\mathbf{r}, t)|^2$, in contrast, is called the **random intensity** or **instantaneous intensity**. For deterministic light, the averaging operation is superfluous since all trials produce exactly the same wavefunction, in which case (2.7-2) is equivalent to (2.7-1).

The average intensity may be time independent or it may be a function of time, as illustrated in Figs. 2.7-2(a) and (b), respectively. The former situation is operative when the optical wave is statistically **stationary**, i.e., when its statistical averages are invariant to time. Clearly, stationarity does not necessarily imply constancy; rather, it implies constancy only of the average properties. An example of stationary random light is that emitted by an incandescent lamp whose filament is heated by a constant electric current. The average intensity $I(\mathbf{r})$ is then a function of distance from the lamp, but it does not vary with time. On the other hand, the random intensity $|U(\mathbf{r}, t)|^2$ fluctuates with both position and time, as illustrated in the figure.

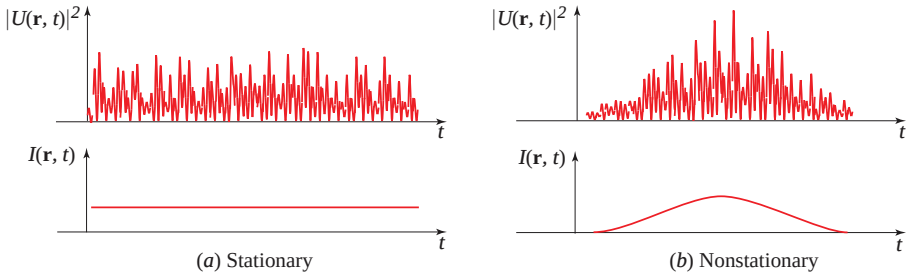


Figure 2.7-2 (a) A statistically stationary wave has an average intensity $I(\mathbf{r})$ that does not vary with time. (b) A statistically nonstationary wave has an average intensity $I(\mathbf{r}, t)$ that varies with time. These plots represent, for example, the intensity of light produced by an incandescent lamp driven by (a) a constant electric current, and (b) a pulse of electric current.

When the light is stationary, the ensemble average over many realizations of the instantaneous intensity, as prescribed by (2.7-2), is usually equivalent to the time average over

a long duration, which is expressed as

$$I(\mathbf{r}) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |U(\mathbf{r},t)|^2 \, dt. \tag{2.7-3}$$

Average Intensity (Time Average)

Random processes for which the time average tends to the ensemble average are said to be **ergodic processes**.

Temporal Coherence Function

We now proceed to further explore the fluctuations of stationary light at a fixed position $\mathbf{r}$, as a function of time. Since $\mathbf{r}$ is fixed, for brevity we refer to the stationary random wavefunction as $U(t) \equiv U(\mathbf{r},t)$ and to the constant intensity as $I(\mathbf{r}) \equiv \langle |U(\mathbf{r},t)|^2 \rangle$. The random fluctuations of $U(t)$ are characterized by a time scale that represents the "memory" of the random function. For times separated by an interval longer than this memory time, the process "forgets" itself and the fluctuations are independent. Within the memory time, the wavefunction appears to be relatively smooth but when viewed over longer time scales, it appears "erratic."

This temporal behavior is captured by a statistical average known as the autocorrelation function. This quantitative measure describes the extent to which the wavefunction fluctuates in unison at two instants of time separated by a given time delay, and thus serves to establish the time scale of the process that characterizes the wavefunction. The **autocorrelation function** of a stationary complex random function $U(t)$ is defined as the ensemble average of the product of $U^*(t)$ and $U(t+\tau)$, as a function of the time delay τ,

$$G(\tau) = \langle U^*(t)\, U(t+\tau) \rangle. \tag{2.7-4}$$

Temporal Coherence Function

When expressed as a time average, the autocorrelation function is written as

$$G(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} U^*(t)\, U(t+\tau) \, dt. \tag{2.7-5}$$

In the language of optical coherence theory, the autocorrelation function $G(\tau)$ is known as the **temporal coherence function**.

To understand the significance of the definition presented in (2.7-4), consider the case in which the average value of the complex wavefunction $\langle U(t) \rangle = 0$. This arises when the phase of the phasor $U(t)$ is equally likely to have any value between 0 and 2π, as illustrated in Fig. 2.7-3.

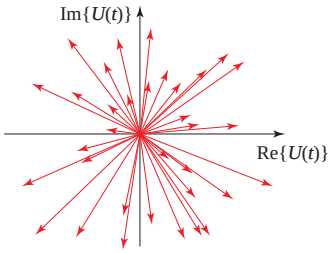


Figure 2.7-3 Variation of the phasor $U(t)$ with time when its argument is uniformly distributed between 0 and 2π. The average values of its real and imaginary parts are zero, so that $\langle U(t) \rangle = 0$.

The phase of the product $U^*(t)U(t+\tau)$ is the angle between the phasors $U(t)$ and $U(t+\tau)$ so when these two quantities are uncorrelated, the angle between their phasors varies randomly between 0 and 2π. The phasor $U^*(t)U(t+\tau)$ then has an angle that is totally uncertain and equally likely to take any direction, so that its average, the autocorrelation function $G(\tau)$, vanishes. On the other hand, if $U(t)$ and $U(t+\tau)$ are correlated for a given value of τ, their phasors maintain some relationship and their fluctuations are linked together, so that the product phasor $U^*(t)U(t+\tau)$ has a preferred direction and its average $G(\tau)$ does not vanish. It is readily shown that $G(\tau)$ is a function with Hermitian symmetry, $G(-\tau) = G^*(\tau)$, and that the intensity I , defined by (2.7-2), is given by

$$I = G(0). \tag{2.7-6}$$

Complex Degree of Temporal Coherence

The temporal coherence function $G(\tau)$ carries information about both the intensity $I = G(0)$ and the degree of correlation (coherence) of stationary light. A measure of coherence that is independent of the intensity is provided by the normalized autocorrelation function,

$$\boxed{g(\tau) = \frac{G(\tau)}{G(0)} = \frac{\langle U^*(t)U(t+\tau)\rangle}{\langle U^*(t)U(t)\rangle},} \tag{2.7-7}$$

Complex Degree of Temporal Coherence

which is called the **complex degree of temporal coherence**. Its absolute value cannot exceed unity,

$$0 \leqslant |g(\tau)| \leqslant 1. \tag{2.7-8}$$

The value of $|g(\tau)|$ is a measure of the degree of correlation between $U(t)$ and $U(t+\tau)$. When the light is monochromatic and deterministic, i.e., when $U(t) = \mathfrak{a}_0 \exp(j2\pi\nu_0 t)$ where $\mathfrak{a}_0$ is a constant, (2.7-7) yields

$$g(\tau) = \exp(j2\pi\nu_0\tau), \tag{2.7-9}$$

so that $|g(\tau)| = 1$ for all τ. The variables $U(t)$ and $U(t+\tau)$ are then totally correlated for all time delays τ. For most sources of light, $|g(\tau)|$ decreases from its maximum value $|g(0)| = 1$ as τ increases, and the fluctuations become uncorrelated when τ substantially exceeds the memory time of the process.

Coherence Time

If $|g(\tau)|$ decreases monotonically with time delay, the value τ_c at which it decreases to a prescribed value ($1/2$ or $1/e$, for example) serves as a measure of the memory time of the fluctuations. The quantity τ_c, called the **coherence time**, is illustrated in Fig. 2.7-4. For $\tau < \tau_c$ the fluctuations are "strongly" correlated whereas for $\tau > \tau_c$ they are "weakly" correlated. The quantity τ_c is the width of the function $|g(\tau)|$; although the width of a function can be defined in many ways, as discussed in Sec. A.2 of Appendix A, the power-equivalent width is most commonly used in conjunction with the definition of coherence time:

$$\boxed{\tau_c = \int_{-\infty}^{\infty} |g(\tau)|^2 \, d\tau} \tag{2.7-10}$$

Coherence Time

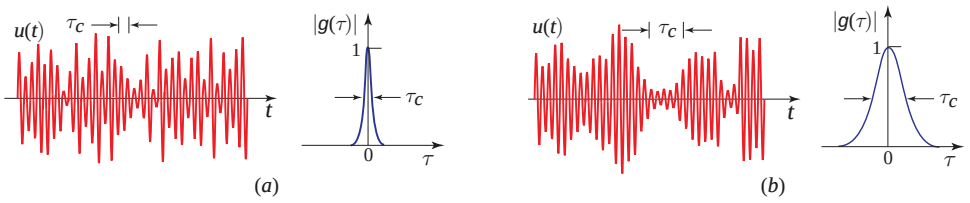


Figure 2.7-4 Illustrating the wavefunction $u(t)$, the magnitude of the complex degree of temporal coherence $|g(\tau)|$, and the coherence time τ_c for a random optical wave with: (a) short coherence time, and (b) long coherence time. The amplitudes and phases of the wavefunctions vary randomly, with time constants roughly established by τ_c (which is assumed to be greater than the duration of an optical cycle). For intervals shorter than the coherence time, the wave is reasonably predictable and can be approximated by a sinusoid. On the other hand, given the amplitude and phase of the wave at a particular time, these quantities cannot be predicted at time delays that stretch beyond a coherence time.

☐ **Consistency of Coherence-Time and Temporal-Coherence Definitions Provided in (2.7-10).**

(a) For a degree of temporal coherence that decreases exponentially, $g(\tau) \equiv \exp(-|\tau|/\tau_c)$:
$\tau_c \equiv \int_{-\infty}^{\infty} |g(\tau)|^2\, d\tau = \int_{-\infty}^{\infty} \exp(-2|\tau|/\tau_c)\, d\tau = 2\int_0^{\infty} \exp(-2\tau/\tau_c)\, d\tau = \tau_c$. ✓
Note that $|g(\tau)|$ decreases by a factor of $1/e = 0.368$ at $\tau = \tau_c$.

(b) For a degree of temporal coherence that decreases as a Gaussian, $g(\tau) \equiv \exp(-\pi\tau^2/2\tau_c^2)$:
$\tau_c \equiv \int_{-\infty}^{\infty} |g(\tau)|^2\, d\tau = \int_{-\infty}^{\infty} \exp(-\pi\tau^2/\tau_c^2)\, d\tau = \tau_c$. ✓
Note that $|g(\tau)|$ decreases by a factor of $\exp(-\pi/2) = 0.208$ at $\tau = \tau_c$. ■

The coherence time of monochromatic light source is infinite since $|g(\tau)| = 1$ everywhere. Practially speaking, however, light for which the coherence time τ_c is much greater than the differences of any time delays encountered in an optical system is **effectively coherent**. Equivalently, light is effectively coherent if its **coherence length** l_c is much greater than all optical pathlength differences encountered in the system:

$$\boxed{l_c = c\tau_c\,.} \qquad (2.7\text{-}11)$$
Coherence Length

Spectral Density

A determination of the average spectrum of random light is attained by carrying out a Fourier decomposition of the random function $U(t)$. As discussed in Sec. A.1 of Appendix A, the amplitude of the component of frequency ν is its Fourier transform,

$$V(\nu) = \int_{-\infty}^{\infty} U(t) \exp(-j2\pi\nu t)\, dt. \qquad (2.7\text{-}12)$$

The average energy per unit area of those components whose frequencies lie in the interval between ν and $\nu + d\nu$ is $\langle |V(\nu)|^2 \rangle\, d\nu$, so that $\langle |V(\nu)|^2 \rangle$ represents the average energy spectral density of the light (energy per unit area per unit frequency).

Since an exemplary stationary function $U(t)$ is eternal and carries infinite energy, we direct our attention instead the **power spectral density**. We begin by determining the truncated Fourier transform of the function $U(t)$ observed over a window of time duration T,

$$V_T(\nu) = \int_{-T/2}^{T/2} U(t) \exp(-j2\pi\nu t)\, dt, \tag{2.7-13}$$

which leads to the (truncated) energy spectral density $\langle |V_T(\nu)|^2 \rangle$. Since the power spectral density is the energy spectral density per unit time, in the limit $T \to \infty$ we have

$$S(\nu) = \lim_{T\to\infty} \frac{1}{T} \langle |V_T(\nu)|^2 \rangle. \tag{2.7-14}$$

However, because $U(t)$ is expressly defined in (2.7-2) such that $|U(t)|^2$ represents power per unit area (intensity), $S(\nu)\, d\nu$ represents the average power per unit area in a band of frequencies lying between ν and $\nu + d\nu$. Strictly speaking, therefore, $S(\nu)$ represents the intensity spectral density (W/m^2-Hz). This quantity is readily converted to the power spectral density (W/Hz) via multiplication by an effective area A_{eff}, and is often referred to simply as the **spectral density** or **spectrum**. Because the complex wavefunction $U(t)$ is defined such that $V(\nu) = 0$ for negative ν, $S(\nu)$ is nonzero only for positive frequencies so that the total average intensity is given by

$$I = \int_0^\infty S(\nu)\, d\nu. \tag{2.7-15}$$

The autocorrelation function $G(\tau)$ defined by (2.7-4), and the spectral density $S(\nu)$ defined by (2.7-13) and (2.7-14), are readily shown to form a Fourier transform pair,

$$S(\nu) = \int_{-\infty}^{\infty} G(\tau) \exp(-j2\pi\nu\tau)\, d\tau, \tag{2.7-16}$$

Spectral Density (Wiener–Khinchin Theorem)

a relationship known as the **Wiener–Khinchin theorem**.

The light entering the eye is usually characterized by a **wavelength-based power spectral density** (or **spectral radiant flux**) $S_\lambda(\lambda_0)$, rather than by its frequency-based counterpart $S_\nu(\nu)$. Wavelength-based power spectral densities are sketched in Fig. 2.7-5 for the light reflected from three locations on an Herbin abstract oil-on-canvas. Each of these spectral densities evokes a perceived color, as detailed in Sec. 9.6.

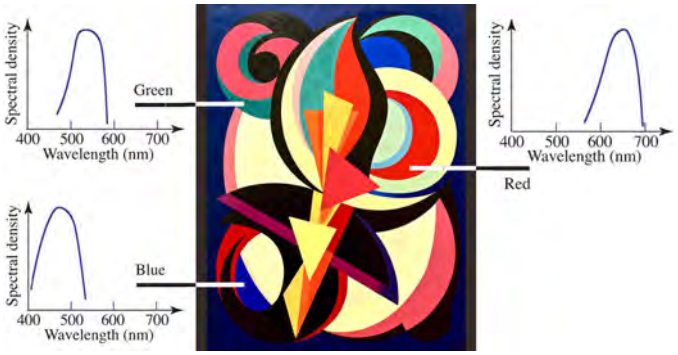


Figure 2.7-5 Wavelength-based power spectral densities $S_\lambda(\lambda_0)$ for the light reflected from three locations on an Herbin abstract painting, plotted as a function of the free-space wavelength λ_0. (Auguste Herbin (1882–1960), *Composition*, Oil-on-Canvas, 1939, Museu Coleção Berardo, Centro Cultural de Belém, Lisboa, Portugal, Pedro Ribeiro Simões via Wikimedia Commons.)

Spectral Width

The spectrum of light is often confined to a narrow band centered about a central frequency ν_0. The **spectral width**, or **linewidth**, of light is the width $\Delta\nu$ of the spectral density $S(\nu)$. Because of the Fourier-transform relation between $S(\nu)$ and $G(\tau)$, the widths of these two functions, $\Delta\nu$ and τ_c, respectively, are inversely related (Sec. A.2 of Appendix A). Hence, as illustrated in Fig. 2.7-6, a light source of broad spectral width has a short coherence time, whereas a light source of narrow spectral width has a long coherence time. In the limiting case of monochromatic light, $G(\tau) = I\exp(j2\pi\nu_0\tau)$, so that the corresponding spectral density $S(\nu) = I\delta(\nu - \nu_0)$ contains only a single frequency component ν_0, in which case $\tau_c = \infty$ and $\Delta\nu = 0$. Although the coherence time of a source of light can be increased by passing the light through a narrowband optical filter to reduce its spectral width, the resultant gain in coherence comes at the expense of a reduction in optical intensity.

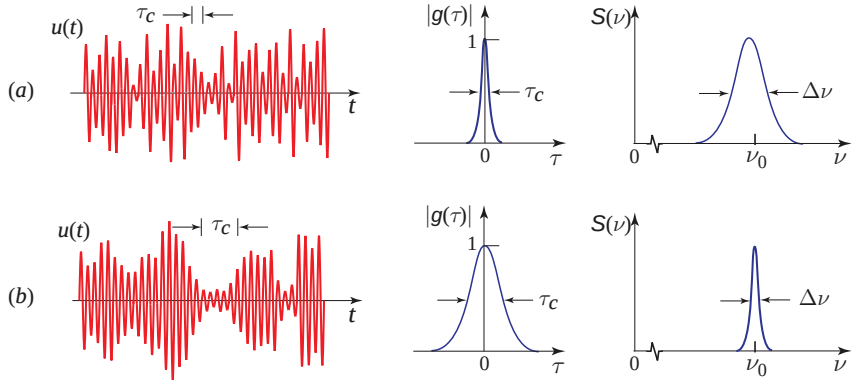


Figure 2.7-6 Examples of two random waves, along with the magnitudes of their complex degrees of temporal coherence $|g(\tau)|$ and spectral densities $S(\nu)$: (a) Narrow complex degree of temporal coherence (short coherence time τ_c) and broad spectral width $\Delta\nu$; (b) Broad complex degree of temporal coherence (long coherence time τ_c) and narrow spectral width $\Delta\nu$. The widths of $|g(\tau)|$ and $S(\nu)$, designated τ_c and $\Delta\nu$, respectively, are inversely related.

A commonly used definition for the spectral width of the function $S(\nu)$ is its full-width at half-maximum (FWHM), denoted $\Delta\nu_{\text{FWHM}} \equiv \Delta\nu$. The relation between $\Delta\nu_{\text{FWHM}}$ and the coherence time τ_c depends on the spectral profile of the source, as displayed in Table 2.7-1.

Table 2.7-1 Relation between spectral width $\Delta\nu_{\text{FWHM}}$ and coherence time τ_c for light with rectangular, Lorentzian, and Gaussian spectral profiles.

SPECTRAL PROFILE :	Rectangular	Lorentzian	Gaussian
Spectral Width $\Delta\nu_{\text{FWHM}}$:	$\dfrac{1}{\tau_c}$	$\dfrac{1}{\pi\tau_c} \approx \dfrac{0.32}{\tau_c}$	$\dfrac{\sqrt{2\ln 2/\pi}}{\tau_c} \approx \dfrac{0.66}{\tau_c}$

It turns out, however, that there is distinct merit in making use of an alternative definition of the spectral width, namely

$$\Delta\nu_c = \frac{\left(\int_0^\infty S(\nu)\,d\nu\right)^2}{\int_0^\infty S^2(\nu)\,d\nu}, \tag{2.7-17}$$

since the spectral width is then the exact inverse of the coherence time, whatever the spectral profile of the light (the derivation is provided below):

$$\boxed{\Delta\nu_c = \frac{1}{\tau_c}.} \qquad (2.7\text{-}18)$$

Spectral Width

□ **Derivation of Relation Between Spectral Width and Coherence Time Provided in (2.7-18).**

Using (2.7-6) and (2.7-15) provides $\int_0^\infty S(\nu)d\nu = G(0)$. Squaring both sides yields

$$\left[\int_0^\infty S(\nu)d\nu)\right]^2 = [G(0)]^2. \qquad (2.7\text{-}19)$$

Since $S(\nu)$ and $G(\tau)$ form a Fourier-transform pair in accordance with (2.7-16), Parseval's theorem may be written as

$$\int_0^\infty S^2(\nu)d\nu = \int_{-\infty}^\infty |G(\tau)|^2 d\tau. \qquad (2.7\text{-}20)$$

Dividing (2.7-19) by (2.7-20), and making use of the definitions for the magnitude of the complex degree of temporal coherence $|g(\tau)|$ provided in (2.7-7), the coherence time τ_c presented in (2.7-10), and the spectral width $\Delta\nu_c$ provided in (2.7-17), leads to

$$\Delta\nu_c = |G(0)|^2 / \int_{-\infty}^\infty |G(\tau)|^2 d\tau = 1/\int_{-\infty}^\infty |g(\tau)|^2 d\tau = 1/\tau_c. \ \surd \qquad (2.7\text{-}21)$$

■

As a particular example, if $S(\nu)$ is a rectangular function extending over the frequency interval from $\nu_0 - B/2$ to $\nu_0 + B/2$, then (2.7-17) yields $\Delta\nu_c = B$. For this particular profile, the coherence time $\tau_c = 1/B$, so that (2.7-18) is obeyed. For the spectral profiles displayed in Table 2.7-1, the two definitions of bandwidth, $\Delta\nu_c$ and $\Delta\nu_{\text{FWHM}}$, differ by a factor that ranges from 0.32 (Lorentzian) to 1 (rectangular).

Representative values of the spectral width $\Delta\nu_c$ for several sources of light, along with their associated coherence times τ_c and coherence lengths $l_c = c_0\tau_c$, are provided in Table 2.7:

SOURCE	$\Delta\nu_c$ (Hz)	$\tau_c = 1/\Delta\nu_c$	$l_c = c_0\tau_c$
Filtered sunlight ($\lambda_0 = 0.4$–0.8 μm)	3.7×10^{14}	2.7 fs	800 nm
Light-emitting diode ($\lambda_0 = 1$ μm, $\Delta\lambda_0 = 50$ nm)	1.5×10^{13}	67 fs	20 μm
Multimode He–Ne laser ($\lambda_0 = 633$ nm)	1.5×10^9	0.7 ns	20 cm
Single-mode He–Ne laser ($\lambda_0 = 633$ nm)	1×10^6	1 μs	300 m

EXAMPLE 2.7-1. ***Coherence Length of Light with Narrow and Broad Spectra.***

1. Light of narrow spectral width: Equations (2.7-11) and (2.7-18) provide that the coherence length l_c and spectral width $\Delta\nu_c$ are related by $l_c = c\tau_c = c/\Delta\nu_c$, where τ_c is the coherence time. Since $\nu = c/\lambda$, we have $|\Delta\nu_c| \approx (c/\lambda^2)|\Delta\lambda|$ for light of narrow spectral width, so that $l_c \approx \lambda^2/\Delta\lambda$.
2. Light with a broad uniform spectrum: As above, $l_c = c\tau_c = c/\Delta\nu_c$. For light with a uniform spectrum that extends between the wavelengths λ_{MIN} and $\lambda_{\text{MAX}} = 2\lambda_{\text{MIN}}$, we have $l_c = c/(\nu_{\text{MAX}} - \nu_{\text{MIN}}) = c/(c/\lambda_{\text{MIN}} - c/\lambda_{\text{MAX}}) = 1/(2/\lambda_{\text{MAX}} - 1/\lambda_{\text{MAX}}) = \lambda_{\text{MAX}}$.

BIBLIOGRAPHY

See also the bibliography on general optics in Chapter 1.

Scalar Waves

B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, 3rd ed. 2019, Chaps. 2, 5, 6, 12.

S. D. Gupta, N. Ghosh, and A. Banerjee, *Wave Optics: Basic Concepts and Contemporary Trends*, CRC Press/Taylor & Francis, 2016.

D. Fleisch and L. Kinnaman, *A Student's Guide to Waves*, Cambridge University Press, 2015.

J. R. Pierce, *Almost All About Waves*, MIT Press, 1974; Dover, reissued 2006.

H. J. Pain, *The Physics of Vibrations and Waves*, Wiley, 6th ed. 2005.

R. H. Webb, *Elementary Wave Optics*, Academic Press, 1969; Dover, reissued 2005.

E. Hecht and A. Zajac, *Optics*, Addison–Wesley, 2nd ed. 1990.

R. W. Wood, *Physical Optics*, Macmillan, 3rd ed. 1934; Optical Society of America, 1988.

H. D. Young, *Fundamentals of Waves, Optics, and Modern Physics*, McGraw–Hill, paperback 2nd ed. 1976.

W. E. Kock, *Sound Waves and Light Waves: The Fundamentals of Wave Motion*, Doubleday/Anchor, 1965.

Electromagnetic Waves

C. A. Bennett, *Principles of Physical Optics*, Wiley, 2nd ed. 2022.

W. Demtröder, *Electrodynamics and Optics*, Springer, 2019.

V. V. Mitin and D. I. Sementsov, *An Introduction to Applied Electromagnetics and Optics*, CRC Press/Taylor & Francis, 2017.

J.-M. Liu, *Principles of Photonics*, Cambridge University Press, 2016.

N. Ida, *Engineering Electromagnetics*, Springer, 3rd ed. 2015.

J. C. Rautio, The Long Road to Maxwell's Equations, *IEEE Spectrum*, vol. 51, no. 12, pp. 36–40 & 54–56, 2014.

F. T. Ulaby and U. Ravaioli, *Fundamentals of Applied Electromagnetics*, Prentice Hall, 7th ed. 2014.

M. N. O. Sadiku, *Elements of Electromagnetics*, Oxford University Press, 6th ed. 2014.

U. S. Inan, A. Inan, and R. Said, *Engineering Electromagnetics and Waves*, Prentice Hall, 2nd ed. 2014.

D. Fleisch, *A Student's Guide to Maxwell's Equations*, Cambridge University Press, 2013.

M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, 7th ed. 2002.

S. A. Akhmanov and S. Yu. Nikitin, *Physical Optics*, Oxford University Press, 1997.

H. A. Haus, *Waves and Fields in Optoelectronics*, Prentice Hall, 1984.

Random Waves

L. Dal Negro, *Waves in Complex Media*, Cambridge University Press, 2022.

J. W. Goodman, *Statistical Optics*, Wiley, 2nd ed. 2015.

O. Korotkova, *Random Light Beams: Theory and Applications*, CRC Press/Taylor & Francis, 2014.

E. Wolf, *Introduction to the Theory of Coherence and Polarization of Light*, Cambridge University Press, 2007.

L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics*, Cambridge University Press, 1995.

L. Mandel and E. Wolf, eds., *Selected Papers on Coherence and Fluctuations of Light (1850–1966)*, SPIE Optical Engineering Press (Milestone Series Volume 19), 1990.

J. Peřina, *Coherence of Light*, Reidel, 1971, 2nd ed. 1985.

L. Mandel and E. Wolf, eds., *Selected Papers on Coherence and Fluctuations of Light*, Volumes 1 and 2, Dover, 1970.

Optical Constants

P. Hartmann, *Optical Glass*, SPIE Optical Engineering Press, 2014.

M. Bass, C. DeCusatis, J. Enoch, V. Lakshminarayanan, G. Li, C. Macdonald, V. Mahajan, and E. van Stryland, eds., *Handbook of Optics*, 3rd ed., McGraw–Hill, 2010.

E. D. Palik, ed., *Handbook of Optical Constants of Solids III*, Academic Press, 1998.

Electromagnetics Classics

J. D. Jackson, *Classical Electrodynamics*, Wiley, 3rd ed. 1999.

S. Ramo, J. R. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*, Wiley, 3rd ed. 1994.

H. A. Haus and J. R. Melcher, *Electromagnetic Fields and Energy*, Prentice Hall, 1989.

L. D. Landau, E. M. Lifshitz, and L. P. Pitaevskii, *Electrodynamics of Continuous Media*, Nauka (Moscow), 2nd ed. 1982; Butterworth–Heinemann, 2nd English ed. 1984, reprinted 2004.

L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields*, Nauka (Moscow), 6th revised ed. 1973; Butterworth–Heinemann, 4th revised English ed. 1975, reprinted with corrections 2000.

J. A. Stratton, *Electromagnetic Theory*, McGraw–Hill, 1941; Wiley–IEEE Classic reissue 2007.

Historical

P. Daukantas, 200 Years of Fresnel's Legacy, *Optics & Photonics News*, vol. 26, no. 9, pp. 40–47, 2015.

T. Levitt, *A Short Bright Flash: Augustin Fresnel and the Birth of the Modern Lighthouse*, Norton, 2013.

F. J. Dijksterhuis, *Lenses and Waves: Christiaan Huygens and the Mathematical Science of Optics in the Seventeenth Century*, 2004, Springer, paperback ed. 2011.

J. Z. Buchwald, *The Rise of the Wave Theory of Light: Optical Theory and Experiment in the Early Nineteenth Century*, University of Chicago Press, paperback ed. 1989.

O. Heaviside, *Electromagnetic Theory*, "The Electrician" Printing and Publishing (London), 1893.

J. C. Maxwell, *A Treatise on Electricity and Magnetism*, Macmillan/Clarendon Press (Oxford), 1873/1881; Wentworth Press, paperback ed. 2019.

C. Huygens, *Treatise on Light*, 1690, University of Chicago Press, 1945; Echo Library, reprinted 2007.

CHAPTER

3

PHOTONS

3.1 THE PHOTON 63

3.2 PHOTON ENERGY, FREQUENCY, AND WAVELENGTH 65

3.3 PHOTON POSITION AND TIME 66

3.4 PHOTON STREAMS 69

3.5 RANDOMNESS OF PHOTON FLOW 73

3.6 PHOTON-NUMBER STATISTICS 75

3.7 RANDOM PARTITIONING OF PHOTON STREAMS 79

Max Planck (1858–1947) suggested that the emission and absorption of light by matter takes the form of quanta of energy.

Albert Einstein (1879–1955) advanced the hypothesis that light itself comprises quanta of energy.

LED Lighting: Devices and Colorimetry. Malvin Carl Teich. Google Books. Published 2024.

From a historical perspective, the theories of optics developed roughly in the following sequence: 1) ray optics → 2) wave optics → 3) electromagnetic optics → 4) quantum optics. These models are progressively more complex and sophisticated, and evolved to provide explanations for the outcomes of increasingly subtle and precise optical experiments. Ray optics, wave optics, and electromagnetic optics are all approximate theories that derive their validity from their successes in generating results that approximate those based on the more rigorous quantum optics, which properly describes almost all known optical phenomena.

In the mathematical framework of **quantum optics**, the vectors $\mathbf{E}$ and $\mathbf{H}$ that represent the electric and magnetic fields of classical electromagnetic optics, respectively, are promoted to operator status in a Hilbert space. These operators are assumed to satisfy certain operator equations and commutation relations that govern their time dynamics and interdependence. Although the equations of quantum optics describe the interactions of electromagnetic fields with matter in much the same way as Maxwell's equations, the results incorporate intrinsic quantum uncertainties. Nevertheless, in spite of its vast successes, quantum optics is not the final arbiter of *all* optical effects. That distinction currently belongs to the **electroweak theory**, which combines quantum electrodynamics with the theory of weak interactions. Ongoing efforts seek to combine electroweak theory with the theories of strong and gravitational interactions in an attempt to forge a general unified field theory that accommodates all four fundamental forces of nature, as they are currently understood.

Although a formal treatment of quantum optics lies beyond the scope of this text, many of the quantum properties of light and its interaction with matter can be described by supplementing electromagnetic optics with several simple relationships that embody the corpuscularity, localization, and fluctuations of quantum fields and energy. This set of rules, called **photon optics**, offers a convenient way of dealing with some quantum-optical phenomena that lie beyond the reach of classical optics, while retaining classical theory as a limiting case. Photon optics also proves useful for elucidating various features of classical light, and is used extensively in the remainder of this text.

From a mathematical perspective, ray optics is the limit of wave optics when the wavelength is infinitesimally small, wave optics is the limit of electromagnetic optics when the polarization properties anchored in its vector character play no role in the problem under consideration, and electromagnetic optics is the limit of quantum optics when the particle-like behavior of light associated with its operator properties can be overlooked. The hierarchy that emerges is depicted in Fig. 3.0-1: quantum optics encompasses electromagnetic optics, which encompasses wave optics, which in turn encompasses ray optics.

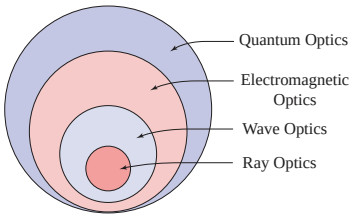


Figure 3.0-1 The theory of quantum optics provides an explanation for virtually all optical phenomena. The electromagnetic theory of light (electromagnetic optics) provides the most complete treatment of light within the confines of classical optics. Wave optics is a scalar approximation of electromagnetic optics. Ray optics is the limit of wave optics when the wavelength is very short.

The point of departure for each model of light is a set of principles, or postulates, from which a large body of results are generated. The postulates of each model are special cases of those of the next higher-level model. In addressing a particular problem in optics, the optimal choice of a model is the simplest that satisfactorily describes a

particular phenomenon to a specified degree of accuracy. Although it can be difficult to make the choice *a priori*, experience often serves as a guide.

We begin by introducing the concept of the photon and examining its properties. Using electromagnetic optics as a point of departure, we then impose a number of rules that govern the behavior of photon energy, position, and time. This is followed by a discussion of the properties of photon streams, including their randomness, photon-number statistics, and partitioning. The interactions of photons with atoms and semiconductors are described in Chapters 4 and 6, respectively.

3.1 THE PHOTON

From a quantum perspective, light consists of particles called **photons** or **quanta**. A photon carries energy and momentum, as well as spin angular momentum associated with its polarization and orbital angular momentum related to the twist of its wavefront. The photon has zero rest mass and travels at c_0, the speed of light in vacuum, and at $c = c_0/n < c_0$ in dielectric media. A photon also has a wavelike character that determines its localization properties in space and time, and the rules by which it interferes and diffracts.

The notion of the photon initially grew out of an attempt by Max Planck (p. 61) in 1900 to resolve a long-standing riddle concerning the spectrum of blackbody radiation emanating from a cavity held at a fixed temperature T (this topic is discussed in Sec. 4.7). Planck ultimately achieved this goal by assuming that the atoms in the walls of the cavity absorbed and emitted energy only as integral multiples of a small unit of energy, i.e., as quanta. In 1905, Albert Einstein (p. 61) extended Planck's notion of energy quantization by considering the light itself to be a collection of light quanta. This enabled Einstein to successfully explain the photoelectric effect, a feat that garnered him the 1921 Nobel Prize in physics. The term **photon**, introduced by Gilbert Lewis in 1926, came to be used to describe what Einstein had originally termed *Lichtquant*.

The concept of the photon and the rules of photon optics are introduced by considering light in an optical cavity. This is a convenient choice because it restricts the space under consideration to a simple geometry. More importantly, the presence of the cavity turns out not to be an important feature of the argument; the results are independent of the form of the cavity, and even of its presence.

Light in a 3D Cavity: Electromagnetic-Optics Perspective

Electromagnetic optics dictates that light in a lossless three-dimensional (3D) cavity of volume V is completely characterized by an electromagnetic field that takes the form of a superposition of discrete orthogonal modes with different spatial distributions, frequencies, and polarizations. The overall electric-field vector, $\mathcal{E}(\mathbf{r}, t) = \text{Re}\{\mathbf{E}(\mathbf{r}, t)\}$, may be expressed in terms of the complex electric field vector $\mathbf{E}(\mathbf{r}, t)$ as

$$\mathbf{E}(\mathbf{r}, t) = \sum_{\mathbf{q}} A_{\mathbf{q}} U_{\mathbf{q}}(\mathbf{r}) \exp(j2\pi\nu_{\mathbf{q}} t)\, \widehat{\mathbf{e}}_{\mathbf{q}}\,, \tag{3.1-1}$$

where the $\mathbf{q}$th mode is characterized by its complex envelope $A_{\mathbf{q}}$, its frequency $\nu_{\mathbf{q}}$, its polarization along the direction of the unit vector $\widehat{\mathbf{e}}_{\mathbf{q}}$, and its spatial distribution characterized by the complex function $U_{\mathbf{q}}(\mathbf{r})$, which is normalized such that $\int_V |U_{\mathbf{q}}(\mathbf{r})|^2\, d\mathbf{r} = 1$.

For convenience, we consider a cubic cavity of dimension d and standing-wave

spatial expansion functions given by

$$U_{\mathbf{q}}(\mathbf{r}) = (2/d)^{3/2} \sin(q_x \pi x/d) \sin(q_y \pi y/d) \sin(q_z \pi z/d), \tag{3.1-2}$$

where the integers q_x, q_y, and q_z are specified by the shorthand notation (q_x, q_y, q_z) [Fig. 3.1-1(a)]. In accordance with (2.6-10), the energy density associated with mode $\mathbf{q}$ is $\frac{1}{2}\epsilon|A_{\mathbf{q}}|^2 |U_{\mathbf{q}}(\mathbf{r})|^2$, so that the energy contained in mode $\mathbf{q}$ is

$$E_{\mathbf{q}} = \tfrac{1}{2}\epsilon \int_V |A_{\mathbf{q}}|^2 |U_{\mathbf{q}}(\mathbf{r})|^2 \, d\mathbf{r} = \tfrac{1}{2}\epsilon |A_{\mathbf{q}}|^2, \tag{3.1-3}$$

where V is the modal volume. In classical electromagnetic theory, the energy $E_{\mathbf{q}}$ can assume an arbitrary nonnegative value, no matter how small, and the total energy is the sum of the energies in all modes.

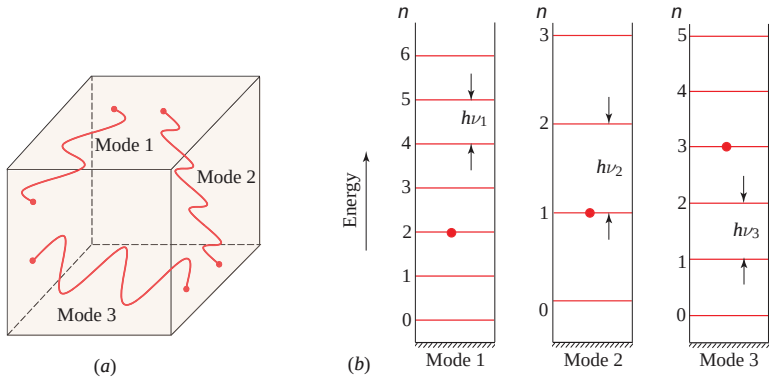


Figure 3.1-1 (a) Schematic of three electromagnetic modes of different frequencies and directions in a cubic cavity. (b) Allowed energy levels of three modes in the context of photon optics. Modes 1, 2, and 3 have frequencies ν_1, ν_2, and ν_3, respectively. In the sketch displayed, modes 1, 2, and 3 contain $n = 2$, 1, and 3 photons, respectively, as represented by the filled circles. The number of photons in a mode can be zero, fixed, or random.

It is important to note that the expansion functions $U_{\mathbf{q}}(\mathbf{r})$, along with $\exp(j2\pi\nu_{\mathbf{q}}t)$, and $\hat{\mathbf{e}}_{\mathbf{q}}$ as specified above, are not unique. Other choices are available, including those comprising polychromatic modes.

Light in a 3D Cavity: Photon-Optics Perspective

The electromagnetic-optics approach described above is preserved in photon optics, but a restriction is placed on the energy that each mode may carry. Rather than comprising a continuous range, with no minimum energy, the modal energy is instead restricted to a ladder of discrete values separated by the fixed energy $h\nu$, where ν is the frequency of the mode, as displayed in Fig. 3.1-1(b). The energy associated with a mode is thus quantized, with only integral units of the fixed energy $h\nu$ permitted. Each unit of energy is carried by a single photon and the mode may carry an arbitrary number of photons.

Light in a cavity comprises a set of modes, each containing an integral number of identical photons. Characteristics of the mode, such as its frequency, spatial distribution, direction of propagation, and polarization, are assigned to the photons.

3.2 PHOTON ENERGY, FREQUENCY, AND WAVELENGTH

Photon Energy and Frequency

Photon optics provides that the energy associated with an electromagnetic mode is quantized to discrete levels separated by the photon energy, as sketched in Fig. 3.1-1(*b*). The energy of a photon in a mode of frequency $\nu = \omega/2\pi$ is

$$E = h\nu = \hbar\omega, \tag{3.2-1}$$

Photon Energy

where $h = 6.6261 \times 10^{-34}$ J·s is **Planck's constant**, and $\hbar \equiv h/2\pi$. Energy may be added to, or subtracted from, a given mode only as individual photons, namely in units of $h\nu$. Frequency and photon energy (specified in units of eV, J, and cm^{-1}) are displayed in Fig. 3.2-1 for the optical and microwave regions of the spectrum.

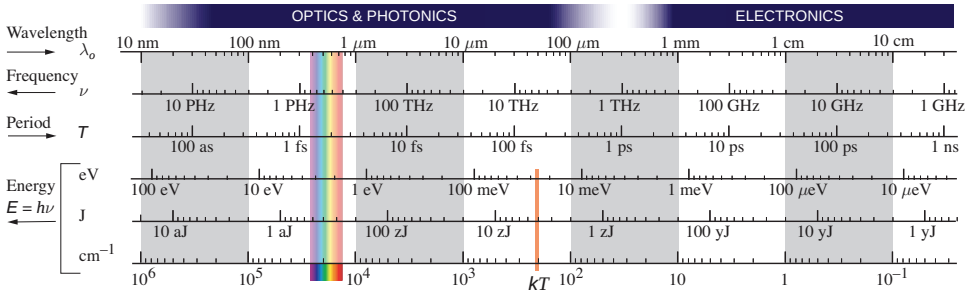


Figure 3.2-1 Relationships among wavelength λ_0, frequency ν, period T, and photon energy E (specified in units of eV, J, and cm^{-1}) for the optical and microwave regions. A photon of free-space wavelength $\lambda_0 = 1$ μm has frequency $\nu = 300$ THz, period $T = 3.33$ fs, and energy $E = 1.24$ eV = 199 zJ = 10^4 cm^{-1}. At room temperature ($T = 300$ K), the thermal energy $kT = 26$ meV = 4.17 zJ = 210 cm^{-1}. Two spectral domains are indicated: 1) optics & photonics, and 2) electronics.

According to quantum optics, all modes, including those carrying zero photons, also carry **zero-point energy** $E_0 = \frac{1}{2}h\nu$, which is associated with the fluctuations of the vacuum field. An electromagnetic mode carrying n photons therefore has total energy

$$E_n = (n + 1/2)\, h\nu, \qquad n = 0, 1, 2, \ldots . \tag{3.2-2}$$

This expression matches that for a quantum-mechanical harmonic oscillator; indeed the two systems are isomorphic. The zero-point energy is seldom directly observed because optical measurements usually involve energy differences (e.g., $E_2 - E_1$). However it is responsible for the spontaneous emission of a photon by an atom, as discussed in Sec. 4.3, and is a source of noise that limits the sensitivity of certain precision measurements. Vacuum fluctuations are also the origin of the **Casimir effect**, a small attractive force that acts between two parallel uncharged conducting plates located in close proximity.

Because photon energy is directly proportional to frequency, the particle nature of light becomes increasingly prevalent as the radiation frequency increases and the wavelength concomitantly decreases. In most interactions, X-rays and gamma-rays, with their high frequencies, behave more like particles than waves, although wavelike effects (such as X-ray diffraction) can be observed. The frequencies of radio waves, in contrast, are so low that they rarely exhibit any particle-like behavior. The optical region lies in an intermediate frequency range such that both particle-like and wavelike behavior are readily observed.

Photon Wavelength and Period

The order of magnitude of the photon energy is easily estimated. An infrared photon of wavelength $\lambda_0 = 1\ \mu\text{m}$ in free space has a frequency $\nu \approx 3 \times 10^{14}$ Hz by virtue of the relation $\lambda_0 \nu = c_0$, and has a period $T = 1/\nu$. Its energy is thus $E = h\nu \approx 1.99 \times 10^{-19}$ J $= 199$ zJ. Since the electron charge is $e \approx 1.6 \times 10^{-19}$ C, the photon energy expressed in electron volts is given by $h\nu/e \approx 1.99 \times 10^{-19}/1.6 \times 10^{-19} \approx 1.24$ eV; this is equivalent to the kinetic energy imparted to an electron when it is accelerated through a potential difference of 1.24 V. For a microwave photon of wavelength of 1 cm, the photon energy is a factor of 10^4 smaller, so that $h\nu = 1.24 \times 10^{-4}$ eV. A convenient approximate conversion formula between free-space wavelength (μm) and photon energy (eV) is therefore

$$E\ (\text{eV}) \approx \frac{1.24}{\lambda_0\ (\mu\text{m})}\,. \tag{3.2-3}$$

The reciprocal wavelength is also used as a unit of energy, particularly in chemistry. It is usually specified in cm^{-1} and is determined by expressing the wavelength in cm and simply taking the inverse. Hence, 1 eV corresponds to $10^4/1.24 \approx 8065\ \text{cm}^{-1}$. Conversions among photon energy, wavelength, frequency, and period in the optical and microwave regions of the electromagnetic spectrum are provided in Fig. 3.2-1.

3.3 PHOTON POSITION AND TIME

Photon Position

Associated with a photon of energy $h\nu$ is a monochromatic wave described by the complex wavefunction $U(\mathbf{r})\exp(j2\pi\nu t)$ of the mode. When such a photon impinges on a detector of small area dA that is normal to the direction of propagation, its indivisibility causes it to either be wholly detected or not detected at all. If detected, the location $\mathbf{r}$ at which the photon is registered is not precisely determined. Rather, it is governed by the optical intensity $I(\mathbf{r}) \propto |U(\mathbf{r})|^2$ at the detector, in accordance with the following probabilistic law:

The probability $p(\mathbf{r})\,dA$ of observing a photon at the position $\mathbf{r}$ within an incremental area dA, at any time, is proportional to the local optical intensity of the mode $I(\mathbf{r}) \propto |U(\mathbf{r})|^2$, so that

$$p(\mathbf{r})\,dA \propto I(\mathbf{r})\,dA. \tag{3.3-1}$$

Photon Position

The photon is therefore more likely to be found at those locations of higher intensity. As an example, a photon in a mode described by a standing wave with the intensity distribution $I(x, y, z) \propto \sin^2(\pi z/d)$, with $0 \leqslant z \leq d$, is most likely to be detected at $z = d/2$, but will never be detected at $z = 0$ or $z = d$. The localized nature of a photon is manifested when it is detected. Unlike a wave, which is extended in space, and a particle, which is localized in space, an optical photon behaves as *both* an extended *and* a localized entity, behavior referred to as **wave–particle duality**.

EXAMPLE 3.3-1. ***Transmittance of a Single Photon at a Beamsplitter.*** An ideal beamsplitter is an optical device that losslessly splits a beam of light into two beams that emerge at right angles to each other. It is characterized by an intensity transmittance $\mathcal{T}$ and an intensity reflectance $\mathcal{R} = 1 - \mathcal{T}$. The intensity of the transmitted wave I_t and the intensity of the reflected wave I_r can be calculated from the intensity of the incident wave I using the electromagnetic relations $I_t = \mathcal{T}I$ and $I_r = (1-\mathcal{T})I$. Because a photon is indivisible, however, it must choose between the two possible exit directions permitted by the beamsplitter. A single photon incident on the device will follow these directions in accordance with the probabilistic photon-position rule (3.3-1). The probability that the photon is transmitted is proportional to I_t and is therefore given by the transmittance $\mathcal{T} = I_t/I$. The probability that it is reflected is $1-\mathcal{T} = I_r/I$. From the perspective of probability theory, the problem is identical to that of flipping a biased coin. Figure 3.3-1 illustrates the process.

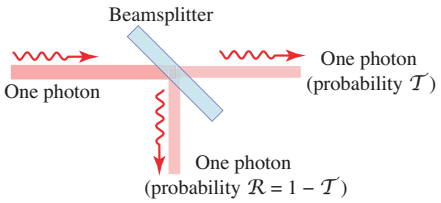


Figure 3.3-1 Probabilistic reflection or transmission of a photon at a lossless beamsplitter.

EXAMPLE 3.3-2. ***Single-Photon Imaging.*** A coherent imaging system is characterized by an impulse response function $h(x, y; x', y')$ that links its output and input fields, $U_o(x, y)$ and $U_i(x, y)$, respectively, via the two-dimensional convolution

$$U_o(x,y) = \int_{-\infty}^{\infty}\int U_i(x',y')\, h(x,y;x',y')\, dx'\, dy'. \tag{3.3-2}$$

The same relationship characterizes the single-photon wavefunctions at the output and input of a single-photon imaging system, where $|U_o(x)|^2$ represents the probability density function of the photon position in the image plane.

Photon Time

The modal expansion provided in (3.1-1) comprises monochromatic modes that are "eternal" harmonic functions of time; a photon in a monochromatic mode is equally likely to be detected at any time. As indicated earlier, however, a modal expansion of the radiation inside (or outside) a cavity is not unique. A more general expansion comprises polychromatic modes such as time-localized wavepackets. The probability of detecting a photon characterized by the complex wavefunction $U(\mathbf{r}, t)$, at any position and in the incremental time interval between t and $t + dt$, is proportional to $I(\mathbf{r}, t)\, dt \propto |U(\mathbf{r}, t)|^2\, dt$. The photon-position rule of photon optics displayed in (3.3-1) may therefore be generalized to include photon time localization, as follows:

The probability $p(\mathbf{r}, t)\, dA\, dt$ of observing a photon at the position $\mathbf{r}$ within an incremental area dA, and during an incremental time interval dt following time t, is proportional to the local optical intensity of the mode $I(\mathbf{r}, t) \propto |U(\mathbf{r}, t)|^2$, so that

$$p(\mathbf{r}, t)\, dA\, dt \propto I(\mathbf{r}, t)\, dA\, dt. \tag{3.3-3}$$

Photon Position and Time

Time–Energy Uncertainty

The time during which a photon in a *monochromatic mode* of frequency ν may be detected is completely uncertain, whereas the value of its frequency ν (and its energy $E = h\nu$) is completely certain.

In contrast, a photon in a *wavepacket mode* with an intensity function $I(t)$ of duration σ_t must be localized within that time. Bounding the photon time in this way is accompanied by an uncertainty in its frequency (and energy) by virtue of the properties of the Fourier transform, and corresponds to a polychromatic photon. Suppressing the position dependence for simplicity, this frequency uncertainty is readily determined by expanding $U(t)$ as a superposition of its harmonic components,

$$U(t) = \int_{-\infty}^{\infty} V(\nu) \exp(j2\pi\nu t)\, d\nu, \tag{3.3-4}$$

where $U(t)$ is the inverse Fourier transform of $V(\nu)$, as defined in (A.1-1).

Denoting the power-RMS temporal width of the function $U(t)$ as σ_t, and the power-RMS spectral width of the function $V(\nu)$ as σ_ν, the product of σ_t and σ_ν for this Fourier-transform pair obeys the duration–bandwidth reciprocity relation set forth (A.2-4), which reads

$$\sigma_t \sigma_\nu \geqslant 1/4\pi\,. \tag{3.3-5}$$

Since angular frequency and frequency are related by $\omega = 2\pi\nu$, (3.3-5) can be written in the alternate form specified in (A.2-7):

$$\sigma_t \sigma_\omega \geqslant 1/2\,. \tag{3.3-6}$$

The definitions of σ_t and σ_ν that accompany these uncertainty relations are provided in (A.2-3) and (A.2-5), respectively. The results presented in (3.3-5) and (3.3-6) follow solely from the properties of the Fourier transform.

Moreover, since the energy of a photon is given by $E = \hbar\omega$, in accordance with (3.2-1), it is not possible to specify the photon energy to an accuracy better than

$$\sigma_E = \hbar\sigma_\omega\,. \tag{3.3-7}$$

It follows from (3.3-6) and (3.3-7) that the time during which a photon may be detected σ_t, and its energy uncertainty σ_E, satisfy the **Heisenberg time–energy uncertainty relation**:

$$\boxed{\sigma_t \sigma_E \geqslant \hbar/2\,.} \tag{3.3-8}$$

Heinsenberg Time–Energy Uncertainty Relation

The inequality set forth in (3.3-8) is analogous to the Heisenberg position–momentum uncertainty relation provided in (A.2-9), which sets a limit on the precision with which the position x and momentum p of a photon can be simultaneously specified. The average energy $\overline{E}$ of the polychromatic photon is $\overline{E} = h\overline{\nu} = \hbar\overline{\omega}$.

Reiterating, a monochromatic photon ($\sigma_\nu \to 0$) has an eternal duration within which it can be observed ($\sigma_t \to \infty$). A photon associated with an optical wavepacket, on the other hand, is localized in time and must therefore be polychromatic, which implies a corresponding energy uncertainty. We conclude by noting that a *wavepacket photon* can be viewed as a confined traveling packet of energy.

Summary: Photon Energy, Momentum, Spin, Position, and Time

Electromagnetic radiation may be described in terms of a sum of modes, one example being monochromatic uniform plane waves of the form:

$$\mathbf{E}(\mathbf{r},t)=\sum_{\mathbf{q}} A_{\mathbf{q}}\exp(-j\mathbf{k}_{\mathbf{q}}\cdot\mathbf{r})\exp(j2\pi\nu_{\mathbf{q}}t)\,\widehat{\mathbf{e}}_{\mathbf{q}}. \tag{3.3-9}$$

Each plane wave has two orthogonal polarization states represented by the vectors $\widehat{\mathbf{e}}_{\mathbf{q}}$ (e.g., vertical/horizontal linearly polarized, right/left circularly polarized). When the energy of a mode is measured, the result is an integer (in general, random) number of photons (energy quanta). Each of the photons associated with the mode $\mathbf{q}$ has the following properties:

- Energy $E=h\nu_{\mathbf{q}}$. (3.3-10)
- Momentum $\mathbf{p}=\hbar\mathbf{k}_{\mathbf{q}}$, with magnitude $p=\hbar k=h/\lambda$. (3.3-11)
- Spin angular momentum (helicity) $\mathcal{S}=\pm\hbar$, if circularly polarized. (3.3-12)
- The photon is equally likely to be found anywhere in space, and at any time, since the wavefunction of the mode is a monochromatic plane wave.

The choice of modes is not unique, however. A modal expansion in terms of nonmonochromatic (quasi-monochromatic), non-plane waves, is also possible:

$$\mathbf{E}(\mathbf{r},t)=\sum_{\mathbf{q}} A_{\mathbf{q}}\,U_{\mathbf{q}}(\mathbf{r},t)\,\widehat{\mathbf{e}}_{\mathbf{q}}. \tag{3.3-13}$$

Each of the photons associated with the mode $\mathbf{q}$ then has the following properties:

- The photon position and time are governed by the complex wavefunction $U_{\mathbf{q}}(\mathbf{r},t)$. The probability of observing the photon at position $\mathbf{r}$ within an incremental area dA, and during an incremental time interval dt following time t, is proportional to $|U_{\mathbf{q}}(\mathbf{r},t)|^2\,dA\,dt$.
- If $U_{\mathbf{q}}(\mathbf{r},t)$ has a finite time duration σ_t, i.e., if the photon is localized in time, then the photon energy $h\nu_{\mathbf{q}}$ has an uncertainty $h\sigma_\nu\geqslant h/4\pi\sigma_t$.
- If $U_{\mathbf{q}}(\mathbf{r},t)$ has a finite spatial extent in the transverse $(z=0)$ plane, i.e., if the photon is localized in the x direction, for example, then the direction of the photon momentum is uncertain. The spread in photon momentum can be determined by analyzing $U_{\mathbf{q}}(\mathbf{r},t)$ as a sum of plane waves; the wave with wavevector $\mathbf{k}$ corresponds to photon momentum $\hbar\mathbf{k}$. Spatial localization of the photon in the transverse plane results in an increase in the uncertainty of the photon-momentum direction.

3.4 PHOTON STREAMS

In Sec. 3.1 we concentrated on the properties and behavior of single photons. We now consider the properties of collections of photons. Photon streams often contain numerous propagating modes. As a result of the processes by means of which photons are created (e.g., atomic emissions, as discussed in Chapter 4), the number of photons occupying any mode is generally random. If an experiment is carried out in which

a weak stream of photons falls on a photosensitive surface, the individual photons are registered (detected) at random localized instants of time and at random points in space, in accordance with (3.3-3). The temporal and spatial behavior of the photon registrations can be highlighted by examining the two features separately.

The *temporal pattern* is revealed by making use of a photodetector such as a single-photon avalanche diode (SPAD), which has good temporal resolution but integrates light over a finite area A, as illustrated in Fig. 3.4-1. Equation (3.3-3), together with the relation $P(t) = \int_A I(\mathbf{r}, t)\, dA$, show that the probability of detecting a photon in the incremental time interval between t and $t + dt$ is proportional to $P(t)$, the optical power at time t. Different forms of partially coherent light exhibit different kinds of intrinsic optical-power fluctuations. Moreover, the optical power can be deliberately manipulated to carry out designer experiments.

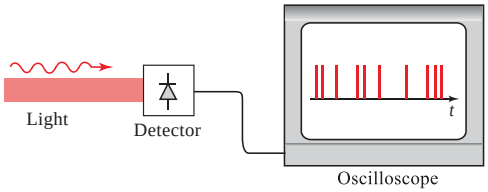


Figure 3.4-1 Individual photon registrations at random localized instants of time for a detector with good temporal resolution that integrates light over an area A.

The *spatial pattern* of photon registrations, on the other hand, is readily manifested by making use of a detector with good spatial resolution that integrates over a fixed exposure time T, such as photographic film. In accordance with (3.3-3), the probability of observing a photon in an incremental area dA surrounding the point $\mathbf{r}$ is proportional to the local time-integrated intensity, $\int_0^T I(\mathbf{r}, t)\, dt$. The random locations of the photon registrations are illustrated in the grainy image of Max Planck provided in Fig. 3.4-2. This image was obtained by rephotographing a high-contrast photograph of Max Planck under very low light conditions. Each white dot in the photograph represents a random photon registration and the density of these registrations follows the local spatial intensity. In 1932, Barnes & Czerny observed that quantum fluctuations could be discerned by the dark-adapted eye at low light levels, which they properly attributed to the shot effect of photons.[†]

Figure 3.4-2 Image of Max Planck under illumination with a sparse stream of photons. The spatial density of the collection of individual photon registrations follows the local integrated intensity. (Adapted from B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, 3rd ed. 2019, Fig. 13.2-2.)

We begin this section by introducing a number of definitions that relate the quantum measures for the mean flow of photons to the classical electromagnetic measures set forth in Secs. 2.1 and 2.6, namely intensity, power, and energy. These definitions are inspired by (3.3-3), which governs the position and time at which individual photons are registered.

[†] R. B. Barnes and M. Czerny, Lässt sich ein Schroteffeckt der Photonen mit dem Auge beobachten?, *Zeitschrift für Physik*, vol. 79, pp. 436–449, 1932.

Mean Photon-Flux Density

Monochromatic light of frequency ν and constant classical intensity $I(\mathbf{r})$ $(\mathrm{W/cm^2})$ carries a mean **photon-flux density**

$$\phi(\mathbf{r}) = \frac{I(\mathbf{r})}{h\nu}. \tag{3.4-1}$$

Mean Photon-Flux Density

Since each photon carries energy $h\nu$, this equation provides a straightforward conversion from a classical measure (energy/s-cm^2) to a quantum measure (photons/s-cm^2). For quasi-monochromatic light of central frequency $\overline{\nu}$, all photons have approximately the same energy $h\overline{\nu}$, so that the mean photon-flux density is approximately

$$\phi(\mathbf{r}) \approx \frac{I(\mathbf{r})}{h\overline{\nu}}. \tag{3.4-2}$$

EXAMPLE 3.4-1. ***Mean Photon-Flux Density and Optical Intensity.*** Typical values of I and ϕ for several common sources of light are provided in Table 3.4-1. It is clear from these values that trillions of photons rain down on each square centimeter of each exposed object each second.

Table 3.4-1 Classical intensity and mean photon-flux density for various sources of light.

SOURCE	Intensity I (J/s-cm^2)	Mean Photon-Flux Density ϕ (photons/s-cm^2)
Starlight	4×10^{-13}	10^6
Moonlight	4×10^{-11}	10^8
Twilight	4×10^{-9}	10^{10}
Indoor light	4×10^{-7}	10^{12}
Sunlight	4×10^{-5}	10^{14}
Laser light[a]	$4 \times 10^{+3}$	10^{22}

[a] A 12-mW green laser beam at a free-space wavelength $\lambda_0 = 500\,\mathrm{nm}$, focused to a 20-$\mu$m-diameter spot.

Mean Photon Flux

The mean**photon flux** Φ (photons/s) is obtained by integrating the mean photon-flux density over a specified area A,

$$\Phi = \int_A \phi(\mathbf{r})\, dA = \frac{P}{h\overline{\nu}}, \tag{3.4-3}$$

Mean Photon Flux

where the optical power is

$$P = \int_A I(\mathbf{r})\, dA, \tag{3.4-4}$$

and $h\overline{\nu}$ is again the average energy of a photon.

EXAMPLE 3.4-2. ***Mean Photon Flux and Optical Power.*** An optical power of 1 nW at a free-space wavelength $\lambda_0 = 0.2\ \mu\text{m}$ corresponds to an average photon flux $\Phi \approx 10^9$ photons/s. Roughly speaking, one photon then strikes the object every nanosecond:

$$1\ \text{nW at}\ \lambda_0 = 0.2\ \mu\text{m} \Rightarrow 1\ \text{photon/ns}. \tag{3.4-5}$$

As a comparison, a photon of wavelength $\lambda_0 = 1\ \mu\text{m}$ carries one-fifth as much energy, in which case 1 nW corresponds to an average of 5 photons/ns.

Mean Photon Number

The mean **photon number** $\overline{n}$ detected over the area A, during the time interval T, is obtained by multiplying the mean photon flux Φ in (3.4-3) by the time duration, which leads to

$$\boxed{\overline{n} = \Phi T = \frac{E}{h\overline{\nu}},} \tag{3.4-6}$$

Mean Photon Number

where $E = PT$ is the optical energy. The photon number is also called the **photon count**. The relationships between the classical and quantum measures of mean photon flow are summarized in the table below:

Summary: Classical and Quantum Measures of Mean Photon Flow

Classical		Quantum	
Optical intensity	$I(\mathbf{r})$	Mean photon-flux density	$\phi(\mathbf{r}) = I(\mathbf{r})/h\overline{\nu}$
Optical power	P	Mean photon flux	$\Phi = P/h\overline{\nu}$
Optical energy	E	Mean photon number	$\overline{n} = E/h\overline{\nu}$

Spectral Measures

For polychromatic light of nonnegligible bandwidth, it is useful to define frequency-based spectral versions of the classical intensity, power, and energy, together with their respective quantum counterparts (these are designated by the subscript ν): spectral photon-flux density, spectral photon flux, and spectral photon number, as indicated in the following table:

Summary: Classical and Quantum Spectral Measures

Classical		Quantum	
I_ν	(W/cm^2-Hz)	$\phi_\nu = I_\nu/h\nu$	(photons/s-cm^2-Hz)
P_ν	(W/Hz)	$\Phi_\nu = P_\nu/h\nu$	(photons/s-Hz)
E_ν	(J/Hz)	$\overline{n}_\nu = E_\nu/h\nu$	(photons/Hz)

As an example, $I_\nu\, d\nu$ represents the spectral intensity in the frequency range between ν and $\nu + d\nu$ while $\overline{n}_\nu\, d\nu$ represents the spectral photon number in the frequency range between ν and $\nu + d\nu$. Hence, a polychromatic flash of light with an intensity that is uniform in space and time, and that comprises a uniform band of optical frequencies of width Γ, carries energy $E = I_\nu \Gamma A T$ and mean photon number $\overline{n} = E/h\bar{\nu} = I_\nu \Gamma A T / h\bar{\nu}$.

Wavelength-based spectral measures, designated by the subscript λ, are widely used for characterizing broadband sources in statistical optics and in radiometry; examples are provided in Fig. 2.7-5 and in Secs. 8.8 and 9.7, respectively.

Time-Varying Light

If the light intensity varies with time, it follows that the mean photon-flux density specified (3.4-1) is also a function of time,

$$\phi(\mathbf{r}, t) = \frac{I(\mathbf{r}, t)}{h\bar{\nu}}\,. \tag{3.4-7}$$

Mean Photon-Flux Density

The mean photon flux and optical power are then functions of time as well,

$$\Phi(t) = \int_A \phi(\mathbf{r}, t)\, dA = \frac{P(t)}{h\bar{\nu}}\,, \tag{3.4-8}$$

Mean Photon Flux

where

$$P(t) = \int_A I(\mathbf{r}, t)\, dA\,. \tag{3.4-9}$$

As a consequence, the mean photon number registered in a time interval between $t = 0$ and $t = T$, obtained by integrating the photon flux, also varies with time

$$\overline{n} = \int_0^T \Phi(t)\, dt = \frac{E}{h\bar{\nu}}\,, \tag{3.4-10}$$

Mean Photon Number

where the mean optical energy (intensity integrated over time and area) is given by

$$E = \int_0^T P(t)\, dt = \int_0^T \int_A I(\mathbf{r}, t)\, dA\, dt\,. \tag{3.4-11}$$

3.5 RANDOMNESS OF PHOTON FLOW

When the classical intensity $I(\mathbf{r}, t)$ is constant, the time and position at which a single photon is detected are governed by (3.3-3), which dictates that the probability density of detecting that photon at the space–time point $(\mathbf{r}, t)$ is proportional to $I(\mathbf{r}, t)$. The classical electromagnetic intensity $I(\mathbf{r}, t)$ also governs the behavior of photon streams, but the interpretation ascribed to $I(\mathbf{r}, t)$ differs:

For photon streams, the classical intensity $I(\mathbf{r}, t)$ determines the mean photon-flux density $\phi(\mathbf{r}, t)$. The fluctuations of $\phi(\mathbf{r}, t)$ are established by the statistical characteristics of the light source emitting the photons.

Photon Arrival Times

Consider a detector that integrates over space, such as that illustrated in Fig. 3.4-1. If the intensity I is constant in time, then so too is the power P. The mean photon-flux density is then $\phi = I/h\bar{\nu}$ and the mean photon flux is $\Phi = P/h\bar{\nu}$. However, the times at which the photons arrive are random, as illustrated schematically in Fig. 3.5-1(a); the statistical properties of the photon arrivals are determined by the nature of the source emitting the photons.

EXAMPLE 3.5-1. ***Random Photon Arrivals.*** The random arrival of photons can be understood as follows. Consider a source with optical power $P = 1$ nW that emits light at a wavelength $\lambda_0 = 1\ \mu$m, so it delivers an *average* photon flux of $\Phi = 5$ photons/ns or 0.005 photons/ps. Since only integral numbers of photons may be detected, this signifies that if 10^5 time intervals are examined, each of duration $T = 1$ ps, then most of the intervals will register zero photons; about 500 of the intervals will register one photon; and very few of the intervals will register two or more photons. The result is a random sequence of discrete events.

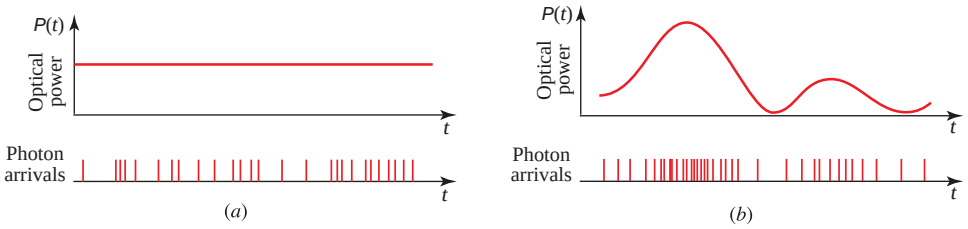


Figure 3.5-1 (a) Constant optical power and a sample function of the randomly arriving photons, the statistics of which are determined by the nature of the source. (b) Time-varying optical power and a sample function of the randomly arriving photons, the statistics of which are established both by the nature of the source and by the fluctuations of the optical power.

If the optical power $P(t)$ also varies with time, the mean density of photon detections tracks the time variation of $P(t)$, as schematically illustrated in Fig. 3.5-1(b). The variations of the mean photon flux $\Phi(t) = P(t)/h\bar{\nu}$ with time reflect the fact that the photons arrive at a greater rate when the optical power is large than when it is small. Such variations in power can arise from intrinsic intensity fluctuations associated with an optical source of a particular kind (e.g., an incandescent source) or from external manipulation of its mean optical power. These fluctuations are over-and-above those exhibited in Fig. 3.5-1(a) for a source of constant optical power, which continue to contribute to the randomness in the photon arrivals for time-varying light.

Photon Arrival Locations

The image of Max Planck portrayed in Fig. 3.4-2 illustrates analogous behavior in the spatial domain. The locations of the detected photons generally follow the classical intensity distribution of the image, exhibiting high photon densities where the intensity is large and low photon densities where it is small. However, there is considerable randomness (also referred to as graininess or spatial noise) in the image that arises from the fluctuations in photon-occurrence locations associated with the source emitting the photons. These fluctuations are most easily discerned when the mean photon-flux density is small, as is the case in Fig. 3.4-2. When the mean photon-flux density becomes large everywhere, as it is in the image of Max Planck on p. 61, the graininess disappears and the classical spatial intensity distribution of the image is recovered.

3.6 PHOTON-NUMBER STATISTICS

An understanding of the photon-number statistics of a source of light is useful for many applications, including low-light imaging. For a *coherent source* of light, such as an ideal laser, the emitted optical power is constant and the arriving photons can be represented as a sequence of independent random occurrences at a rate specified by the photon flux, which is proportional to the optical power. Independent photon arrivals lead to Poisson photon-number statistics. For a *partially coherent source* of light, on the other hand, the optical power fluctuations result in photon arrivals that no longer form a sequence of independent events, and the photon-number statistics can differ substantially from Poisson form. We begin by considering Poisson photon-number statistics and follow this with an examination of doubly stochastic Poisson photon-number statistics. The theory presented in this section is suitable only for classical light.

Poisson Photon-Number Statistics

Coherent light has constant optical power P. The corresponding mean photon flux $\Phi = P/h\bar{\nu}$ (photons/s) is also constant, but the actual photon registration times are random, as portrayed schematically in Fig. 3.6-1. Given a fixed time interval of duration T, called the **counting time** (or **counting window**), the random variable n, called the **photon number** (or **count number**), denotes the number of photons detected within that window.

Equation (3.4-6) specifies that the **mean photon number** (or **count mean**) is $\bar{n} = \Phi T = PT/h\bar{\nu}$. We now seek to establish the **photon-number distribution** (or **counting distribution**), $p(n)$ vs. n, i.e., the probability $p(0)$ of detecting zero photons, the probability $p(1)$ of detecting one photon, and so on, in a sequence of counting windows of duration T.

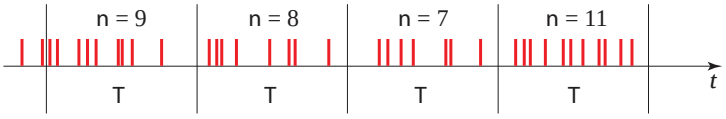


Figure 3.6-1 Random arrival of photons for a coherent light source of power P. Consecutive counting windows of duration T are indicated. Though the optical power is constant, the photon number n observed in each counting window is random.

An expression for the photon-number distribution $p(n)$ vs. n for coherent light can be obtained by assuming that the photon registrations are statistically independent. The result, which is derived below and is known as the **Poisson distribution**, takes the form

$$p(n) = \frac{\bar{n}^n \exp(-\bar{n})}{n!}, \qquad n = 0, 1, 2, \ldots . \tag{3.6-1}$$

Poisson Distribution

Equation (3.6-1) is plotted in Fig. 3.6-2, on both linear and semilogarithmic coordinates, for several values of the mean photon number $\bar{n}$. In addition to coherent light, the Poisson distribution also characterizes the photon statistics associated with a number of other sources of light, including multimode thermal light, as will be discussed in Chapter 4.

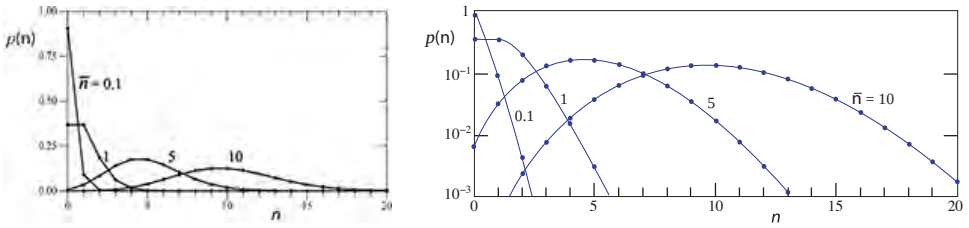
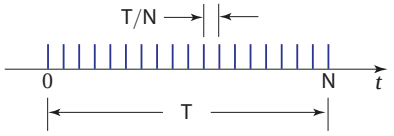


Figure 3.6-2 Poisson photon-number distribution $p(n)$ vs. number of photons n for four values of the mean photon number: $\overline{n} = 0.1$, 1.0, 5.0, and 10. The distribution is plotted on linear coordinates (left) and on semilogarithmic coordinates (right) The curves become progressively broader as $\overline{n}$ increases.

□ **Derivation of the Poisson Distribution.** Divide the counting time T displayed in Fig. 3.6-1 into a large number N of subintervals, each of sufficiently short duration T/N such that each subinterval carries one photon with probability $p = \overline{n}/N$ and zero photons with probability $1-p$. The probability of finding n independent photons in the N subintervals, like the flips of a biased coin, follows the binomial distribution:

$$p(n) = \frac{N!}{n!\,(N-n)!}\,p^{\,n}\,(1-p)^{N-n}$$

$$= \frac{N!}{n!\,(N-n)!}\left(\frac{\overline{n}}{N}\right)^{n}\left(1-\frac{\overline{n}}{N}\right)^{N-n}.$$



As the number of subintervals $N \to \infty$, we have $N!/(N-n)!\,N^{\,n} \to 1$ along with $(1-\overline{n}/N)^{-n} \to 1$ and $(1-\overline{n}/N)^{N} \to \exp(-\overline{n}\,)$, which lead to (3.6-1). ■

Photon-Number Mean and Variance

The **count mean (number mean)** and **count variance (number variance)** are statistics that characterize a discrete counting distribution. The count mean is defined as

$$\overline{n} = \sum_{n=0}^{\infty} n\,p(n), \tag{3.6-2}$$

while the count variance, which is the average of the squared deviation from the mean, is given by

$$\sigma_n^2 = \sum_{n=0}^{\infty} (n-\overline{n}\,)^2\,p(n). \tag{3.6-3}$$

The **count standard deviation** σ_n, defined as the square root of the count variance, is a measure of the width of the distribution. The quantities $p(n)$, $\overline{n}$, and σ_n are collectively called the **counting statistics**. Though the distribution $p(n)$ contains information beyond its mean and variance, these two parameters provide a rough outline of its form.

As derived below, inserting (3.6-1) into (3.6-2) and (3.6-3) confirms that the mean of the Poisson distribution is indeed $\overline{n}$ and that its variance is equal to its mean,

$$\boxed{\sigma_n^2 = \overline{n}\,.} \tag{3.6-4}$$

Mean and Variance
Poisson Distribution

Taking $\overline{n} = 100$ as an example, the standard deviation is $\sigma_n = 10$, which signifies that, on average, the observation of 100 photons is accompanied by an uncertainty of ± 10 photons.

□ **Normalization, Count Mean, and Count Variance of the Poisson Distribution.**

Normalization:
$$\sum_{n=0}^{\infty} p(n) = \sum_{n=0}^{\infty} \frac{\overline{n}^n e^{-\overline{n}}}{n!} = e^{-\overline{n}} \sum_{n=0}^{\infty} \frac{\overline{n}^n}{n!} = e^{-\overline{n}} \cdot e^{\overline{n}} = 1. \checkmark \tag{3.6-5}$$

Mean:
$$\sum_{n=0}^{\infty} np(n) = \sum_{n=0}^{\infty} n \frac{\overline{n}^n e^{-\overline{n}}}{n!} = e^{-\overline{n}} \cdot \overline{n} \sum_{n=1}^{\infty} \frac{\overline{n}^{n-1}}{(n-1)!} = e^{-\overline{n}} \cdot \overline{n} \cdot e^{\overline{n}} = \overline{n}. \checkmark \tag{3.6-6}$$

Variance:
$$\begin{aligned}
\sigma_n^2 &= \sum_{n=0}^{\infty} (n-\overline{n})^2 p(n) = \sum_{n=0}^{\infty} (n-\overline{n})^2 \frac{\overline{n}^n e^{-\overline{n}}}{n!} \\
&= \sum_{n=0}^{\infty} n^2 \frac{\overline{n}^n e^{-\overline{n}}}{n!} - 2\overline{n} \sum_{n=0}^{\infty} n \frac{\overline{n}^n e^{-\overline{n}}}{n!} + \overline{n}^2 \sum_{n=0}^{\infty} \frac{\overline{n}^n e^{-\overline{n}}}{n!} \\
&= \sum_{n=0}^{\infty} n^2 \frac{\overline{n}^n e^{-\overline{n}}}{n!} - \overline{n}^2 \sum_{n=0}^{\infty} \frac{\overline{n}^n e^{-\overline{n}}}{n!} \\
&= \sum_{n=1}^{\infty} n \frac{\overline{n}^n e^{-\overline{n}}}{(n-1)!} - \overline{n}^2 = \sum_{n=1}^{\infty} (n-1) \frac{\overline{n}^n e^{-\overline{n}}}{(n-1)!} + \sum_{n=1}^{\infty} \frac{\overline{n}^n e^{-\overline{n}}}{(n-1)!} - \overline{n}^2 \\
&= \overline{n}^2 e^{-\overline{n}} \sum_{n=2}^{\infty} \frac{\overline{n}^{n-2}}{(n-2)!} + \overline{n} e^{-\overline{n}} \sum_{n=1}^{\infty} \frac{\overline{n}^{n-1}}{(n-1)!} - \overline{n}^2 = \overline{n}^2 + \overline{n} - \overline{n}^2 = \overline{n}. \checkmark
\end{aligned} \tag{3.6-7}$$

■

Signal-to-Noise Ratio

Another counting statistic that useful for determining the performance of a photon-detection system is the count signal-to-noise ratio (SNR). Representing the signal by the mean $\overline{n}$, and the noise by the standard deviation σ_n, the count SNR is defined as

$$\text{SNR} = \frac{(\text{mean})^2}{\text{variance}} = \frac{\overline{n}^2}{\sigma_n^2}\,. \tag{3.6-8}$$

If the light obeys Poisson photon-number statistics, then (3.6-4) provides that $\sigma_n^2 = \overline{n}$, whereupon

$$\boxed{\text{SNR} = \overline{n}\,.} \tag{3.6-9}$$

Signal-to-Noise Ratio
Poisson Distribution

The Poisson signal-to-noise ratio increases linearly with the mean photon number. Although the SNR is often useful for measuring the randomness of a signal, applications that require a determination of the probability of error of a system generally require knowledge of the full probability distribution.

Doubly Stochastic Poisson Photon-Number Statistics

As discussed above, coherent light has constant intensity $I(\mathbf{r}, t)$, constant optical power P, and constant mean photon flux $\Phi = P/h\bar{\nu}$. The arriving photons behave as independent events with a Poisson photon-number distribution $p(n) = \bar{n}^n e^{-\bar{n}}/n!$, where the mean photon number $\bar{n} = \Phi T = PT/h\bar{\nu}$ is constant.

However, if the light is partially coherent because the intensity varies in time, then so too does the optical power [as portrayed in Fig. 3.5-1(b)], the mean photon flux, and the mean photon number $\bar{n}$. In that case, in accordance with (3.4-10) and (3.4-11), the mean photon number, which we denote as w rather than $\bar{n}$ for reasons that will become clear below, can be expressed as

$$w \equiv \bar{n} = \frac{1}{h\bar{\nu}} \int_0^T P(t)\, dt = \frac{1}{h\bar{\nu}} \int_0^T \int_A I(\mathbf{r}, t)\, dA\, dt. \tag{3.6-10}$$

The integrated intensity w, which has units of photon number and is thus dimensionless, therefore varies in time for partially coherent light.

Variations in the mean photon number arising from intensity fluctuations cause the photon-number distribution to depart from Poisson behavior, as we now demonstrate. If the fluctuations of w are described by a probability density function $p(w)$, the applicable photon-number distribution is obtained by averaging the Poisson distribution conditioned on w being constant, $p(n|w) = w^n e^{-w}/n!$, over the range of allowed values of w dictated by $p(w)$. It is now clear that we introduced the symbol w above so that we could co-opt the symbol n for the new photon number.

The resultant photon-number distribution therefore takes the form

$$p(n) = \int_0^\infty \frac{w^n e^{-w}}{n!}\, p(w)\, dw, \tag{3.6-11}$$

Poisson Transform (Mandel's Formula)

which is known as the **Poisson transform** of $p(w)$ and also as **Mandel's formula**. Equation (3.6-11) is also sometimes referred to as a **doubly stochastic photon-number distribution** by virtue of the fact that its randomness arises from two sources: 1) the random arrivals of the photons, which behave locally in Poisson fashion and are present even for sources of constant intensity; and 2) the integrated-intensity fluctuations associated with the time varying nature of the intensity. The sequence of random photon arrivals that underlies doubly stochastic Poisson photon-number statistics is known as a **doubly stochastic Poisson process (DSPP)**.

The photon-number mean and variance for the doubly stochastic photon-number distribution are obtained by using (3.6-2) and (3.6-3) in conjunction with (3.6-11); the results turn out to be

$$\bar{n} = \bar{w} \tag{3.6-12}$$

and

$$\sigma_n^2 = \bar{n} + \sigma_w^2, \tag{3.6-13}$$

respectively, where σ_w^2 represents the variance of w. The photon-number variance is seen to contain two contributions: 1) the basic Poisson contribution $\bar{n}$; and 2) a (positive) contribution arising from the intensity fluctuations. When the intensity is constant, $p(w)$ becomes a delta function, $\sigma_w^2 \to 0$, and $p(n)$ reduces to the Poisson distribution.

EXAMPLE 3.6-1. ***Photon-Number Distribution for an Exponentially Distributed Integrated Intensity.*** As an example of the use of the Poisson transform (3.6-11), we consider an exponentially distributed integrated-intensity probability density function:

$$p(w) = \begin{cases} \dfrac{1}{\overline{w}} \exp\left(-\dfrac{w}{\overline{w}}\right), & w \geqslant 0 \\ 0, & w < 0\,. \end{cases} \tag{3.6-14}$$

Equation (3.6-14) is appropriate for describing quasi-monochromatic light whose real and imaginary complex-field amplitude components are Gaussian, as well as independent and identically distributed. It is applicable for partially coherent light whose spectral width is sufficiently narrow that its coherence time τ_c is much larger than the counting time T. The associated photon-number distribution is determined by substituting (3.6-14) into (3.6-11) and evaluating the integral. The result turns out to be the geometric (Bose–Einstein) photon-counting distribution (4.2-8), which is considered in Sec. 4.8 in connection with thermal light.

3.7 RANDOM PARTITIONING OF PHOTON STREAMS

A photon stream is said to be partitioned when it is subjected to the removal of some of its photons. The process is called **random partitioning** when the removed photons are randomly diverted and **random deletion** when they are annihilated. There are numerous ways in which this can occur. Perhaps the simplest example of random partitioning is provided by an ideal lossless beamsplitter. Each photon incident on an input port of the device is randomly chosen to exit one or the other of the two output ports (Fig. 3.7-1). An example of random deletion is provided by the action of a photodetector. Each photon incident on the photosensitive material is chosen either to be absorbed and to create a photoelectron in the process, or to pass through the material and be lost.

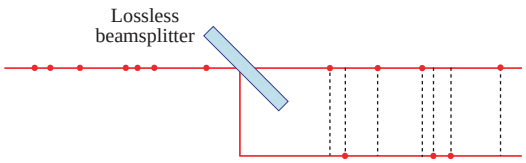


Figure 3.7-1 Random partitioning of a stream of photons by a beamsplitter.

The treatment presented here is restricted to situations in which the partitioning behaves in accordance with a sequence of independent **Bernoulli trials** (coin tosses), each associated with an incident photon. This is applicable when the photon stream impinges on only one of the input ports, as portrayed in Fig. 3.7-1. When photon streams enter both ports, they can interfere and violate the independent-trial assumption. The behavior of a single photon impinging on a lossless beamsplitter with intensity transmittance $\mathcal{T}$ and intensity reflectance $\mathcal{R} = 1 - \mathcal{T}$ was considered in Example 3.3-1. As illustrated in Fig. 3.3-1, it was ascertained that the probability of the photon being transmitted was equal to the transmittance of the beamsplitter $\mathcal{T}$, while the probability of it being reflected was equal to $1 - \mathcal{T}$.

We now consider a photon stream of mean photon flux Φ incident on the beamsplitter, so that the mean number of impinging photons in the time interval T is $\overline{n} = \Phi T$. The

mean number of photons transmitted and reflected is then $\mathcal{T}\overline{n}$ and $(1-\mathcal{T})\overline{n}$, respectively. We proceed to determine the photon-number statistics after partitioning. We begin by assuming that the incident stream consists of precisely n photons, so that the probability $p(m)$ that m photons are transmitted is the same as that of flipping a biased coin n times and obtaining m heads, where the probability of obtaining a head is $\mathcal{T}$. The probability $p(m)$ is then characterized by the binomial distribution

$$p(m) = \binom{n}{m} \mathcal{T}^m (1-\mathcal{T})^{n-m}, \qquad m = 0, 1, \ldots, n, \tag{3.7-1}$$

where $\binom{n}{m} = n!/m!\,(n-m)!$. By symmetry, the result for the reflected photons is the same, with $1-\mathcal{T}$ replacing $\mathcal{T}$.

The statistics of the binomial distribution dictate that the mean number of transmitted photons is

$$\overline{m} = \mathcal{T} n \tag{3.7-2}$$

and the photon-number variance is

$$\sigma_m^2 = \mathcal{T}(1-\mathcal{T})n = (1-\mathcal{T})\,\overline{m}. \tag{3.7-3}$$

The count signal-to-noise ratio specified in (3.6-8) is then SNR $= \overline{m}^{\,2}/\sigma_m^2 = \overline{m}/(1-\mathcal{T})$, which increases linearly with the mean number of transmitted photons $\overline{m}$. In the limit where the incident photon flux is large, the photons will therefore be partitioned between the transmitted and reflected beams in good agreement with $\mathcal{T}$ and $(1-\mathcal{T})$, respectively, as predicted by classical optics.

The calculation for an arbitrary photon-number distribution $p_0(n)$ at the input to the beamsplitter proceeds by recognizing that the number of photons n at the input is random rather than fixed. The photon-number probability distribution for the transmitted stream is therefore a weighted sum of binomial distributions, with the weighting established by the probability of n photons being present at the input. The photon-number distribution $p(m)$ at the output of the beamsplitter, for an input photon-number distribution $p_0(n)$, is therefore $p(m) = \sum_n p(m|n)\,p_0(n)$, where the observation of m photons conditioned on n having a particular value is the binomial distribution $p(m|n) = \binom{n}{m} \mathcal{T}^m (1-\mathcal{T})^{n-m}$. Finally, then, we arrive at the formula that specifies $p(m)$ in terms of $p_0(n)$ and $\mathcal{T}$:

$$\boxed{p(m) = \sum_{n=m}^{\infty} \binom{n}{m} \mathcal{T}^m (1-\mathcal{T})^{n-m}\, p_0(n).} \tag{3.7-4}$$

Photon-Number Distribution Under Random Partitioning

The same formula applies for the random deletion of photons.

When the photon-number distribution $p_0(n)$ at the input to the beamsplitter is Poisson, it is demonstrated below that the partitioned photon-number distribution $p(m)$ remains Poisson; however, its photon-number mean is reduced by the factor $\mathcal{T}$. Hence, the signal-to-noise ratio for a randomly partitioned Poisson stream is SNR = $\overline{n}\,\mathcal{T}$. Since $\mathcal{T} \leqslant 1$, random partitioning decreases the signal-to-noise ratio or, stated differently, introduces noise.

□ **Random Partitioning of a Poisson Photon Stream.**

(a) The photon-number distribution $p(m)$ for photons whose initial counting distribution $p_0(n)$ is Poisson retains its Poisson form under random partitioning, but with a reduced mean $\overline{m} = \overline{n}\,\mathcal{T}$:

$$p(m) = \sum_{n=m}^{\infty} \binom{n}{m} \mathcal{T}^m (1-\mathcal{T})^{n-m}\, p_0(n) = \sum_{n=m}^{\infty} \binom{n}{m} \mathcal{T}^m (1-\mathcal{T})^{n-m}\, \frac{\overline{n}^{\,n} e^{-\overline{n}}}{n!}$$

$$= \sum_{n=m}^{\infty} \frac{\overline{n}^{\,n} e^{-\overline{n}}}{m!(n-m)!} \mathcal{T}^m (1-\mathcal{T})^{n-m} = \frac{(\overline{n}\,\mathcal{T})^m}{m!}\, e^{-\overline{n}} \sum_{n=m}^{\infty} \frac{\overline{n}^{\,n-m}}{(n-m)!} (1-\mathcal{T})^{n-m}$$

$$= \frac{(\overline{n}\,\mathcal{T})^m}{m!}\, e^{-\overline{n}} \sum_{k=0}^{\infty} \frac{[\overline{n}\,(1-\mathcal{T})]^k}{k!} = \frac{(\overline{n}\,\mathcal{T})^m}{m!}\, e^{-\overline{n}}\, e^{\overline{n}(1-\mathcal{T})} = \frac{(\overline{n}\,\mathcal{T})^m}{m!}\, e^{-(\overline{n}\,\mathcal{T})}. \quad (3.7\text{-}5)$$

(b) The signal-to-noise ratio of a randomly partitioned Poisson photon stream is established by making use of the definition provided in (3.6-8):

$$\mathrm{SNR} = \frac{(\mathrm{mean})^2}{\mathrm{variance}} = \frac{\overline{m}^{\,2}}{\sigma_m^2} = \frac{(\overline{n}\,\mathcal{T})^2}{\overline{n}\,\mathcal{T}} = \overline{n}\,\mathcal{T}. \quad (3.7\text{-}6)$$

■

BIBLIOGRAPHY

Quantum Optics

C. C. Gerry and P. L. Knight, *Introductory Quantum Optics*, Cambridge University Press, 2nd ed. 2024.

C. Fabre, *Quantum Processes & Measurement: Theory & Experiment*, Cambridge University Press, 2023.

P. Meystre, *Quantum Optics: Taming the Quantum*, Springer, 2021.

H.-A. Bachor and T. C. Ralph, *A Guide to Experiments in Quantum Optics*, Wiley–VCH, 3rd ed. 2019.

B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, 3rd ed. 2019, Chaps. 11, 13.

Z.-Y. J. Ou, *Quantum Optics for Experimentalists*, World Scientific, 2017.

G. S. Agarwal, *Quantum Optics*, Cambridge University Press, 2012.

G. Grynberg, A. Aspect, and C. Fabre, *Introduction to Quantum Optics: From the Semi-Classical Approach to Quantized Light*, Cambridge University Press, 2010.

D. F. Walls and G. J. Milburn, *Quantum Optics*, Springer, 2nd ed. 2008.

J. C. Garrison and R. Y. Chiao, *Quantum Optics*, Oxford University Press, 2008.

P. Meystre and M. Sargent III, *Elements of Quantum Optics*, Springer, 4th ed. 2007.

M. Fox, *Quantum Optics: An Introduction*, Oxford University Press, 2006.

W. Vogel and D.-G. Welsch, *Quantum Optics*, Wiley–VCH, 3rd ed. 2006.

W. P. Schleich, *Quantum Optics in Phase Space*, Wiley–VCH, 2001.

R. Loudon, *The Quantum Theory of Light*, Oxford University Press, 3rd ed. 2000.

M. O. Scully and M. S. Zubairy, *Quantum Optics*, Cambridge University Press, paperback ed. 1997.

W. H. Louisell, *Quantum Statistical Properties of Radiation*, Wiley, 1973; Wiley–VCH, reprinted 1990.

J. R. Klauder and E. C. G. Sudarshan, *Fundamentals of Quantum Optics*, Benjamin, 1968; Dover, reissued 2006.

Quantum Coherence, Photon Statistics, Two-Photon Light, and Quantum Imaging

R. de J. León-Montiel, M. A. Quiroz-Juarez, O. Magana-Loaiza, and J. Torres, Quantum Light for Imaging, Sensing and Spectroscopy, *Frontiers in Physics*, DOI:10.3389/978-2-83250-394-2, 2022.

D. S. Simon, G. Jaeger, A. V. Sergienko, *Quantum Metrology, Imaging, and Communication*, Springer, 2017.

B. R. Masters, Satyendra Nath Bose and Bose–Einstein Statistics, *Optics & Photonics News*, vol. 24, no. 4, pp. 40–47, 2013.

M. C. Teich, B. E. A. Saleh, F. N. C. Wong, and J. H. Shapiro, Variations on the Theme of Quantum Optical Coherence Tomography: A Review, *Quantum Information Processing*, vol. 11, pp. 903–923, 2012.

R. J. Glauber, Nobel Lecture: One Hundred Years of Light Quanta, *Reviews of Modern Physics*, vol. 78, pp. 1267–1278, 2006.

B. E. A. Saleh, M. C. Teich, and A. V. Sergienko, Wolf Equations for Two-Photon Light, *Physical Review Letters*, vol. 94, 223601, 2005.

M. B. Nasr, B. E. A. Saleh, A. V. Sergienko, and M. C. Teich, Dispersion-Cancelled and Dispersion-Sensitive Quantum Optical Coherence Tomography, *Optics Express*, vol. 12, pp. 1353–1362, 2004.

J. S. Bell (with an introduction by A. Aspect), *Speakable and Unspeakable in Quantum Mechanics*, Cambridge University Press, 2nd ed. 2004.

A. F. Abouraddy, B. E. A. Saleh, A. V. Sergienko, and M. C. Teich, Entangled-Photon Fourier Optics, *Journal of the Optical Society of America B*, vol. 19, pp. 1174–1184, 2002.

J. Peřina, ed., *Coherence and Statistics of Photons and Atoms*, Wiley, 2001.

A. F. Abouraddy, B. E. A. Saleh, A. V. Sergienko, and M. C. Teich, Role of Entanglement in Two-Photon Imaging, *Physical Review Letters*, vol. 87, 123602, 2001.

L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics*, Cambridge University Press, 1995.

M. C. Teich and B. E. A. Saleh, Squeezed and Antibunched Light, *Physics Today*, vol. 43, no. 6, pp. 26–34, 1990 (Erratum: vol. 43, no. 11, pp. 123–124, 1990).

M. C. Teich and B. E. A. Saleh, Squeezed States of Light, *Quantum Optics: Journal of the European Physical Society B*, vol. 1, pp. 153–191, 1989.

R. A. Campos, B. E. A. Saleh, and M. C. Teich, Quantum-Mechanical Lossless Beam Splitter: SU(2) Symmetry and Photon Statistics, *Physical Review A*, vol. 40, pp. 1371–1384, 1989.

M. C. Teich and B. E. A. Saleh, Photon Bunching and Antibunching, in E. Wolf, ed., *Progress in Optics*, North-Holland, vol. 26, pp. 1–104, 1988.

J. Peřina, *Coherence of Light*, Reidel, 2nd ed. 1985.

R. P. Feynman (with an introduction by A. Zee), *QED: The Strange Theory of Light and Matter*, Princeton University Press, 1985, reissued 2014.

M. C. Teich, Role of the Doubly Stochastic Neyman Type-A and Thomas Counting Distributions in Photon Detection, *Applied Optics*, vol. 20, pp. 2457–2467, 1981.

D. N. Klyshko, *Photons and Nonlinear Optics*, Nauka (Moscow), 1980 [Translation: Gordon and Breach, New York, 1988].

E. Wolf, Einstein's Researches on the Nature of Light, *Optics News*, vol. 5, no. 1, pp. 24–39, 1979.

B. E. A. Saleh, *Photoelectron Statistics*, Springer, 1978.

L. Mandel and E. Wolf, Coherence Properties of Optical Fields, *Reviews of Modern Physics*, vol. 37, pp. 231–287, 1965.

Collections of Selected Papers

R. J. Glauber, *Quantum Theory of Optical Coherence: Selected Papers and Lectures*, Wiley–VCH, 2007.

G. S. Agarwal, ed., *Selected Papers on Fundamentals of Quantum Optics*, SPIE Optical Engineering Press (Milestone Series Volume 103), 1995.

J. Peřina, ed., *Selected Papers on Photon Statistics and Coherence in Nonlinear Optics*, SPIE Optical Engineering Press (Milestone Series Volume 39), 1991.

L. Mandel and E. Wolf, eds., *Selected Papers on Coherence and Fluctuations of Light (1850–1966)*, SPIE Optical Engineering Press (Milestone Series Volume 19), 1990.

L. Mandel and E. Wolf, eds., *Selected Papers on Coherence and Fluctuations of Light*, Volumes 1 and 2, Dover, 1970.

Seminal Publications

R. E. Slusher, L. W. Hollberg, B. Yurke, J. C. Mertz, and J. F. Valley, Observation of Squeezed States Generated by Four-Wave Mixing in an Optical Cavity, *Physical Review Letters*, vol. 55, pp. 2409–2412, 1985.

M. C. Teich and B. E. A. Saleh, Observation of Sub-Poisson Franck–Hertz Light at 253.7 nm, *Journal of the Optical Society of America B*, vol. 2, pp. 275–282, 1985.

R. Short and L. Mandel, Observation of Sub-Poissonian Photon Statistics, *Physical Review Letters*, vol. 51, pp. 384–387, 1983.

A. Aspect, P. Grangier, and G. Roger, Experimental Tests of Realistic Local Theories via Bell's Theorem, *Physical Review Letters*, vol. 47, pp. 460-463, 1981.

H. J. Kimble, M. Dagenais, and L. Mandel, Photon Antibunching in Resonance Fluorescence, *Physical Review Letters*, vol. 39, pp. 691–695, 1977.

S. Weinberg, Light as a Fundamental Particle, *Physics Today*, vol. 28, no. 6, pp. 32–37, 1975.

D. C. Burnham and D. L. Weinberg, Observation of Simultaneity in Parametric Production of Optical Photon Pairs, *Physical Review Letters*, vol. 25, pp. 84–87, 1970.

D. Magde and H. Mahr, Study in Ammonium Dihydrogen Phosphate of Spontaneous Parametric Interaction Tunable from 4400 to 16000 Å, *Physical Review Letters*, vol. 18, pp. 905–907, 1967.

S. E. Harris, M. K. Oshman, and R. L. Byer, Observation of Tunable Optical Parametric Fluorescence, *Physical Review Letters*, vol. 18, pp. 732–734, 1967.

J. S. Bell, On the Einstein Podolsky Rosen Paradox, *Physics Physique Fizika* (Long Island City), vol. 1, no. 3, pp. 195–200, 1964.

R. J. Glauber, Coherent and Incoherent States of the Radiation Field, *Physical Review*, vol. 131, pp. 2766–2788, 1963.

R. J. Glauber, The Quantum Theory of Optical Coherence, *Physical Review*, vol. 130, pp. 2529–2539, 1963.

L. Mandel, Fluctuations of Light Beams, in E. Wolf, ed., *Progress in Optics*, North-Holland, vol. 2, pp. 181–248, 1963.

A. Einstein, B. Podolsky, and N. Rosen, Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?, *Physical Review*, vol. 47, pp. 777–780, 1935.

E. Schrödinger, Die gegenwärtige Situation in der Quantenmechanik, *Die Naturwissenschaften*, vol. 23, pp. 807–812, 823–828, 844–849; 1935 [Translation: The Present Situation in Quantum Mechanics: A Translation of Schrödinger's 'Cat Paradox' Paper, *Proceedings of the American Philosophical Society*, vol. 124, pp. 323–338, 1980].

S. N. Bose, Plancks Gesetz und Lichtquantenhypothese (Planck's Law and the Light-Quantum Hypothesis), *Zeitschrift für Physik*, vol. 26, pp. 178-181, 1924.

CHAPTER

4

THERMAL LIGHT

4.1 TEMPERATURE AND EQUIPARTITION OF ENERGY 85

4.2 OCCUPATION OF ENERGY LEVELS 89

4.3 INTERACTIONS OF PHOTONS WITH ATOMS 96

4.4 SPONTANEOUS EMISSION 99

4.5 ABSORPTION AND STIMULATED EMISSION 103

4.6 LINE BROADENING 108

4.7 BLACKBODY RADIATION 111

4.8 THERMAL RADIATION 116

In 1848, **Lord Kelvin (William Thompson) (1824–1907)** proposed a temperature scale whose increments matched those of the Celsius scale but whose zero corresponded to the temperature at which all internal motion in a material is minimized.

In 1868, **Ludwig Boltzmann (1844–1906)**, in the course of studying the statistical mechanics of gases in thermal equilibrium, conceived of the distribution that bears his name. He also derived the Stefan–Boltzmann law, which had been formulated previously by Josef Stefan.

LED Lighting: Devices and Colorimetry. Malvin Carl Teich.
Google Books. Published 2024.

Before LED lighting came into widespread use in the early 2000s, the principal source of artificial illumination was thermal light. In particular, the electric incandescent filament lamp became the workhorse of artificial lighting in 1879, shortly after its invention by Thomas Edison and his British rival, Joseph Swan. An unfortunate limitation associated with the incandescent lamp, however, is that only about 5% of the power it radiates is in the visible region; roughly 95% is radiated in the infrared and is lost as heat. Nevertheless, by virtue of its simple construction, convenience, and low cost, incandescent lighting maintained its primacy until the arrival of LED lighting, and it still serves as a benchmark because of its excellent color rendering quality.

The fundamental principles underlying the generation of thermal light are presented in this chapter. We begin with a basic introduction to the concepts of temperature, thermal equilibrium, and the equipartion of energy (Sec. 4.1), and consider how thermal excitations cause the atoms of matter to constantly undergo upward and downward transitions among their allowed energy levels via the absorption and emission of photons (Sec. 4.2). A photon interacts with an atom if its energy matches the energy difference between two atomic levels (Sec. 4.3), in accordance with the rules of quantum mechanics.

If the atom is initially in the lower energy level, the photon may impart its energy to the atom and raise it to the higher level via a process called **absorption**. Or, if the atom is initially in the higher energy level, the photon may stimulate the atom to undergo a transition to the lower level and emit a second photon of the same energy via a process known as **stimulated emission**. An atom in the higher energy level can also transition to the lower level in the absence of an initiating photon via a process called **spontaneous emission** (Sec. 4.4). The relationship among these three processes was first established by Einstein (p. 61) in 1917 (Sec. 4.5). Spontaneous emission, endowed with a particular lineshape function (Sec. 4.6), is responsible for the operation of light-emitting diodes (Chapter 6).

The interaction of many photons with many atoms, under conditions of thermal equilibrium and steady state, can take place in an object known as a blackbody, a concept introduced by Kirchhoff in 1860. All blackbodies with temperatures greater than absolute zero emit a universal form of radiation called blackbody radiation, whose spectrum obeys the iconic radiation law introduced by Planck in 1900 (Sec. 4.7). The peak frequency of the Planck spectrum shifts toward higher frequencies (shorter wavelengths) as the temperature of the blackbody increases, by virtue of the increased population of higher atomic energy levels. The designation **thermal radiation** (thermal light) is an umbrella term that encompasses **blackbody radiation**, along with its closely related cousin **graybody radiation**, which includes incandescent light (Sec. 4.8).

Processes other than thermal ones can also result in the emission of light. These include laser action, Čerenkov radiation (emitted by charged particles traveling faster than the speed of light in a medium), Bremsstrahlung (emitted by the deceleration of charged particles as they penetrate matter), and luminescence radiation. Of the many forms of luminescence radiation, photoluminescence is of principal importance in the context of LED lighting. As will become clear in Sec. 10.2, photoluminescence is a process whereby a molecular system excited to a higher energy level by the absorption of a photon decays to a lower level via the emission of a lower-frequency photon in conjunction with a nonradiative transition.

4.1 TEMPERATURE AND EQUIPARTITION OF ENERGY

Our points of departure are temperature (and the scales commonly used to measure it) and thermal equilibrium. We then proceed to provide a brief introduction to the ideal gas law and the kinetic theory of gases, which allow temperature to be related to the internal

energy of a system from a microscopic perspective. We conclude with a discussion of the equipartition theorem of statistical mechanics, which can also be used to define temperature and derive the ideal gas law.

Temperature Scales and Thermal Equilibrium

Temperature is measured with the help of a thermometer, an instrument that registers a change in some physical property of a material in response to a change in its temperature. Thermometers are often implemented by making use of the observation that the volumes of many materials increase with increasing temperature, the old-fashioned household mercury thermometer being a familiar example.

Celsius and Fahrenheit Scales. Temperature units are defined by: 1) selecting two convenient temperatures, such as the freezing and boiling points of water; 2) ascribing arbitrary temperature values to those two points; and 3) constructing a scale between them with equal divisions. For the Celsius (Fahrenheit) scale, the freezing and boiling points of water are set at $0\ ^\circ$C ($32\ ^\circ$F) and $100\ ^\circ$C ($212\ ^\circ$F), respectively, and the scale is endowed with 100 (180) divisions.

Kelvin Scale. As will become clear in the sequel, the temperature of a substance is related to the microscopic motion of its constituent molecules. In 1848, Lord Kelvin (p. 84) constructed a temperature scale endowed with divisions that correspond to the Celsius scale but in which $T = 0$, called **absolute zero**, represents the point at which all of the internal motion in a material is at a minimum and the object can get no colder. Conducting a set of experiments with a thermometer calibrated on the Celsius scale, he established that this temperature was $-273.15\ ^\circ$C. The scale he proposed is called the **Kelvin temperature scale** and the temperature is measured in kelvins. One kelvin represents the same temperature difference as one Celsius degree, but kelvin temperatures are measured from absolute zero rather than from the freezing point of water. Figure 4.1-1 displays the numerical values for a collection of well-known temperature markers using the three temperature scales: Fahrenheit (°F), Celsius (°C), and Kelvin (K).

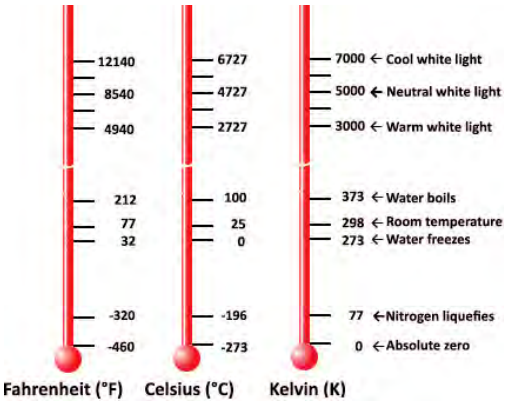


Figure 4.1-1 Numerical values for well-known temperature markers using three temperature scales: Fahrenheit, Celsius, and Kelvin. Absolute zero (0 K) corresponds to $-273.15\ ^\circ$C and $-459.67\ ^\circ$F. Matter cannot be cooled below absolute zero. The upper three entries are representative temperatures for thermal sources that generate visible light for illumination (via a tungsten-filament incandescent lamp, for example), annotated with the standard terminology used to designate those temperatures in the context of LED lighting, as will be explained in Secs. 9.7 and 9.8.

Example 4.1-1 reports the temperatures of selected objects on earth and in the cosmos in kelvins.

EXAMPLE 4.1-1. ***Temperatures (K) of Selected Objects on Earth and in the Cosmos.***

On earth:

- Experimental ^{23}Na Bose–Einstein condensate: $T \approx 4.5 \times 10^{-10}$ K.
- Helium liquefies: $T \approx 4$ K.
- Nitrogen liquefies: $T \approx 77$ K.
- Earth: $T \approx 300$ K.
- Human body temperature: $T \approx 310$ K.
- Lead melts: $T \approx 620$ K.
- Hot lava: $T \approx 1300$ K.
- Laser-induced burning plasma in hohlraum: $T \approx 6 \times 10^{7}$ K.
- Experimental quark–gluon plasma: $T \approx 6 \times 10^{12}$ K.

In the cosmos:

- Boomerang nebula: $T \approx 1$ K.
- Cosmic microwave background radiation: $T \approx 2.725$ K.
- Gaseous matter between stars and galaxies: $T \approx 3$ K.
- Surface of Uranus: $T \approx 60$ K.
- Brown dwarf in Lyra constellation: $T \approx 300$ K.
- Surface of the sun: $T \approx 5800$ K.
- Eta Carinae stellar system: $T \approx 3.8 \times 10^{4}$ K.
- White dwarf in Red Spider nebula: $T \approx 3 \times 10^{5}$ K.
- Newly formed neutron star: $T \approx 10^{11}$ K.

Thermal Equilibrium. Two systems are said to be in **thermal equilibrium** if there is no net flow of heat between them when they are brought into thermal contact. The motions of the constituent molecules in both objects are then in steady state and their fluctuations are, on average, independent of time. In accordance with the zeroth law of thermodynamics, two systems individually in thermal equilibrium with a third system are also in thermal equilibrium with each other. Since the third system can be a thermometer, two systems at the same temperature are in thermal equilibrium with each other. The time required to reach thermal equilibrium is known as the *thermal relaxation time*. A system is in **thermal quasi-equilibrium** when it is considered over a period of time that is short in comparison with the thermal relaxation time.

Temperature and Internal Energy

Our initial discussion of temperature was focused on the mechanics of how it is measured. We now proceed to link temperature to the average internal energy of the system in which it is measured. This connection is forged by making use of the ideal gas law, the kinetic theory of gases, and Newton's laws of motion.

Ideal Gas Law. The ideal gas law, written as

$$PV = NkT, \tag{4.1-1}$$

is an empirical relationship that relates several macroscopic thermodynamic variables in a gas: the pressure P, volume V, number of particles N, and temperature T measured in kelvins. The quantity $k \approx 1.38 \times 10^{-23}$ J/K is **Boltzmann's constant** (a photo of Boltzmann stands on p. 84). At $T = 300$ K (room temperature), $kT = 0.026$ eV = 4.14 zJ = 209 cm^{-1} (Fig. 3.2-1). Equation (4.1-1) offers a fine description for many real gases (particularly monatomic ones), as long as their densities are sufficiently low that the constituent molecules interact little, and it is widely used in practice. It is apparent from this equation that at fixed volume, a perfect gas exhibits $P = 0$ at $T = 0$. However, although this law relates macroscopic variables such as pressure and temperature, it provides no insight into the molecular underpinnings of the relation. To establish those connections, we appeal to the kinetic theory of gases.

Kinetic Theory of Gases. Kinetic theory provides a bridge that links pressure, a macroscopic thermodynamic property characterizing a large collection of particles in a container, to the internal kinetic energy the constituent particles, a microscopic property. We make use of a simple version of kinetic theory in which the individual particles constantly undergo random elastic collisions with the walls of the container, as depicted in Fig. 4.1-2. We then make use of Newton's laws of motion and the ideal gas law to relate temperature to internal energy. Interactions among the constituent particles are ignored and steady-state conditions are assumed to prevail. Kinetic theory is an elementary form of statistical mechanics, which is based in atomic physics, and provides the physical underpinnings of thermodynamics.

Newton's laws provide the explanation of how an individual gas particle, in colliding with a wall of the container, exerts a force on the wall as it rebounds. The large number of gas particles striking the walls in this manner exert a collective force, and the force per unit area represents the pressure of the gas. The ideal gas law in turn provides that the pressure of the gas is proportional to its temperature. In short, kinetic theory reveals that the macroscopic Kelvin temperature is proportional to the microscopic internal average translational kinetic energy of a particle in an ideal monatomic gas, whatever the pressure and volume of the container.

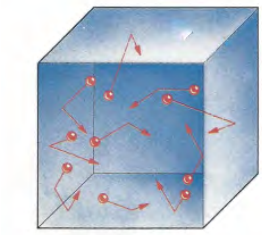


Figure 4.1-2 Individual gas particles bouncing off the walls of a container exert forces that are the origin of gas pressure.

The expression that relates these quantities is

$$\tfrac{1}{2}mv_{\text{RMS}}^2 = \tfrac{3}{2}kT, \tag{4.1-2}$$

where m is the mass of the particle and v_{RMS} is its root-mean-square velocity, as defined in (A.2-1) of Appendix A. The Boltzmann constant k serves as the constant of proportionality between the temperature of the system and the internal kinetic energy.

□ **Relation Between System Temperature and Internal Kinetic Energy in an Ideal Gas.** Consider a collection of N identical particles in a cubical container of side length L and volume L^3, and focus on the motion of a single particle of mass m and velocity v_x traveling along the x axis in a direction normal to a wall of the container. Since a collision with the wall is elastic, the speed of the particle is the same both before and after the collision, and the change of momentum of the particle resulting from an encounter is $\Delta p = +mv_x - (-mv_x) = 2mv_x$. Because the particle rebounds from the wall once during its round-trip travel time along the x axis, which is $\Delta t = 2L/v_x$, the force exerted on the wall by the collision is $\mathcal{F} = \Delta p/\Delta t = mv_x^2/L$. However, the N particles actually move about randomly in three dimensions, so only $N/3$ of them on average strike the particular wall. Hence, the total force exerted by the N particles on the wall is $\mathcal{F} = (N/3)(mv_{\text{RMS}}^2/L)$, where v_{RMS} represents the root-mean-square speed [see (A.2-1) of Appendix A] associated with the Maxwell velocity distribution (also called the Maxwell–Boltzmann distribution). We conclude that the force per unit area, or pressure, is given by $P = (N/3)(mv_{\text{RMS}}^2/L^3) = \frac{2}{3}N(\frac{1}{2}mv_{\text{RMS}}^2/V)$. Rewriting this as $PV = \frac{2}{3}N(\frac{1}{2}mv_{\text{RMS}}^2)$, and observing that the ideal gas law $PV = NkT$ expresses PV in terms of temperature, we arrive at $\frac{1}{2}mv_{\text{RMS}}^2 = \frac{3}{2}kT$, as set forth in (4.1-2). Temperature is a property of a *collection* of particles in thermal equilibrium; it is not to be associated with an individual particle. ■

Equipartition of Energy

The equipartition theorem of classical statistical mechanics specifies that in thermal equilibrium an equal share of energy is to be associated with each and every form of energy in the system that can be expressed as a quadratic function of a coordinate or velocity component. Each such form of energy is called a **degree-of-freedom (DOF)**. A particle moving through free space, for example, has three DOFs since its kinetic energy is expressed as a sum of three quadratic functions of velocity of the form $mv_x^2/2$. According to the equipartition theorem, each component of rotational kinetic energy, vibrational kinetic energy, and elastic potential energy is eligible for its share since the associated energies assume the quadratic forms $\mathcal{I}\omega_x^2/2$, $mv_x^2/2$, and $\kappa x^2/2$, respectively, where $\mathcal{I}$ is the moment of inertia, ω_x is a component of the angular velocity, and κ is the molecular spring constant.

The equipartition theorem specifies that at temperature T, the average thermal energy ascribable to each such quadratic degree-of-freedom is $\frac{1}{2}kT$. Hence, an ideal monatomic gas, with three translational DOFs, has total thermal energy $\frac{3}{2}kT$, as expressed in (4.1-2). Similarly, a harmonic mode has two DOFs, one for its vibrational kinetic energy and the other for its potential energy, for a total of kT. The same is true for a classical harmonic electromagnetic mode. In a crystalline solid, each constituent atom can vibrate in three orthogonal dimensions with respect to its neighbors in the lattice, which gives rise to six DOFs per atom, three from kinetic and three from potential energy.

In general, tabulating the DOFs that contribute to the thermal energy in a classical system can be tricky, although equipartition-theorem violations at lower temperatures are reasonably well-understood. At sufficiently high temperatures, strong interparticle interactions serve to excite all available DOFs, but this is frequently not the case at moderate temperatures, where the unexcited DOFs do not contribute to the overall thermal energy and are said to be *frozen out*. The equipartition theorem is not applicable for static contributions, such as the energies stored in chemical bonds, nor should it be used for broken bonds or phase transitions. The equipartition principle fails for quantum systems where the energy levels do not form a smooth continuum. The quintessential example of this breakdown is the failure of the Rayleigh–Jeans law to describe the spectrum of blackbody radiation, as discussed in Sec. 4.7.

4.2 OCCUPATION OF ENERGY LEVELS

As mandated by the laws of statistical physics, an object such as an individual atom drawn from a collection of identical objects in thermal equilibrium, continuously undergoes random transitions among its various energy levels. The principal determinant of the magnitude of the energy-level occupancy fluctuations, and of the average behavior, is the internal energy of the system as characterized by its temperature.

Energy Levels

The atoms of matter may exist in relative isolation, as in the case of a dilute atomic gas, or they may interact strongly with neighboring atoms to form molecules, liquids, and solids. The energy levels of simple forms of matter are determined by solving the **Schrödinger equation** of quantum mechanics, subject to a potential energy $V(\mathbf{r}, t)$ that characterizes the environment. Atomic and ionic energy levels, for example, are established by determining the potential energies of the electrons in the presence of the atomic nucleus and all other electrons, along with the potential energies associated with the orbital and spin angular momenta, which are usually small in comparison with those involving charges. Molecules, liquids, and solids obey more complex versions of the Schrödinger equation, in which the potential energy contains terms that accommo-

date interactions among the constituent atoms, as well as contributions from externally applied fields.

The energy levels can be discrete (as for an atom), or continuous (as for a free particle such as an electron), or they can comprise sets of densely packed discrete levels called bands (as discussed in Sec. 5.1 for a semiconductor). The presence of thermal excitations, or of an external field such as light illuminating the material, can induce the system to move from one of its energy levels to another. These interactions provide the means via which the system can exchange energy with the outside world.

Boltzmann Distribution

Consider a system comprising a collection of distinguishable objects (such as atoms) that form a dilute gas, where each object is in one of its allowed discrete energy states, $E_1, E_2, \ldots, E_m, \ldots$. If the system is in thermal equilibrium at temperature T, the probability $P(E_m)$ that an object is in energy level E_m is characterized by the **Boltzmann distribution**

$$P(E_m) \propto \exp\left(-\frac{E_m}{kT}\right), \qquad m = 1, 2, 3, \ldots, \tag{4.2-1}$$

Boltzmann Distribution

which is parameterized by the energy kT, where again k is Boltzmann's constant. The coefficient of proportionality in (4.2-1) is determined by imposing the normalization condition $\sum_m P(E_m) = 1$.

The occupation probability $P(E_m)$ vs. E_m specified in (4.2-1) is an exponentially decreasing function of E_m, as displayed in Fig. 4.2-1.

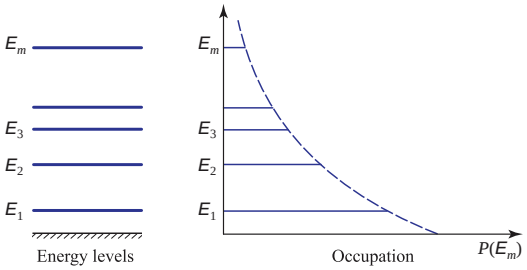


Figure 4.2-1 The Boltzmann distribution $P(E_m)$ (plotted along the *abscissa*) specifies the probability that the energy level E_m of an entity with an arbitrary collection of discrete energy levels (plotted along the *ordinate*), is occupied. $P(E_m)$ is an exponentially decreasing function of E_m.

□ **Form of the Boltzmann Distribution.** An understanding of the form assumed by the Boltzmann distribution can be attained by considering a system of many identical entities that share a fixed total energy E. The entities are isolated from their surroundings but are in thermal equilibrium, exchanging energy among themselves via a bath at temperature T. The divisions of energy are taken to be distinguishable if they involve different energy states, and all possible divisions of the total energy are assumed to occur with equal probability. If one of the entities takes a large share of the total energy, less is available for the remaining entities so there are fewer possible divisions. Consequently, large energies are less probable than small ones. A quantitative description is provided by considering two entities: The probability of finding one with energy E_1 and the other with energy E_2 is the product $P(E_1)P(E_2)$ since they are independent. However, if the sum of the energies of the two entities is fixed at the value $E_1 + E_2$, then $P(E_1)P(E_2)$ must be a function of $(E_1 + E_2)$, which uniquely specifies that the probability takes the form of an exponential function. The result for multiple energy levels follows by induction. The energies are measured in units of the equipartition energy kT. ■

Now consider the Boltzmann distribution in the context of a large number of atoms N. If N_m is the number of atoms occupying energy level E_m, the fraction $N_m/N \approx P(E_m)$. Hence, if N_1 atoms occupy level 1 and N_2 atoms occupy a higher level 2, in thermal equilibrium the population ratio is, on average,

$$\frac{N_2}{N_1} = \exp\left(-\frac{E_2 - E_1}{kT}\right). \tag{4.2-2}$$

This ratio depends on the temperature T. At $T = 0$ K, we have $N_2/N_1 = 0$ and all atoms are in the lowest energy level (ground state). As the temperature increases, the populations of the higher energy levels grow, but the average population of a given energy level always remains greater than that of a higher-lying level. This condition need not hold under non-equilibrium conditions, however, where a higher energy level can have a greater average population than a lower energy level. This latter condition is known as a *population inversion* and is the basis of laser action.

It has been assumed in the foregoing that there is a unique way in which an atom can find itself in one of its energy levels. It is sometimes the case, however, that two or more states (e.g., different states of angular momentum) correspond to the same energy. To account for such degenerate states, (4.2-2) can be written in the more general form

$$\frac{N_2}{N_1} = \frac{g_2}{g_1} \exp\left(-\frac{E_2 - E_1}{kT}\right), \tag{4.2-3}$$

where the so-called **degeneracy factors** g_2 and g_1 represent the numbers of states corresponding to the energy levels E_2 and E_1, respectively.

Fermions and Bosons. Fundamental particles in physics are divided into two broad classes of indistinguishable particles: **Fermions**, such as electrons, protons, neutrons, and other material particles, are endowed with spin that is a half-integer multiple of $\hbar$. Fermions obey **Fermi–Dirac statistics**. Fermions with overlapping wavefunctions, such as electrons in a multielectron atom or in a semiconductor material, are subject to the **Pauli exclusion principle**, which asserts that no two identical particles can simultaneously be in the same state. In contrast, **bosons**, such as photons and other force-carrier particles have a spin that is an integer multiple of $\hbar$, as do quasiparticles such as plasmons, polaritons, and phonons. Bosons obey **Bose–Einstein statistics** and are not subject to the Pauli exclusion principle

Fermi–Dirac Statistics. The probability of occupancy of a state of energy E for a collection of fermions in thermal equilibrium is represented by the **Fermi–Dirac distribution** (or **Fermi function**),

$$\boxed{f_{\text{FD}}(E) = \frac{1}{\exp\left[(E - E_f)/kT\right] + 1},} \tag{4.2-4}$$

Fermi–Dirac Statistics

which is illustrated in Fig. 4.2-2. The occupancy probability decreases monotonically with increasing E, falling to a value of $1/2$ at the **Fermi energy** $E = E_f$. As a result of the Pauli exclusion principle, the Fermi–Dirac distribution represents a sequence of probabilities, each with a value that lies between 0 and 1 for every value of E. The condition $f_{\text{FD}}(E) = 1$ indicates that a state is definitely occupied whereas the condition $f_{\text{FD}}(E) = 0$ indicates that it is definitely unoccupied.

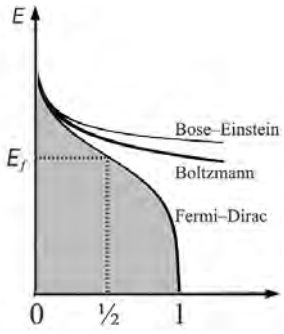


Figure 4.2-2 The Fermi–Dirac distribution $f_{\text{FD}}(E)$ (plotted on the *abscissa*) represents the probability of occupancy of a state of energy E (displayed on the *ordinate*). This distribution is applicable for systems containing particles with overlapping wavefunctions in which the Pauli exclusion principle is operative. The Bose–Einstein distribution $f_{\text{BE}}(E)$, which behaves very differently, is plotted in the same way. Both distributions may be approximated by the Boltzmann probability distribution $P(E_m)$ in the domain $E \gg E_f$ and $E \gg kT$, where the probability of occupancy is low.

For $E \gg E_f$ and $E \gg kT$, the occupancy probability is sufficiently small that the issue of indistinguishability is not relevant and the Fermi–Dirac distribution (4.2-4) reduces to the Boltzmann probability distribution,

$$P(E) \propto \exp(-E/kT). \tag{4.2-5}$$

The Boltzmann approximation is generally applicable for valence electrons in the outer subshells of atoms and ions so that the populations of optically active electrons are essentially governed by it. The Fermi function is considered further in Chapter 5 in connection with semiconductors.

Bose–Einstein Statistics. The probability of occupancy of a state of energy E for bosons in thermal equilibrium is represented by the **Bose–Einstein distribution**,

$$f_{\text{BE}}(E) = \frac{1}{\exp\left[(E - E_\mu)/kT\right] - 1}, \tag{4.2-6}$$

Bose–Einstein Statistics

where the chemical potential $E_\mu = 0$ for photons since their number is not conserved. This distribution is plotted in Fig. 4.2-2 along with the Fermi–Dirac distribution. For $E \gg kT$, the occupancy probability is sufficiently small that the issue of indistinguishability is again irrelevant and the Bose–Einstein distribution, like the Fermi–Dirac distribution in (4.2-4), reduces to the Boltzmann probability distribution (4.2-5), as displayed in Fig. 4.2-2.

Photon-Number Statistics for Thermal Light in a Cavity

Single-Mode Thermal Light. Thermal light is generated in an optical cavity whose walls are maintained at a fixed temperature T and whose atoms emit photons into the modes of the cavity. As discussed in connection with (4.2-6), when $E \gg kT$ the probability of occupancy for a collection of photons in thermal equilibrium, which is known as a **photon gas**, with energy levels E_n follows the Boltzmann probability distribution. Replacing the atomic energy-level designator m in (4.2-1) with the photon-number subscript n provides $P(E_n) \propto \exp(-E_n/kT)$, $n = 1, 2, 3, \ldots$.

This exponentially decreasing distribution is sketched in Fig. 4.2-3 with $P(E_n)$ plotted along the abscissa and E_n plotted along the ordinate. The occupancy of each energy level is random with higher energies relatively less probable than lower energies.

We proceed by assuming that a collection of photons in a mode of frequency ν has allowed energy levels specified by $E_n = (n + \frac{1}{2})h\nu$, as provided in (3.2-2) and illustrated

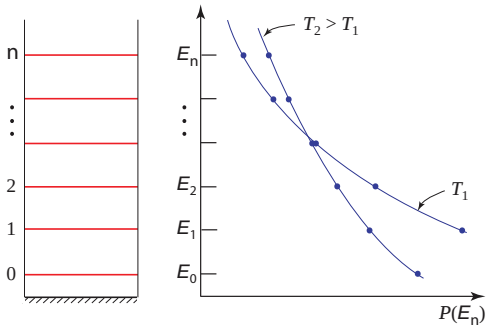


Figure 4.2-3 *Left*: Allowed energy levels of a collection of photons in a mode of frequency ν. *Right*: Boltzmann probability distribution $P(E_n)$ (plotted along the abscissa) versus energy E_n (plotted along the ordinate) for two values of the temperature T. The higher the temperature, the more likely that higher energy levels are occupied.

in Fig. 4.2-3. It follows that the probability of finding n photons in the mode is

$$p(n) \propto \exp\left(-\frac{nh\nu}{kT}\right) = \left[\exp\left(-\frac{h\nu}{kT}\right)\right]^n, \qquad n = 0, 1, 2, \ldots. \tag{4.2-7}$$

The zero-point energy $E_0 = \frac{1}{2}h\nu$ disappears into the normalization and does not affect the results. Equation (4.2-7) is normalized by imposing the condition $\sum_{n=0}^{\infty} p(n) = 1$, which yields the normalization constant $[1 - \exp(-h\nu/kT)]$.

The probability distribution for the number of photons n in a cavity mode of frequency ν given in (4.2-7) is more simply written in terms of the mean photon number $\overline{n}$ as

$$p(n) = \frac{1}{\overline{n}+1}\left(\frac{\overline{n}}{\overline{n}+1}\right)^n, \qquad n = 0, 1, 2, \ldots, \tag{4.2-8}$$

Bose–Einstein (Geometric) Photon-Number Distribution

where

$$\overline{n} = \frac{1}{\exp(h\nu/kT) - 1}, \tag{4.2-9}$$

as is readily demonstrated with the help of (3.6-2). It will become apparent in Sec. 4.7 that (4.2-9) accords with the mean photon number (4.7-7) calculated for a collection of photons interacting with atoms in thermal equilibrium, which is reassuring.

In the parlance of probability theory, the distribution presented in (4.2-8) is known as the **geometric distribution** since $p(n)$ is a geometrically decreasing function of n. In the physics literature, it is generally referred to as the **Bose–Einstein distribution** since it was first set forth by Bose based on a statistical argument for counting the states of indistinguishable particles such as photons. Einstein recognized that (4.2-8) was also applicable for describing bosons whose numbers are conserved, and he predicted the possibility of a condensation to the lowest energy state in a bosonic atomic gas cooled below a critical temperature.

The Bose–Einstein distribution is displayed on a semilogarithmic plot in Fig. 4.2-4 for several values of the mean photon number $\overline{n}$ [or, equivalently, for several values of the temperature T via (4.2-9)]. Its exponential character is apparent from the straight-line behavior on this semilogarithmic plot. Comparing Fig. 4.2-4 with Fig. 3.6-2 for the Poisson distribution demonstrates that the photon-number distributions for thermal light decrease monotonically from $n = 0$ and are far broader than those for coherent light.

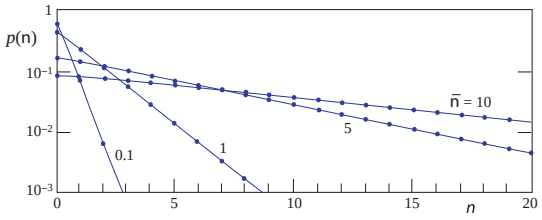


Figure 4.2-4 Semilogarithmic plot of the Bose–Einstein photon-number distribution, $p(n)$ vs. number of photons n, for four values of the mean photon number: $\overline{n}$ =0.1, 1.0, 5.0, and 10. The curves broaden substantially as $\overline{n}$ increases and the maxima always fall at $n = 0$.

The photon-number variance of the Bose–Einstein distribution, which is calculated using (3.6-3), is determined to be

$$\sigma_n^2 = \overline{n} + \overline{n}^2, \tag{4.2-10}$$

Photon-Number Variance
Bose–Einstein Distribution

where $\overline{n}$ is the photon-number mean.

□ **Normalization, Count Mean, and Count Variance of the Bose–Einstein Distribution.** The calculations are facilitated by using the substitutions $q = \overline{n}/(\overline{n}+1)$ and $1-q = 1/(\overline{n}+1)$ in (4.2-8), which convert it to the form $p(n) = (1-q)q^n$. The sum of an infinite geometric series is expressed as $\sum_{n=0}^{\infty} q^n = 1/(1-q)$, and it follows that $\sum_{n=1}^{\infty} q^n = q/(1-q)$ and $\sum_{n=2}^{\infty} q^n = q^2/(1-q)$. The orders of summation and differentiation are interchangeable since $q < 1$ and the series converge.

Normalization:
$$\sum_{n=0}^{\infty} p(n) = (1-q)\sum_{n=0}^{\infty} q^n = \frac{1-q}{1-q} = 1. \checkmark \tag{4.2-11}$$

Mean:
$$\sum_{n=0}^{\infty} np(n) = (1-q)\sum_{n=0}^{\infty} nq^n = (1-q)q\sum_{n=1}^{\infty} nq^{n-1} = (1-q)q\frac{\partial}{\partial q}\left(\sum_{n=1}^{\infty} q^n\right)$$
$$= (1-q)q\frac{\partial}{\partial q}\left(\frac{q}{1-q}\right) = \frac{q}{1-q} = \overline{n}. \checkmark \tag{4.2-12}$$

Variance:
$$\sigma_n^2 = \sum_{n=0}^{\infty}(n-\overline{n})^2\, p(n) = \sum_{n=0}^{\infty}(n^2 - n + n - \overline{n}^2)\, p(n) = \sum_{n=0}^{\infty} n(n-1)\, p(n) + \overline{n} - \overline{n}^2$$
$$= (1-q)\sum_{n=0}^{\infty} n(n-1)q^n + \overline{n} - \overline{n}^2 = (1-q)q^2\sum_{n=2}^{\infty} n(n-1)q^{n-2} + \overline{n} - \overline{n}^2$$
$$= (1-q)q^2\frac{\partial^2}{\partial q^2}\left(\sum_{n=2}^{\infty} q^n\right) + \overline{n} - \overline{n}^2 = (1-q)q^2\frac{\partial^2}{\partial q^2}\left(\frac{q^2}{1-q}\right) + \overline{n} - \overline{n}^2$$
$$= 2\left(\frac{q}{1-q}\right)^2 + \overline{n} - \overline{n}^2 = 2\overline{n}^2 + \overline{n} - \overline{n}^2 = \overline{n} + \overline{n}^2. \checkmark \tag{4.2-13}$$

■

Comparing the Bose–Einstein and Poisson variances given in (4.2-10) and (3.6-4), respectively, reveals that, for $\overline{n} > 1$, the former grows quadratically with $\overline{n}$ while the latter grows linearly. The photon-number fluctuations of the Bose–Einstein distribution are clearly far greater than those of the Poisson distribution, as is apparent by comparing Figs. 4.2-4 and 3.6-2. This large variability is consistent with the random nature of thermal light, as described in Sec. 2.7. The noisiness of the Bose–Einstein distribution is crisply highlighted by its signal-to-noise ratio, which, in accordance with (3.6-8), is given by

$$\text{SNR} = \overline{n}/(\overline{n}+1)\,. \tag{4.2-14}$$

Hence, the Bose–Einstein SNR always remains smaller than unity no matter how large the mean $\overline{n}$.

□ **Average Energy of a Cavity Mode in Thermal Equilibrium.** The average number of photons $\overline{n}$ of frequency ν, for a single mode of thermal light under conditions of thermal equilibrium at temperature T, is given by (4.2-9). Since the average energy per photon of frequency ν is $h\nu$, the average energy associated with the mode is $\overline{E} = h\nu\overline{n}$, and

$$\overline{E} = kT\frac{h\nu/kT}{\exp(h\nu/kT)-1}\,. \tag{4.2-15}$$

The dependences of $\overline{E}$ on $h\nu$ for $T = 300$ K ($kT = 0.026$ eV) and for $T = 600$ K ($kT = 0.052$ eV) are displayed in Fig. 4.2-5. In the limit $h\nu/kT \ll 1$, i.e., when the photon energy is much smaller

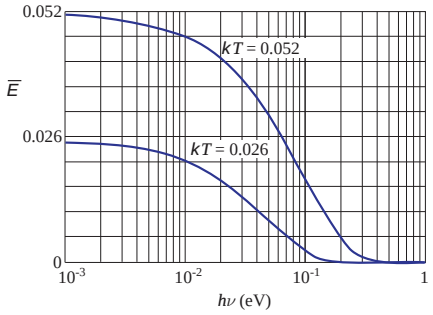


Figure 4.2-5 Average energy $\overline{E}$ of a cavity mode in thermal equilibrium vs. photon energy $h\nu$ for $T = 300$ K ($kT = 0.026$ eV) and for $T = 600$ K ($kT = 0.052$ eV).

than the unit of thermal energy, a Taylor-series expansion provides $\exp(h\nu/kT) \approx 1 + (h\nu/kT)$, whereupon (4.2-15) reduces to $\overline{E} \approx kT$. The average energy of the mode then matches that expected from the classical equipartition theorem, as would obtained were the light not quantized. ■

Multimode Thermal Light. Multimode thermal light in a cavity is taken to be a collection of $\mathcal{M}$ independent thermal modes sufficiently close to each other in frequency that each obeys a Bose–Einstein distribution with the same mean photon number $\overline{n} = 1/[\exp(h\nu/kT)-1]$, as provided in (4.2-9). Since this light comprises a sum of random numbers of photons contributed by independent individual modes, the overall photon-number distribution is the $\mathcal{M}$-fold self-convolution of the Bose–Einstein single-mode distribution set forth in (4.2-8). The result is the **negative-binomial distribution**

$$p(m) = \binom{m+\mathcal{M}-1}{m}\frac{(\overline{m}/\mathcal{M})^m}{(1+\overline{m}/\mathcal{M})^{m+\mathcal{M}}}\,, \tag{4.2-16}$$

Negative-Binomial Photon-Number Distribution

with overall mean count

$$\overline{m} = \mathcal{M}\overline{n}. \tag{4.2-17}$$

It is straightforward to show that the negative-binomial distribution reduces to the Bose–Einstein distribution (4.2-8) for $\mathcal{M} = 1$ and to the Poisson distribution (3.6-1) as $\mathcal{M} \to \infty$.

The variance of the overall photon number σ_m^2 is the sum of the variances of the individual modes, as provided in (4.2-10), which may be written in terms of the overall multimode count mean $\overline{m}$ as

$$\sigma_m^2 = \mathcal{M}(\overline{n} + \overline{n}^2) = \overline{m} + \frac{\overline{m}^2}{\mathcal{M}} . \tag{4.2-18}$$

Photon-Number Variance
Negative-Binomial Distribution

Since $\mathcal{M} \geqslant 1$, the photon-number variance for multimode thermal light is reduced below that for single-mode thermal light of the same mean, a result that arises from averaging.

4.3 INTERACTIONS OF PHOTONS WITH ATOMS

An atom may emit (create) or absorb (annihilate) a photon by undergoing a downward or upward transition between pairs of its energy levels, while conserving energy in the process. The elementary laws that govern such emissions and absorptions are described. The interaction of photons with electrons and holes in semiconductor materials is considered in detail in Chapter 6.

Elementary Interactions

Consider an atom with two energy levels, E_1 and E_2, placed in an optical cavity of volume V that can sustain a number of electromagnetic modes. We are particularly interested in the interaction between the atom and the photons of a *prescribed* radiation mode of frequency $\nu \approx \nu_0$, where $h\nu_0 = E_2 - E_1$, since photons of this energy match the atomic energy-level difference. A formal study of such interactions relies on quantum electrodynamics; we present the key results that emerge from such an analysis below, without proof.

Three forms of interaction are possible — spontaneous emission, absorption, and stimulated emission, as schematized in Figs. 4.3-1, 4.3-2, and 4.3-3, respectively, which we consider in turn.

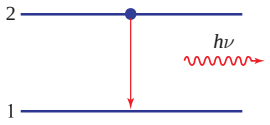


Figure 4.3-1 Spontaneous emission of a photon into a mode of frequency ν via a transition from atomic energy level 2 to energy level 1.

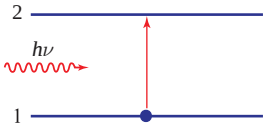


Figure 4.3-2 Absorption is a process whereby a photon of energy $h\nu$ induces the atom to undergo an upward transition from level 1 to level 2.

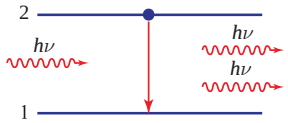


Figure 4.3-3 In stimulated emission, a photon of energy $h\nu$ induces the emission of a clone photon as the atom transitions from level 2 to level 1.

Spontaneous Emission. If the atom is initially in the upper energy level, it may decay spontaneously to the lower energy level and release its energy in the form of a photon (Fig. 4.3-1). The photon energy $h\nu \approx E_2 - E_1$ is added to the energy of the electromagnetic mode. The process is called **spontaneous emission** because the transition is independent of the number of photons that may already be in the mode.

In a cavity of volume V, the probability density (per second), or rate, for a spontaneous transition depends on ν in a way that characterizes that atomic transition,

$$p_{\text{sp}} = \frac{c}{V}\,\sigma(\nu). \tag{4.3-1}$$

Spontaneous Emission (into a Prescribed Mode)

The quantity $\sigma(\nu)$, known as the **transition cross section**, is a function of ν centered about the atomic resonance frequency ν_0. The significance of this quantity will become apparent subsequently, but it is clear that σ has dimensions of cm^2 (since the dimensions of p_{sp}, c, and V are s^{-1}, cm/s, and cm^3, respectively). In principle, $\sigma(\nu)$ can be determined using fundamental quantum mechanics but the calculations are generally onerous and suffer from inaccuracies, so it is usually determined empirically. Equation (4.3-1) applies separately to every mode, with a transition cross section σ that depends on the angle θ between the dipole moment of the atom and the field direction of the mode, in accordance with

$$\sigma = \sigma_{\text{max}} \cos^2\theta. \tag{4.3-2}$$

The maximum cross section σ_{max} is attained when the dipole moment and field align.

The term "probability density" signifies that the probability of an emission taking place in an incremental time interval between t and $t + \Delta t$ is simply $p_{\text{sp}}\,\Delta t$. Because it is a probability density, p_{sp} can have a numerical value greater than $1\ \text{s}^{-1}$, although of course $p_{\text{sp}}\Delta t \leqslant 1$. Thus, if there are a large number N of such atoms, a fraction of approximately $\Delta N = (p_{\text{sp}}\,\Delta t)N$ atoms will undergo this transition within the time interval Δt. Consequently, we can write $dN/dt = -p_{\text{sp}}\,N$, which indicates that the number of atoms $N(t) = N(0)\exp(-p_{\text{sp}}t)$ decays exponentially with time constant $1/p_{\text{sp}}$, as illustrated in Fig. 4.3-4.

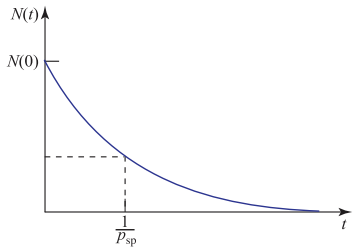


Figure 4.3-4 Spontaneous emission into a single mode results in an exponential decrease of the number of excited atoms, with time constant $1/p_{\text{sp}}$.

Absorption. If the atom is initially in the lower energy level and the radiation mode contains a photon, the photon may be annihilated and the atom concomitantly raised to the upper energy level (Fig. 4.3-2). This process, which is *induced* by the photon, is called **absorption**. It is also referred to as or **induced absorption** or **stimulated absorption**. It can occur only when the mode contains a photon.

The probability density for the absorption of a photon from a given mode of frequency ν, in a cavity of volume V, is governed by the *same law* that governs spontaneous emission into that mode, namely

$$p_{\text{ab}} = \frac{c}{V}\,\sigma(\nu). \tag{4.3-3}$$

However, if there are n photons in the mode, the probability density that the atom absorbs *one* photon is n times greater since the events are mutually exclusive, i.e.,

$$P_{\text{ab}} = n\,\frac{c}{V}\,\sigma(\nu). \tag{4.3-4}$$

Absorption of One Photon (from a Mode with n Photons)

Stimulated Emission. Finally, if the atom is in the upper energy level and the mode contains a photon, the atom may be *induced* to emit another photon into the same mode. This process, known as **stimulated emission** or **induced emission**, is the inverse of absorption. The presence of a photon in a mode of specified frequency, propagation direction, and polarization stimulates the emission of a duplicate ("clone") photon with precisely the same characteristics as the original (Fig. 4.3-3). This photon amplification process underlies the operation of laser amplifiers and lasers.

The probability density p_{st} that this process occurs in a cavity of volume V is governed by the *same law* that governs spontaneous emission and absorption:

$$p_{\text{st}} = \frac{c}{V}\,\sigma(\nu). \tag{4.3-5}$$

If the mode originally carries n photons, the probability density that the atom is stimulated to emit an additional photon is, just as in the case of absorption,

$$P_{\text{st}} = n\,\frac{c}{V}\,\sigma(\nu). \tag{4.3-6}$$

Stimulated Emission of One Photon (into a Mode with n Photons)

For notational convenience, we use the common designator W_i for both the absorption of one photon and the stimulated emission of one photon:

$$W_i \equiv P_{\text{ab}} = P_{\text{st}}. \tag{4.3-7}$$

Probability Density for One Photon Absorption and Stimulated Emission

Inasmuch as spontaneous emission is present along with stimulated emission, combining (4.3-1) and (4.3-6) leads to an overall probability density for the atom emitting a photon into the mode, i.e., $p_{\text{sp}} + P_{\text{st}} = (n+1)(c/V)\sigma(\nu)$. From a quantum-electrodynamic point of view, spontaneous emission may be regarded as stimulated emission induced by the zero-point fluctuations associated with the mode (Sec. 3.2). Because the zero-point energy plays no role in absorption, however, P_{ab} is proportional to n rather than to $(n+1)$.

Lineshape Function and Transition Strength

It is clear from the foregoing that the transition cross section $\sigma(\nu)$ characterizes the interaction of the atom with the photon. Its shape governs the relative magnitude of the interaction of the atom with photons over a range of frequencies, while its area,

$$S = \int_0^\infty \sigma(\nu)\,d\nu, \tag{4.3-8}$$

known as the **transition strength**, represents the strength of the interaction. The area S, which has units of cm^2-Hz, can be readily separated from the shape (profile) of $\sigma(\nu)$ by defining a normalized **lineshape function** $g(\nu) = \sigma(\nu)/S$, which has unity area, $\int_0^\infty g(\nu)\,d\nu = 1$, and units of Hz^{-1}. The transition cross section can then be written in terms of its strength and profile as

$$\sigma(\nu) = Sg(\nu). \tag{4.3-9}$$

The lineshape function $g(\nu)$ is centered about the resonance frequency ν_0, where $\sigma(\nu)$ is largest, and decreases sharply as ν deviates from ν_0. Transitions are therefore most likely to occur for photons of frequency $\nu \approx \nu_0$. The width of the function $g(\nu)$ is known as the **transition linewidth** $\Delta\nu$, which is usually defined as the full-width at half-maximum (FWHM) value of $g(\nu)$ (see Sec. A.2 of Appendix A). Since $g(\nu)$ has unity area, its width is inversely proportional to its central value,

$$\Delta\nu \propto 1/g(\nu_0). \tag{4.3-10}$$

It is also useful to define a **peak cross section** at the resonance frequency, $\sigma_0 \equiv \sigma(\nu_0)$. As illustrated in Fig. 4.3-5, the transition cross section $\sigma(\nu)$ is then characterized by four features: 1) its height σ_0; 2) its width $\Delta\nu$; 3) its area S; and 4) its profile $g(\nu)$.

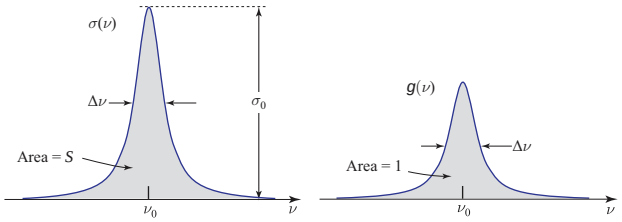


Figure 4.3-5 Features of the transition cross section $\sigma(\nu)$ and the lineshape function $g(\nu)$.

The process of spontaneous emission is reexamined in Sec. 4.4 in the context of photon emission into *any* available mode, rather than into a prescribed mode. In the same vein, the processes of absorption and stimulated emission are revisited in Sec. 4.5 from the perspective of transitions induced by *broadband*, rather than monochromatic, light. A number of line-broadening mechanisms and their lineshape functions are discussed in Sec. 4.6.

4.4 SPONTANEOUS EMISSION

Spontaneous Emission into Any Available Mode

Equation (4.3-1) provides the probability density p_{sp} for spontaneous emission into a *prescribed* mode of frequency ν, without regard to whether the mode contains photons. Of paramount interest, however, is the probability density for the spontaneous emission of one photon of frequency ν into *any* available optical mode, as illustrated schematically in Fig. 4.4-1.

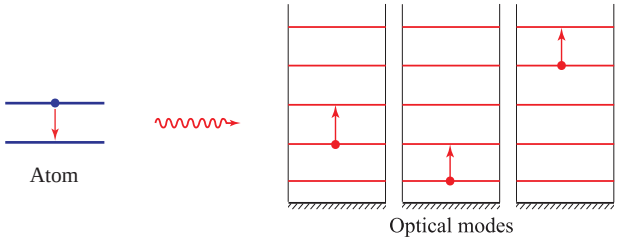


Figure 4.4-1 An atom may spontaneously emit a photon into any one (but only one) of the many available optical modes with frequencies $\nu \approx \nu_0$.

To determine the probability density for the total spontaneous emission into all modes, it is required to have knowledge not only of the probability density for spontaneous emission into a specific mode, but also the density of modes. We proceed to demonstrate that the density of modes for a three-dimensional cavity increases quadratically with frequency.

Density of Modes in a Three-Dimensional Cavity

Planar-Mirror Cavity. A 3D planar-mirror cavity is constructed from three pairs of parallel mirrors that form the walls of a closed rectangular box of dimensions d_x, d_y, and d_z. The structure is a three-dimensional cavity, as depicted in Fig. 4.4-2(a). By virtue of the boundary conditions, standing-wave electric-field solutions within the cavity require that the components of the wavevector $\mathbf{k} = (k_x, k_y, k_z)$ are discretized and obey

$$k_x = q_x \frac{\pi}{d_x}, \quad k_y = q_y \frac{\pi}{d_y}, \quad k_z = q_z \frac{\pi}{d_z}, \qquad q_x, q_y, q_z = 1, 2, \ldots, \tag{4.4-1}$$

where q_x, q_y, and q_z are positive integers representing the respective mode numbers. The k-space construct for a cubic cavity with $d_x = d_y = d_z = d$ is illustrated in Fig. 4.4-2(b). Each mode $\mathbf{q}$, characterized by the three integers (q_x, q_y, q_z), is represented by a dot in (k_x, k_y, k_z)-space. The spacing between the dots in a given direction is inversely proportional to the width of the cavity along that direction.

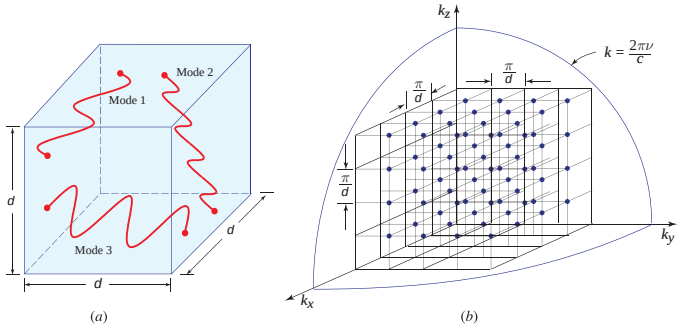


Figure 4.4-2 (a) Waves in a three-dimensional cubic cavity ($d_x = d_y = d_z = d$). (b) The endpoints of the wavevectors (k_x, k_y, k_z) of the modes in a three-dimensional cavity are indicated by dots. The wavenumber k of a mode is the distance from the origin to the dot. Each point in k-space occupies a volume $(\pi/d)^3$. All modes of frequency smaller than ν lie inside the positive octant of a sphere of radius $k = 2\pi\nu/c$.

The values of the wavenumbers k, and the corresponding resonance frequencies ν, satisfy

$$k^2 = k_x^2 + k_y^2 + k_z^2 = \left(\frac{2\pi\nu}{c}\right)^2. \tag{4.4-2}$$

The surface of constant frequency ν is a sphere of radius $k = 2\pi\nu/c$. The resonance frequencies are determined from (4.4-1) and (4.4-2), so that

$$\nu_{\mathbf{q}} = \sqrt{q_x^2\,\nu_{Fx}^2 + q_y^2\,\nu_{Fy}^2 + q_z^2\,\nu_{Fz}^2}\,, \qquad q_x, q_y, q_z = 1, 2, \ldots, \tag{4.4-3}$$

Resonance Frequencies

where

$$\nu_{Fx} = \frac{c}{2d_x}, \quad \nu_{Fy} = \frac{c}{2d_y}, \quad \nu_{Fz} = \frac{c}{2d_z} \tag{4.4-4}$$

are frequency spacings that are inversely proportional to the cavity widths in the x, y, and z directions, respectively. For cavities whose dimensions are much larger than a wavelength, the frequency spacings are much smaller than the optical frequency. For example, if $d = 1$ cm and $n = 1$, we have $\nu_F = 15$ GHz.

Density of Modes. When all the dimensions of the cavity are much greater than a wavelength, the mode spacing (free spectral range) $\nu_F = c/2d$ is small, and it is difficult to analytically enumerate the modes. In this case, however, we can resort to a continuous approximation and introduce the concept of the **density of modes**, the validity of which depends on the relative values of the frequency bandwidth of interest and the frequency interval between successive modes.

The number of modes that lie in the frequency interval between 0 and ν corresponds to the number of points that lie in the volume of the positive octant of a sphere of radius k in the k-space diagram portrayed in Fig. 4.4-2(b). The number of modes in the positive octant of a sphere of radius k is $2(\frac{1}{8})(\frac{4}{3}\pi k^3)/(\pi/d)^3 = (k^3/3\pi^2)d^3$. The initial factor of 2 accommodates the two possible polarizations of each mode, while the denominator $(\pi/d)^3$ represents the volume in k-space per point. It follows that the number of modes with wavenumbers between k and $k + \Delta k$, per unit volume, is $\varrho(k)\Delta k = [(d/dk)(k^3/3\pi^2)]\Delta k = (k^2/\pi^2)\Delta k$, so that the density of modes in k-space is $\varrho(k) = k^2/\pi^2$. This derivation is identical to that used for determining the density of allowed quantum states for electron waves confined within perfectly reflecting walls in a bulk semiconductor, as provided in (5.4-1).

Since $k = 2\pi\nu/c$, the number of modes lying between 0 and ν is $[(2\pi\nu/c)^3/3\pi^2]d^3$ = $(8\pi\nu^3/3c^3)d^3$. The number of modes in the incremental frequency interval lying between ν and $\nu + \Delta\nu$ is therefore $(d/d\nu)[(8\pi\nu^3/3c^3)d^3]\Delta\nu = (8\pi\nu^2/c^3)d^3\Delta\nu$. The density of modes $M(\nu)$, i.e., the number of modes per unit bandwidth surrounding the frequency ν, per unit cavity volume, is thus

$$\boxed{M(\nu) \approx \frac{8\pi\nu^2}{c^3}.} \tag{4.4-5}$$

Density of Modes (3D Cavity)

Equation (4.4-5) provides a suitable approximation for the number of modes of frequency ν, per unit volume of the cavity, per unit frequency bandwidth, that may be used when the mode spacing is sufficiently small that a continuous approximation can be used for counting. Although the density of modes was derived on the basis of cubic geometry, the results are applicable for arbitrary geometries, provided that the cavity dimensions are large in comparison with the wavelength. Equation (4.4-5) is useful in various areas of physics and is an integral part of the calculation to determine the spectrum of blackbody radiation (Sec. 4.7).

Since $M(\nu)$ increases quadratically with frequency, the number of modes within a fixed frequency bandwidth $\Delta\nu$ increases with the frequency ν in the manner sketched in Fig. 4.4-3. As an example, at $\nu = 3 \times 10^{14}$ ($\lambda_0 = 1$ μm), $M(\nu) = 0.08$ modes/cm^3-Hz. Within a frequency band of width 1 GHz, there are then $\approx 8 \times 10^7$ modes/cm^3. The number of modes per unit volume within an arbitrary frequency interval $\nu_1 < \nu < \nu_2$ is given by the integral $\int_{\nu_1}^{\nu_2} M(\nu)\, d\nu$.

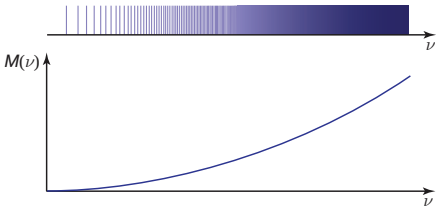


Figure 4.4-3 The density of modes $M(\nu)$ vs. ν for a three-dimensional optical cavity increases as a quadratic function of frequency. As sketched in the inset, the frequency spacing between adjacent modes decreases as the frequency increases.

Total Spontaneous Emission into All Modes

Equation (4.4-5) for the density of modes makes it possible to calculate the probability density for spontaneous emission into all modes, which is the probability density for spontaneous emission into a specific mode (4.3-1), weighted by the modal density (4.4-5). Since the modes at each frequency have an isotropic distribution of directions, each with two polarizations, we make use of the average transition cross section $\overline{\sigma}(\nu)$. If θ is the angle between the dipole moment of the atom and the field direction, (4.3-2) provides

$$\overline{\sigma}(\nu) = \tfrac{1}{3}\sigma_{\max} \tag{4.4-6}$$

since $\langle \cos^2\theta \rangle = 1/3$, where the symbol $\langle \cdot \rangle$ represents averaging in 3D space.

The total spontaneous-emission probability density may therefore be written as

$$P_{\text{sp}} = \int_0^\infty \left[\frac{c}{V}\,\overline{\sigma}(\nu)\right] [V M(\nu)]\, d\nu = c \int_0^\infty \overline{\sigma}(\nu)\, M(\nu)\, d\nu. \tag{4.4-7}$$

Because the function $\overline{\sigma}(\nu)$ is sharply peaked, it is narrow in comparison with the quadratic function $M(\nu) = 8\pi\nu^2/c^3$. Since $\overline{\sigma}(\nu)$ is centered about ν_0, $M(\nu)$ is approximately constant with value $M(\nu_0)$, and can thus be removed from the integral. The probability density for the spontaneous emission of one photon into *any* mode is therefore

$$P_{\text{sp}} = M(\nu_0)\, c\overline{S} = \frac{8\pi\nu_0^2 \overline{S}}{c^2} = \frac{8\pi\overline{S}}{\lambda^2}, \tag{4.4-8}$$

where $\lambda = c/\nu_0$ is the wavelength of the light in the medium and $\overline{S} = \int_0^\infty \overline{\sigma}(\nu)\, d\nu$. We define a time constant t_{sp}, known as the **spontaneous lifetime** for the $2 \rightarrow 1$ transition, such that $1/t_{\text{sp}} \equiv P_{\text{sp}}$, so

$$\boxed{P_{\text{sp}} = \frac{1}{t_{\text{sp}}},} \tag{4.4-9}$$

Spontaneous Emission (Broadband)

which is independent of the cavity volume V.

Combining (4.4-8) and (4.4-9) leads to

$$\overline{S} = \frac{\lambda^2}{8\pi t_{\text{sp}}}, \tag{4.4-10}$$

which enables the transition strength to be determined from an empirical measurement of the spontaneous lifetime t_{sp}. For a well-known transition, such as that between the first excited and ground states of atomic hydrogen, we have $t_{\text{sp}} \approx 10^{-8}$ s; however, t_{sp} can vary over a range that extends from femtoseconds to seconds. Equation (4.4-10) is useful because a first-principles calculation of $\overline{S}$ would require intimate knowledge of the quantum-mechanical behavior of the system, which is not always available, or easy to compute if it is available.

Relation Between Transition Cross Section and Spontaneous Lifetime

Using (4.4-10), together with the formula $\overline{\sigma}(\nu) = \overline{S}g(\nu)$, which derives from (4.3-9), leads to a relation that connects the average transition cross section with the spontaneous lifetime and the lineshape function:

$$\boxed{\overline{\sigma}(\nu) = \frac{\lambda^2}{8\pi t_{\rm sp}}\, g(\nu).} \qquad \begin{matrix}(4.4\text{-}11)\\ \text{Average Transition}\\ \text{Cross Section}\end{matrix}$$

This formula is known as the *Füchtbauer–Ladenburg equation*. The average transition cross section at the central frequency ν_0 is therefore

$$\overline{\sigma}_0 \equiv \overline{\sigma}(\nu_0) = \frac{\lambda^2}{8\pi t_{\rm sp}}\, g(\nu_0). \tag{4.4-12}$$

Because $g(\nu_0)$ is inversely proportional to $\Delta\nu$, for a given value of $t_{\rm sp}$ the peak transition cross section $\overline{\sigma}_0$ is inversely proportional to the linewidth $\Delta\nu$, in accordance with (4.3-10).

Frequencies of Spontaneously Emitted Photons. The probability density ($\rm s^{-1}$) of an excited atom spontaneously emitting a photon into any of the modes in the frequency band ν to $\nu + d\nu$ is specified by the integrand of (4.4-7), namely $P_{\rm sp}(\nu)\, d\nu = (c/V)\,\overline{\sigma}(\nu)\, V M(\nu)\, d\nu$. The average transition cross section in turn is given by $\overline{\sigma}(\nu) = (\lambda^2/8\pi t_{\rm sp})\, g(\nu)$ in accordance with (4.4-11), and the density of modes per unit volume is expressed as $M(\nu) = 8\pi\nu^2/c^3$ as specified in (4.4-5). Combining these three equations yields the probability density that a spontaneously emitted photon has a frequency lying between ν and $\nu + d\nu$,

$$P_{\rm sp}(\nu)\, d\nu = (1/t_{\rm sp})\, g(\nu)\, d\nu, \tag{4.4-13}$$

which is proportional to $g(\nu)d\nu$. Hence, when many photons are spontaneously emitted, the distribution of their frequencies follows the lineshape function $g(\nu)$.

4.5 ABSORPTION AND STIMULATED EMISSION

We now turn from spontaneous emission to absorption and stimulated emission. We begin by considering the interaction of single-mode light with an atom when a stream of photons impinges on it, rather than when the atom resides in a cavity of volume V, as considered earlier. We then investigate atomic photon absorption and emission induced by broadband light and relate the ensuing transition rates to those for spontaneous emission.

Transitions Induced by Monochromatic Light

Let monochromatic light of frequency ν, intensity I, and mean photon-flux density (photons/$\rm cm^2$-s)

$$\phi = I/h\nu \tag{4.5-1}$$

interact with an atom whose resonance frequency is ν_0. We seek to determine the probability densities for absorption and stimulated emission, $W_i \equiv P_{\rm ab} = P_{\rm st}$, in this configuration.

The number of photons n involved in the interaction is determined by constructing a volume in the form of a cylinder of base area A, height $c \times 1$ s, and volume $V = cA$. The axis of the cylinder is parallel to $\mathbf{k}$, the direction of propagation of the light. The photon flux that crosses the cylinder base is $\Phi = \phi A$ (photons/s). Because photons travel at the speed of light c, all of the photons within the volume of the cylinder cross its base within one second. It follows that, at any time, the cylinder contains $n = \phi A = \phi V/c$ photons so that

$$\phi = n \frac{c}{V} . \tag{4.5-2}$$

To determine W_i, we substitute (4.5-2) into (4.3-4) or (4.3-6), and make use of (4.3-7), to obtain

$$\boxed{W_i = \phi \sigma(\nu).} \tag{4.5-3}$$

It is apparent that $\sigma(\nu)$ is the coefficient of proportionality between the probability density of an induced transition and the photon-flux density. This relationship informs us that the appellation "transition cross section" is apt: ϕ is the photon-flux density (cm^{-2}·s^{-1}) while $\sigma(\nu)$ is the effective cross-sectional area of the atom (cm^2), so that $\phi\,\sigma(\nu)$ represents the probability density (s^{-1}) that a photon in the stream is "captured" by the "cross section" of the atom for the purpose of absorption or stimulated emission.

It is clear from (4.3-4), (4.3-6), and (4.4-7) that the probability densities for absorption, stimulated emission, and spontaneous emission are all proportional to $\sigma(\nu)$. As discussed above, stimulated emission involves decay only into those modes that contain photons. Although the expression for $\overline{\sigma}(\nu)$ set forth in (4.4-11) was obtained for spontaneous emission into multiple modes, it is convenient to use it in conjunction with (4.5-3) to determine the probability density for induced transitions as well, since $t_{\rm sp}$ is readily determined experimentally.

The use of the quantity $\overline{\sigma}(\nu)$ instead of $\sigma(\nu)$ in (4.4-11) is a result of averaging over the angle between the dipole moment of the atom and the field direction [see (4.3-2) and (4.4-6)]; it is appropriate for spontaneous emission into all modes. However, when such averaging is not called for, as in the case of stimulated emission into a particular mode and a fixed angle θ, $\sigma(\nu)$ and σ_0 are used in place of $\overline{\sigma}(\nu)$ and $\overline{\sigma}_0$. Any change in $\sigma(\nu)$ required for averaging with a particular induced-transition configuration can be readily accommodated by modifying $t_{\rm sp}$, which is then referred to as the **effective spontaneous lifetime**. For simplicity, we shall henceforth not distinguish between $t_{\rm sp}$ for spontaneous emission and its effective value for stimulated emission.

Transitions Induced by Broadband Light

Consider now an atom in a cavity of volume V containing multimode polychromatic light of spectral energy density $\varrho(\nu)$ (energy per unit frequency per unit volume) that is broadband in comparison with the atomic linewidth. The average number of photons in the frequency band from ν to $\nu + d\nu$ is $[\varrho(\nu)V/h\nu]\,d\nu$; each of these has a probability density $(c/V)\sigma(\nu)$ of initiating an atomic transition. As with spontaneous emission, the modes at each frequency are taken to be isotropically distributed with two polarizations, so that the overall probability of absorption or stimulated emission is

$$W_i = \int_0^\infty \frac{\varrho(\nu)V}{h\nu} \left[\frac{c}{V}\,\overline{\sigma}(\nu)\right] d\nu. \tag{4.5-4}$$

Since the radiation is broadband, the function $\varrho(\nu)$ varies slowly in comparison with the sharply peaked transition cross section $\overline{\sigma}(\nu)$. We can therefore replace $\varrho(\nu)/h\nu$ under the integral with $\varrho(\nu_0)/h\nu_0$, which leads to

$$W_i = \frac{\varrho(\nu_0)}{h\nu_0}\, c \int_0^\infty \overline{\sigma}(\nu)\, d\nu = \frac{\varrho(\nu_0)}{h\nu_0}\, c\overline{S}. \tag{4.5-5}$$

Using (4.4-10), we therefore have

$$W_i = \frac{\lambda^3}{8\pi h t_{\rm sp}}\, \varrho(\nu_0), \tag{4.5-6}$$

where $\lambda = c/\nu_0$ is the wavelength in the medium at the central frequency ν_0. Defining

$$\overline{n} = \frac{\lambda^3}{8\pi h}\, \varrho(\nu_0), \tag{4.5-7}$$

which represents the mean number of photons per mode, allows us to write (4.5-6) in the convenient form

$$\boxed{W_i = \overline{n}/t_{\rm sp}\,.} \tag{4.5-8}$$

Induced Transition (Broadband Light)

The interpretation of $\overline{n}$ as the mean number of photons per mode follows from the form of the ratio [see (4.4-8), (4.5-5), and (4.5-6)]

$$\frac{W_i}{P_{\rm sp}} = \frac{\lambda^3 \varrho(\nu_0)}{8\pi h t_{\rm sp}} \frac{1}{M(\nu_0) c\overline{S}} = \frac{\varrho(\nu_0)}{h\nu_0 M(\nu_0)}\,; \tag{4.5-9}$$

the quantity $\varrho(\nu_0)/h\nu_0$ represents the mean number of photons per unit volume in the vicinity of the frequency ν_0 while $M(\nu_0)$ is the number of modes per unit volume in the vicinity of ν_0. The probability density W_i is thus a factor of $\overline{n}$ greater than that for spontaneous emission, as provided in (4.4-9), since each mode contains an average of $\overline{n}$ photons. Broadband absorption and stimulation emission evidently share a close relationship with broadband spontaneous emission.

Einstein A and B Coefficients. Although Einstein did not have knowledge of (4.5-6), in 1917 he carried out an analysis of the energy exchange between atoms and radiation that led him to general expressions for the probability densities of spontaneous and stimulated transitions. He assumed that the atoms interacted with broadband radiation of spectral energy density $\varrho(\nu)$, under conditions of thermal equilibrium, and obtained the following expressions:

$$\boxed{\begin{aligned} P_{\rm sp} &= A \\ W_i &= B\,\varrho(\nu_0). \end{aligned}} \tag{4.5-10}$$

(4.5-11)

Einstein's Postulates

The constants A and B are known as the **Einstein A and B coefficients**.

Comparison of (4.5-10) and (4.5-11) with (4.4-9) and (4.5-6), respectively, reveals that the A and B coefficients correspond to

$$A = \frac{1}{t_{\rm sp}} \tag{4.5-12}$$

$$B = \frac{\lambda^3}{8\pi h t_{\rm sp}}, \tag{4.5-13}$$

which are associated with spontaneous and stimulated transitions, respectively. Their ratio is given by

$$\frac{B}{A} = \frac{\lambda^3}{8\pi h}. \tag{4.5-14}$$

The relation between the A and B coefficients is a result of the microscopic probability laws of interaction between an atom and the photons of each mode. An analysis similar to that provided by Einstein will be presented in Sec. 4.7 in connection with blackbody radiation.

EXAMPLE 4.5-1. ***Comparison Between Spontaneous and Stimulated Emission Rates.*** Whereas the rate of spontaneous emission for an atom in the upper state is constant at $A = 1/t_{\rm sp}$, the rate of stimulated emission in the presence of broadband light, $B\varrho(\nu_0)$, is proportional to the spectral energy density of the light, $\varrho(\nu_0)$. The two rates are equal when $\varrho(\nu_0) = A/B = 8\pi h/\lambda^3$; for larger values of the spectral energy density, the rate of stimulated emission exceeds that of spontaneous emission. If $\lambda = 1\ \mu$m, for example, $A/B = 1.66 \times 10^{-14}$ J/m^3-Hz. This corresponds to an intensity spectral density $c\varrho(\nu_0) \approx 5 \times 10^{-6}$ W/m^2-Hz in free space. Thus, for a linewidth $\Delta\nu = 10^7$ Hz, the optical intensity at which the stimulated emission rate equals the spontaneous emission rate is 50 W/m^2 or 5 mW/cm^2.

Summary: Transition Cross Section

An atomic transition may be characterized by its resonance frequency $\nu_0 = (E_2 - E_1)/h$, spontaneous lifetime $t_{\rm sp}$, and lineshape function $g(\nu)$, which has linewidth $\Delta\nu$. The average transition cross section is

$$\overline{\sigma}(\nu) = \overline{S}g(\nu) = \frac{\lambda^2}{8\pi t_{\rm sp}}\, g(\nu). \tag{4.4-11}$$

Summary: Spontaneous Emission

- If the atom is in the upper level and in a cavity of volume V, the probability density (per second) of emitting spontaneously into one *prescribed* mode of frequency ν is

$$p_{\mathrm{sp}} = \frac{c}{V}\,\sigma(\nu). \tag{4.3-1}$$

- The probability density of spontaneous emission into *any* of the available modes is

$$P_{\mathrm{sp}} = \frac{8\pi\overline{S}}{\lambda^2} = \frac{1}{t_{\mathrm{sp}}} = A\,. \tag{4.4-9}$$

- The probability density of emitting into modes lying only in the frequency band between ν and $\nu + d\nu$ is $P_{\mathrm{sp}}\,d\nu = (1/t_{\mathrm{sp}})g(\nu)\,d\nu$.

Summary: Stimulated Emission and Absorption

- If the atom in the cavity is in the upper level and a radiation mode contains n photons of frequency ν, the probability density of emitting a photon into that mode is

$$W_i = n\,\frac{c}{V}\,\sigma(\nu). \tag{4.3-6}$$

  If the atom is instead in the lower level, and a mode contains n photons, the probability density of absorption of a photon from that mode is also given by (4.3-6).
- If instead of being in a cavity, the atom is illuminated by a monochromatic beam of light of frequency ν, with mean photon-flux density ϕ (photons per second per unit area), the probability density of stimulated emission (if the atom is in the upper level) or absorption (if the atom is in the lower level) is

$$W_i = \phi\,\sigma(\nu). \tag{4.5-3}$$

- If the light illuminating the atom is polychromatic, but narrowband in comparison with the atomic linewidth, and has a mean spectral photon-flux density ϕ_ν (photons per second per unit area per unit frequency), the probability density of stimulated emission/absorption is

$$W_i = \int \phi_\nu\,\sigma(\nu)\,d\nu. \tag{4.5-15}$$

- If the light illuminating the atom has a spectral energy density $\varrho(\nu)$ that is broadband in comparison with the atomic linewidth, the probability density of stimulated emission/absorption is

$$W_i = B\,\varrho(\nu_0), \tag{4.5-11}$$

  where $B = \lambda^3/8\pi h t_{\mathrm{sp}}$ is the Einstein B coefficient.

In all of these formulas, $c = c_0/n$ is the velocity of light and $\lambda = \lambda_0/n$ is the wavelength of light, in the atomic medium, and n is the refractive index.

The processes of spontaneous emission, absorption, and stimulated emission discussed in the foregoing sections, together with the principles of photon optics set forth in Chapter 3, serve as the basis for understanding the origin and properties of blackbody and thermal radiation, as discussed in Secs. 4.7 and 4.8, respectively. In semiconductor photonics, these same processes underlie the operation of LEDs, as described in Chapter 6.

4.6 LINE BROADENING

Because the lineshape function $g(\nu)$ plays a central role in atom–photon interactions, we conduct a brief foray into some of the mechanisms that lead to line broadening. The same lineshape function applies for spontaneous emission, absorption, and stimulated emission.

Lifetime Broadening

Atoms can undergo transitions between energy levels by both radiative and nonradiative means. Radiative transitions are associated with photon absorption and emission, whereas nonradiative transitions permit energy transfer to take place via mechanisms such as lattice vibrations, inelastic collisions among constituent atoms, and inelastic collisions with the walls of the vessel. Each atomic energy level has a lifetime τ, which is the inverse of the rate at which its population decays, radiatively or nonradiatively, to all lower levels.

The lifetime τ_2 of energy level 2 displayed in Fig. 4.3-1, for example, represents the inverse of the rate at which the population of that level decays to level 1 and to all other lower energy levels (none of which are shown in the figure), by either radiative or nonradiative means. Since $1/t_{\text{sp}}$ is the radiative decay rate from level 2 to level 1, the overall decay rate $1/\tau_2$ must be larger, i.e., $1/\tau_2 \geqslant 1/t_{\text{sp}}$, thereby corresponding to a shorter decay time, $\tau_2 \leqslant t_{\text{sp}}$. The lifetime τ_1 of level 1 is defined similarly. Clearly, if level 1 is the lowest allowed energy level (the ground state), then it will never decay and $\tau_1 = \infty$.

Lifetime broadening is, in essence, a Fourier transform effect. The lifetime τ of an energy level is related to the time uncertainty of the occupation of that level. As discussed in Appendix A, the Fourier transform of an exponentially decaying harmonic field $e^{-t/2\tau}e^{j2\pi\nu_0 t}$, whose energy decays as $e^{-t/\tau}$ with time constant τ, is proportional to $1/[1+j4\pi(\nu-\nu_0)\tau]$. The full-width at half-maximum (FWHM) of the absolute square of this Lorentzian function of frequency is $\Delta\nu = 1/2\pi\tau$. This spectral uncertainty corresponds to an energy uncertainty $\Delta E = h\Delta\nu = h/2\pi\tau$. We conclude that a lifetime-broadened energy level with lifetime τ has an energy spread $\Delta E = h/2\pi\tau$, provided that the decay process can be modeled as a simple exponential. In this picture, spontaneous emission can be viewed in terms of a damped harmonic oscillator that generates an exponentially decaying harmonic function.

Hence, if the energy spreads of levels 1 and 2 are $\Delta E_1 = h/2\pi\tau_1$ and $\Delta E_2 = h/2\pi\tau_2$, respectively, the spread in the energy difference corresponding to the transition between the two levels is

$$\Delta E = \Delta E_1 + \Delta E_2 = \frac{h}{2\pi}\left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right) = \frac{h}{2\pi}\frac{1}{\tau}, \tag{4.6-1}$$

where τ is the transition lifetime and $\tau^{-1} = (\tau_1^{-1} + \tau_2^{-1})$. The corresponding spread of the transition frequency, which is called the lifetime-broadening linewidth, is therefore

$$\Delta\nu = \frac{1}{2\pi}\left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right). \tag{4.6-2}$$

Lifetime-Broadening Linewidth

This spread is centered about the frequency $\nu_0 = (E_2 - E_1)/h$, and the lineshape function has a Lorentzian profile:

$$g(\nu) = \frac{\Delta\nu/2\pi}{(\nu - \nu_0)^2 + (\Delta\nu/2)^2}. \tag{4.6-3}$$

Lorentzian Lineshape Function

EXAMPLE 4.6-1. ***Peak Transition Cross Section for Lorentzian Lineshape Function.*** The value of the Lorentzian lineshape function $g(\nu)$ specified in (4.6-3) at its central frequency ν_0 is

$$g(\nu_0) = 2/\pi\Delta\nu, \tag{4.6-4}$$

so that the peak transition cross section, in accordance with (4.4-12), is given by

$$\overline{\sigma}_0 = \frac{\lambda^2}{2\pi}\frac{1}{2\pi t_{\rm sp}\Delta\nu}. \tag{4.6-5}$$

Peak Cross Section

The largest transition cross section occurs under ideal conditions when the decay is entirely radiative so that $\tau_2 = t_{\rm sp}$ and $1/\tau_1 = 0$ (which is the case when level 1 is the ground state from which no decay is possible). From (4.6-2), we then have $\Delta\nu = 1/2\pi t_{\rm sp}$, whereupon

$$\overline{\sigma}_0 = \lambda^2/2\pi\,, \tag{4.6-6}$$

indicating that the peak cross section is of the order of one square wavelength. When level 1 is not the ground state, or when nonradiative transitions are significant, we have $\Delta\nu \gg 1/2\pi t_{\rm sp}$ in which case $\overline{\sigma}_0$ can be significantly smaller than $\lambda^2/2\pi$. For optical transitions in the range $\lambda = 0.1$ to $10\ \mu$m, calculated values of $\lambda^2/2\pi$ typically lie between 10^{-11} and 10^{-7} cm^2, whereas observed values of σ_0 generally fall in the range between 10^{-20} and 10^{-12} cm^2.

EXAMPLE 4.6-2. ***Sequence of Wavepackets Emitted by Atoms at Poisson Times.*** A sequence of wavepackets emitted by a collection of atoms at random times is a source of partially coherent light (Fig. 4.6-1). The frequencies of all wavepackets are assumed to be identical and their decays are assumed to arise from the finite atomic lifetimes. Each wavepacket is taken to have a random phase since it is emitted by a different atom.

Figure 4.6-1 Light comprising wavepackets emitted at Poisson times from a collection of atoms has a coherence time τ_c that is equal to the duration of a wavepacket. The magnitude of the complex degree of coherence $|g(\tau)|$ is a double-sided exponential, corresponding to a power spectral density $S(\nu)$ with Lorentzian form.

The individual wavepackets may, for example, be considered to be harmonic functions with an exponentially decaying envelope representing the atomic lifetime, so that at a given position a wavepacket emitted at $t = 0$ has the complex wavefunction

$$U_p(t) = \begin{cases} A_p \exp(-t/\tau_c) \exp(j2\pi\nu_0 t), & t \geqslant 0 \\ 0, & t < 0. \end{cases} \tag{4.6-7}$$

The independent random phases of the different emissions are included in A_p . The statistical properties of the overall wavefunction are determined by carrying out the appropriate averaging operations in accordance with the rules of mathematical statistics. This process yields a complex degree of coherence $g(\tau) = \exp(-|\tau|/\tau_c)\exp(j2\pi\nu_0\tau)$, whose magnitude is a double-sided exponential function. In accordance with the Wiener–Khinchin theorem (2.7-16), the corresponding power spectral density is $S(\nu) \propto (\Delta\nu/2\pi)/[(\nu - \nu_0)^2 + (\Delta\nu/2)^2]$, which is Lorentzian with $\Delta\nu = 1/\pi\tau_c$. The coherence time τ_c turns out to be the width of a wavepacket, signifying that the light is correlated over that time.

Homogeneous and Inhomogeneous Broadening

Lifetime broadening is an example of **homogeneous broadening**, in which the interacting atoms of a medium are all taken to be identical, with the same lineshape functions and center frequencies.

Some media exhibit **inhomogeneous broadening** as well, in which different subsets of interacting atoms exhibit different behavior, either because of differences in their local environment, different dynamical behavior, or different origins. A commonly encountered example of inhomogeneous broadening is Doppler broadening. As a result of the Doppler effect, an atom moving with velocity v along a given direction exhibits a lineshape function that is shifted by the frequency $\pm(v/c)\nu_0$ when viewed along that direction, where ν_0 is its central frequency.

For inhomogeneously broadened media, we can define an average lineshape function

$$\overline{g}(\nu) = \langle g_\beta(\nu) \rangle, \tag{4.6-8}$$

where $\langle\cdot\rangle$ represents an average with respect to the variable β, which labels the subset of atoms with the homogeneously broadened lineshape function $g_\beta(\nu)$. The average lineshape function is obtained by weighting the $g_\beta(\nu)$, which are known as *spectral packets*, by the fraction of the atomic population endowed with the property β, as pictured in Fig. 4.6-2.

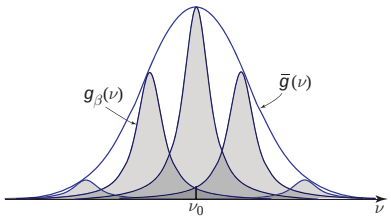


Figure 4.6-2 The average lineshape function $\overline{g}(\nu)$ for an inhomogeneously broadened collection of atoms. The underlying homogeneously broadened spectral packets are denoted $g_\beta(\nu)$.

Some atom–photon interactions exhibit broadening intermediate between pure homogeneous and pure inhomogeneous. Such mixed broadening can be modeled by making use of an intermediate lineshape function, such as the Voigt profile.

It will become apparent in Sec. 4.7 that the spectrum of blackbody radiation is inhomogeneously broadened as a result of the different center frequencies of the cavity modes. It will be seen in Chapter 6 that the spectrum of the spontaneous emission from an LED is also inhomogeneously broadened, in this case by virtue of the different center frequencies of the recombination photons as required by the Pauli exclusion principle.

4.7 BLACKBODY RADIATION

The term **blackbody** was initially introduced by Gustav Kirchhoff in 1860. Its definition, as updated in 1914 by Max Planck (p. 61), is an opaque object of arbitrary composition, in thermal equilibrium, that absorbs all incident radiation, at whatever wavelength and angle of incidence, and re-emits it. The quintessential example of a blackbody radiator is an opaque cavity at temperature T, with a small hole through which the radiation interior to the cavity can escape and be sampled. Blackbodies emit a universal form of isotropic radiation known as **blackbody radiation**. We proceed to investigate the spectrum and total power radiated by a blackbody by examining the interactions among a collection of photons and atoms in a cavity, in steady state and thermal equilibrium. In the course of our study, we obtain expressions for the Planck spectrum, Wien's law, and the Stefan–Boltzmann law.

Thermal Equilibrium Between Photons and Atoms

A macroscopic rate-equation approach that balances spontaneous emission, absorption, and stimulated emission, under conditions of thermal equilibrium, leads to the spectral energy density of blackbody radiation. We begin our analysis by considering (4.4-9) and (4.5-8), which govern spontaneous emission and induced transitions, respectively, in the presence of broadband light.

Consider a 3D closed cavity of unit volume whose walls consist of large numbers of atoms, each with two energy levels denoted 1 and 2, that are separated by an energy difference $h\nu$. The cavity, which is maintained at temperature T, supports broadband radiation that can be observed through a small hole. Let $N_1(t)$ and $N_2(t)$ represent the numbers of atoms per unit volume occupying energy levels 1 and 2, at time t, respectively. Since some of the atoms are initially in level 2, as ensured by the finite temperature, spontaneous emission creates radiation in the cavity. This radiation in turn can induce absorption and stimulated emission. The three processes coexist and it is assumed that steady-state (equilibrium) conditions are attained. We further assume that an average of $\overline{n}$ photons occupies *each* of the radiation modes whose frequencies lie within the atomic linewidth, as established in (4.5-8).

We first treat spontaneous emission alone. The probability that a single atom in the upper energy level undergoes spontaneous emission into any of the modes, within the time increment from t to $t + \Delta t$, is $P_{\text{sp}}\Delta t = \Delta t/t_{\text{sp}}$. There are $N_2(t)$ such atoms so that the average number of emitted photons within Δt is $N_2(t)\Delta t/t_{\text{sp}}$. This is also the number of atoms that depart from level 2 during the time interval Δt. Hence, the (negative) rate of increase of $N_2(t)$ arising from spontaneous emission is described by the differential equation

$$\frac{dN_2}{dt} = -\frac{N_2}{t_{\text{sp}}}. \tag{4.7-1}$$

The solution, $N_2(t) = N_2(0)\exp(-t/t_{\text{sp}})$, is an exponentially decaying function of time, as displayed in Fig. 4.7-1. Given sufficient time, the number of atoms in the upper level N_2 will decay to zero with time constant t_{sp}, the energy carried off by the spontaneously emitted photons.

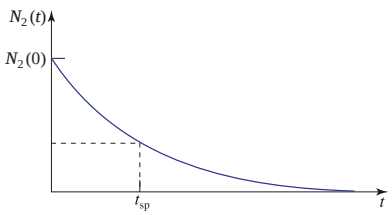


Figure 4.7-1 Decay of the upper-level population caused by spontaneous emission alone.

We now incorporate absorption and stimulated emission, which contribute to changes in the populations. Since there are N_1 atoms capable of absorption, the rate of increase of the population of atoms in the upper energy level arising from absorption is, based on (4.5-8),

$$\frac{dN_2}{dt} = N_1 W_i = \frac{\overline{n}\, N_1}{t_{\text{sp}}}\,. \tag{4.7-2}$$

Similarly, stimulated emission gives rise to a (negative) rate of increase of atoms in the upper state, expressed as

$$\frac{dN_2}{dt} = -N_2 W_i = -\frac{\overline{n}\, N_2}{t_{\text{sp}}}\,. \tag{4.7-3}$$

The rates of atomic absorption and stimulated emission are both proportional to $\overline{n}$, the average number of photons in each mode.

Combining (4.7-1), (4.7-2), and (4.7-3) to accommodate spontaneous emission, absorption, and stimulated emission together, yields the rate equation

$$\boxed{\frac{dN_2}{dt} = -\frac{N_2}{t_{\text{sp}}} + \frac{\overline{n}\, N_1}{t_{\text{sp}}} - \frac{\overline{n}\, N_2}{t_{\text{sp}}}\,.} \tag{4.7-4}$$

Rate Equation (Broadband Light)

This result ignores transitions into or out of level 2 that arise from extraneous effects, such as interactions with energy levels other than level 1, nonradiative transitions, and external sources of excitation. Steady-state operation demands that $dN_2/dt = 0$, which leads to

$$\frac{N_2}{N_1} = \frac{\overline{n}}{1+\overline{n}}\,, \tag{4.7-5}$$

which is clearly $\leqslant 1$. In addition to the requirement for steady-state operation, we now impose the requirement that the two energy states are in thermal equilibrium, and posit that their populations approximately obey the Boltzmann distribution (4.2-2):

$$\frac{N_2}{N_1} = \exp\left(-\frac{E_2 - E_1}{kT}\right) = \exp\left(-\frac{h\nu}{kT}\right). \tag{4.7-6}$$

Substituting (4.7-6) into (4.7-5) then leads to the following expression for the mean number of photons per mode near frequency ν:

$$\overline{n} = \frac{1}{\exp(h\nu/kT) - 1}\,. \tag{4.7-7}$$

The foregoing derivation is predicated on the interaction of two energy levels coupled by absorption, stimulated emission, and spontaneous emission, at a frequency near ν. Its applicability is, however, far broader. This may be understood by considering a cavity whose walls are made of solid materials that possess a continuum of energy levels at all energy separations, and therefore all values of ν. Atoms in the walls spontaneously emit into the cavity. The emitted light subsequently interacts with the atoms in the walls, giving rise to absorption and stimulated emission. If the walls are maintained at temperature T, the combined system of atoms and radiation reaches thermal equilibrium, whatever the nature of the walls and whatever the cavity shape.

Equation (4.7-7) is identical to (4.2-9), the expression for the mean photon number in a mode of thermal light for which the occupation of the modal energy levels follows the distribution $p(n) \propto \exp(-nh\nu/kT)$. This indicates that our analysis is self-consistent. Photons interacting with atoms in thermal equilibrium at temperature T are themselves in thermal equilibrium as a photon gas at the same temperature T.

Blackbody Radiation Spectrum

Based on the foregoing discussion, the average energy $\overline{E}$ of a blackbody radiation mode is $\overline{n}h\nu$, where $\overline{n}$ is given by (4.7-7), so that

$$\overline{E} = \frac{h\nu}{\exp(h\nu/kT) - 1}. \tag{4.7-8}$$

Average Energy
(Mode in Thermal Equilibrium)

The dependence of $\overline{E}$ on ν, which is the same as that set forth in (4.2-15), is portrayed in Fig. 4.7-2.

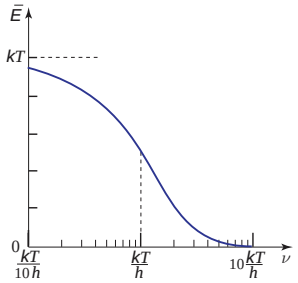


Figure 4.7-2 Semilogarithmic plot of the average energy $\overline{E}$ of an electromagnetic mode in thermal equilibrium at temperature T, as a function of the modal frequency ν. Plots of $\overline{E}$ vs. photon energy $h\nu$ for $T = 300$ K ($kT = 0.026$ eV) and for $T = 600$ K ($kT = 0.052$ eV) are presented in Fig. 4.2-5.

Multiplying $\overline{E}$ in (4.7-8) (average energy per mode) by the 3D modal density $M(\nu) = 8\pi\nu^2/c^3$ provided in (4.4-5) (number of modes per unit frequency per unit cavity volume) gives rise to the spectral energy density $\varrho_\nu(\nu) = M(\nu)\overline{E}$ (energy per unit frequency bandwidth per unit cavity volume):

$$\varrho_\nu(\nu, T) = \frac{8\pi h\nu^3}{c^3} \frac{1}{\exp(h\nu/kT) - 1}. \tag{4.7-9}$$

Blackbody Spectral Energy Density
(Frequency Parameterization)

This formula, known as the **blackbody radiation spectrum**, and also as **Planck's radiation law**, is sketched as a function of frequency on linear coordinates in Fig. 4.7-3. The formal plot presented in Fig. 4.7-4 is the iconic representation of (4.7-9) on logarthmic coordinates, with temperature as a parameter. As the blackbody temperature increases, the mean number of photons per mode increases in accordance with (4.7-7). Photons can emerge from, or disappear into, the walls of the cavity. The radiation is unpolarized since the medium is assumed to be isotropic. Though bosons, the photons in a blackbody cavity are not conserved and therefore do not form a Bose–Einstein condensate.

Entropy. The Planck radiation law is the unique maximum entropy energy distribution for a photon gas, much as the Maxwell–Boltzmann distribution is the unique maximum entropy energy distribution for a gas of material particles. While the properties of a photon gas depend only on temperature, however, those of a material gas depend on the masses and numbers of particles, as well as on temperature.

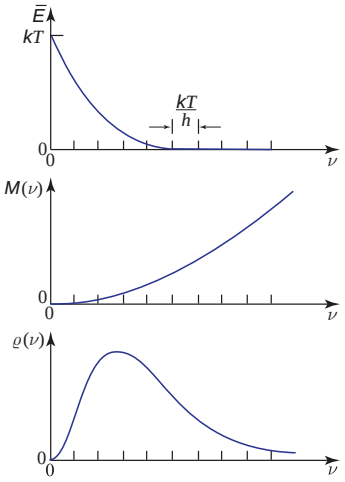


Figure 4.7-3 Frequency dependence of the energy per mode $\bar{E}$, the density of modes $M(\nu)$, and the spectral energy density $\varrho_\nu(\nu) = M(\nu)\bar{E}$ for blackbody radiation, on linear coordinates.

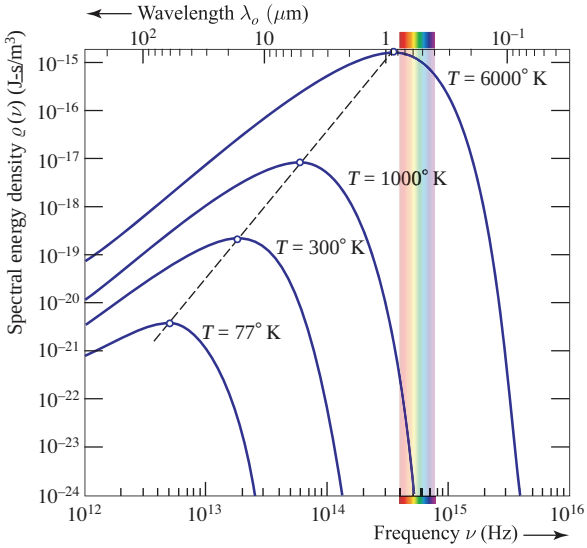


Figure 4.7-4 Dependence of the blackbody spectral energy density $\varrho_\nu(\nu)$ on the frequency ν for several different temperatures, on logarithmic coordinates. The scale on the upper abscissa is the wavelength $\lambda_0 = c_0/\nu$.

Failure of the Equipartition Theorem. The formula for the spectrum of blackbody radiation played a central role in the discovery of the photon nature of light. Based on classical electromagnetic theory, the modal density for a three-dimensional cavity was known to be $M(\nu) = 8\pi\nu^2/c^3$, as provided in (4.4-5). Moreover, in the context of classical statistical mechanics, the equipartition theorem discussed in Sec. 4.1 had long dictated that the average energy per mode was fixed at $\bar{E} = kT$, independent of the modal frequency. This led to a theoretical expression for the blackbody spectrum, $\varrho_\nu(\nu) = M(\nu)\bar{E} = 8\pi\nu^2 kT/c^3$, a result known as the **Rayleigh–Jeans formula**. However, this formula disagreed with experiment in the ultraviolet region and at frequencies beyond, and its integral with respect to frequency diverged. The failure was considered to be so profound that it was named the *ultraviolet catastrophe* by the physicist Paul Ehrenfest.

Max Planck resolved the conundrum in 1900 by allowing the energy levels of the atoms in the walls of the cavity to be quantized. That led to an expression for the average energy per mode given by (4.7-8) [or (4.2-15)] and thence to the correct blackbody spectral energy density (4.7-9). The energy quantization effectively gave rise to a progressive *chilling out* of the modes as their frequency increased, as is apparent in the top panel of Fig. 4.7-3, thereby averting the ultraviolet catastrophe.

The Rayleigh–Jeans formula is recovered from (4.7-9) in the high-temperature (classical) limit $h\nu \ll kT$ by using the Taylor-series approximation $\exp(h\nu/kT) \approx 1 + h\nu/kT$ in (4.7-8), which yields $\bar{E} = kT$, in accord with the classical equipartition theorem. Einstein subsequently extended Planck's explanation by proposing that the quantization imposed on the atomic energy levels be imposed directly on the electromagnetic modal energies, a leap that helped solidify the concept of the photon.

Wavelength Parameterization for the Blackbody Radiation Spectrum. The spectral energy density $\varrho_\nu(\nu)$ set forth in (4.7-9) is readily transformed into the spectral energy density $\varrho_\lambda(\lambda)$, parameterized in terms of wavelength, where $\varrho_\lambda(\lambda)\,d\lambda$ represents the energy per unit volume in the wavelength region between λ and $\lambda + d\lambda$. Since $\nu =$

c/λ, the two densities are related by $\varrho_\lambda(\lambda)\,d\lambda = \varrho_\nu(\nu)\,d\nu$, so that $\varrho_\lambda = \varrho_\nu\,|d\nu/d\lambda| = \varrho_\nu(\lambda)\cdot c/\lambda^2$. This yields $\varrho_\lambda(\lambda) = [8\pi h(c/\lambda)^3/c^3]\cdot[1/(\exp(hc/\lambda kT)-1)]\cdot[c/\lambda^2]$, which provides

$$\boxed{\varrho_\lambda(\lambda,T) = \frac{8\pi hc}{\lambda^5}\,\frac{1}{\exp(hc/\lambda kT)-1}\,,} \qquad \begin{array}{r}(4.7\text{-}10)\\ \text{Blackbody Spectral Energy Density}\\ \text{(Wavelength Parameterization)}\end{array}$$

with units of $\mathrm{J\cdot nm^{-1}\cdot m^{-3}}$.

All things being equal, the frequency parameterization set forth in (4.7-9) is preferred to the wavelength parameterization of (4.7-10) since the frequency is invariant as the radiation passes through media of different refractive indices. However, it turns out that the wavelength-based spectral radiance $L_\lambda(\lambda,T) = (c/4\pi)\varrho_\lambda(\lambda,T)$, set forth in (9.7-1) and plotted in Fig. 9.7-1(a), enjoys widespread use in radiometry and LED lighting.

EXAMPLE 4.7-1. ***Frequency and Wavelength of Maximum Spectral Energy Density.***

(a) *Peak frequency* ν_p: Substituting $x = h\nu/kT$ into the blackbody radiation law specified in (4.7-9) provides $\varrho_\nu(x) = [8\pi(kT)^3/c^3h^2][x^3/(e^x-1)]$. The frequency ν_p at which $\varrho_\nu(\nu)$ is maximized is obtained by setting the derivative $d\varrho_\nu(x)/dx$ equal to zero, which yields $x = 3(1-e^{-x})$, where $x = h\nu_\mathrm{p}/kT$. Numerical solution of this nonlinear equation provides $x \approx 2.821$. The peak frequency is therefore $\nu_\mathrm{p} = xkT/h$, which, at $T = 6000$ K, corresponds to $\nu_\mathrm{p} \approx 352$ THz, comporting with the lower abscissa of Fig. 4.7-4. The corresponding wavelength is given by $c_0/\nu_\mathrm{p} \approx 850$ nm, which comports with the upper abscissa in this figure.

(b) *Peak wavelength* λ_p: Substituting $y = hc/\lambda kT$ into the form of the blackbody radiation law specified in (4.7-10) leads to $\varrho_\lambda(y) = [8\pi(kT)^5/c^4h^4][y^5/(e^y-1)]$. The wavelength λ_p at which $\varrho_\lambda(\lambda)$ is maximized is determined by setting the derivative $d\varrho_\lambda(y)/dy = 0$. This then leads to $y = 5(1-e^{-y})$, where $y = hc/\lambda_\mathrm{p}kT$. Numerically solving this nonlinear equation provides $y \approx 4.965$. The peak wavelength is therefore $\lambda_\mathrm{p} = hc/ykT$, which, at $T = 6000$ K, corresponds to $\lambda_\mathrm{p} \approx 483$ nm.

(c) *Some features of the spectral energy density depend on the parameterization employed*: The frequency-based spectral density (4.7-9) and the wavelength-based spectral density (4.7-10) have distinct functional forms; their ratio is given by $\varrho_\lambda(\lambda)/\varrho_\nu(\nu) = c/\lambda^2$. Forming a product of the expressions for the frequency and wavelength maxima, $\nu_\mathrm{p} = xkT/h$ and $\lambda_\mathrm{p} = hc/ykT$, respectively, yields

$$\lambda_\mathrm{p}\nu_\mathrm{p} = (x/y)\,c \neq c\,, \tag{4.7-11}$$

where $x/y \approx 2.821/4.965 \approx 0.568$. The numerical calculations carried out in parts (a) and (b) above are in accord with $\lambda_\mathrm{p} = (x/y)\,(c/\nu_\mathrm{p})$, as provided in (4.7-11): At $T = 6000$ K, the wavelength corresponding to the peak frequency was calculated in (a) to be $c/\nu_\mathrm{p} \approx 850$ nm, while the actual peak wavelength was determined in (b) to be $\lambda_\mathrm{p} \approx 483$ nm, and indeed $483\ \mathrm{nm} \approx 0.568 \times 850$ nm. The shapes, peak locations, and some other features of the two density functions clearly depend on the parameterization selected, although other features do not. In particular, at any given temperature, the integral of the wavelength density over the interval $[\lambda_1,\lambda_2]$ returns a value that is the same as the integral of the frequency density over the corresponding interval $[c/\lambda_2, c/\lambda_1]$.

EXAMPLE 4.7-2. ***Total Blackbody Energy Density Per Unit Volume.*** The total energy density (per unit cavity volume) of a blackbody is determined by integrating the frequency-parameterized spectral energy density (4.7-9) over all frequencies, or, equivalently, by integrating the wavelength-parameterized spectral energy density (4.7-10) over all wavelengths. For the frequency-parameterized case, use of the substitution $x = h\nu/kT$ in (4.7-9) leads to

$$\int_0^\infty \varrho_\nu(\nu,T)\,d\nu = \frac{8\pi h}{c^3}\int_0^\infty \frac{\nu^3\,d\nu}{\exp(h\nu/kT)-1} = \frac{8\pi k^4T^4}{c^3h^3}\int_0^\infty \frac{x^3}{e^x-1}\,dx = \frac{8\pi^5k^4T^4}{15c^3h^3}\,, \tag{4.7-12}$$

since the definite integral $\int_0^\infty dx\, x^3/(e^x-1) = \pi^4/15$. For the wavelength-parameterized case, using the substitution $y = hc/\lambda kT$ in (4.7-10) yields the identical result,

$$\int_0^\infty \varrho_\lambda(\lambda, T)\, d\lambda = 8\pi hc \int_0^\infty \frac{\lambda^{-5}\, d\lambda}{\exp(hc/\lambda kT) - 1} = \frac{8\pi k^4 T^4}{c^3 h^3} \int_0^\infty \frac{y^3}{e^y - 1}\, dy = \frac{8\pi^5 k^4 T^4}{15 c^3 h^3}. \quad (4.7\text{-}13)$$

Wien's Law

The expression for the peak wavelength of the blackbody energy density provided in Example 4.7-1(b), $\lambda_p = hc/ykT$, with $y \approx 4.965$, establishes that the product of the peak wavelength and temperature of a blackbody radiator is given by

$$\boxed{\lambda_p T = b,} \quad (4.7\text{-}14)$$
Wien's Law

where the constant $b \equiv hc/yk \approx 2.90 \times 10^6$ nm·K is known as **Wien's constant**. Equation (4.7-14), which was established a number of years before Planck developed his general formula, is known as **Wien's law**. Versions of this law can also be fashioned using wavelength markers other than the peak wavelength (e.g., the median wavelength).

Stefan–Boltzmann Law

The **Stefan–Boltzmann law** characterizes the temperature dependence of the power radiated by a blackbody per unit area of its surface, P/A, which increases with temperature as T^4. This expression is derived by multiplying the total energy density per unit volume provided in (4.7-12) [or in (4.7-13)] by the factor $c/4$, which serves to convert energy density to power per unit area, resulting in $P/A = (c/4)(8\pi^5 k^4 T^4/15c^3h^3) = 2\pi^5 k^4 T^4/15c^2h^3$. The Stefan–Boltzmann law is usually cast in the form

$$\boxed{P/A = \sigma_{SB} T^4,} \quad (4.7\text{-}15)$$
Stefan–Boltzmann Law
(Blackbody Radiation)

where $\sigma_{SB} \equiv 2\pi^5 k^4/15c^2h^3 \approx 5.67 \times 10^{-8}$ W·m^{-2}·K^{-4} is known as the **Stefan–Boltzmann constant**.

4.8 THERMAL RADIATION

The presentation provided in Sec. 4.7 demonstrated that a collection of atoms and radiation in a cavity behaves as a blackbody when steady state and thermal equilibrium prevail. As illustrated in Fig. 4.8-1(a), a blackbody radiator is an opaque object that absorbs all of the radiation incident upon it and re-emits it isotropically as blackbody radiation, with a spectrum that depends only on the temperature of the object. Based on conservation of energy, Kirchhoff established in 1906 that the ability of a blackbody radiator to absorb radiant energy is equal to its ability to radiate energy.

Because the blackbody spectrum depends only on the temperature of the object, as specified in (4.7-9), the terms **thermal radiation** (*Wärmestrahlung* in German) and **thermal light** (thermal radiation whose spectrum lies principally in the visible region)

are often used as synonyms for blackbody radiation. While all objects emit thermal radiation as a result of their finite temperatures, not all objects emit blackbody radiation, as will become clear in the sequel.

We begin by considering graybody radiation, which has properties that are closely related (but not identical) to blackbody radiation. We then discuss the properties of incandescent light, which can be approximated by graybody radiation in the visible region. We follow this by introducing a technique called thermography, which highlights the features of various sources of thermal radiation. Thermography uses the thermal radiation emitted by an object to generate a self-temperature image. The reader is asked to take notice that in practice all three terms, *blackbody radiation*, *graybody radiation*, and *thermal radiation*, are often used interchangeably. Finally, we present the photon-number statistics for a beam of thermal light.

Graybody Radiation

Real-world objects are seldom blackbody radiators; far more often they behave as **graybody radiators**. As portrayed in Fig. 4.8-1(b), a graybody radiator partially reflects, and partially transmits, portions of the radiation incident upon it, but it mimics a blackbody radiator in that it emits all of the energy that it absorbs. A generalization of Kirchhoff's 1860 radiation law accommodates such partially transparent and partially reflecting bodies.

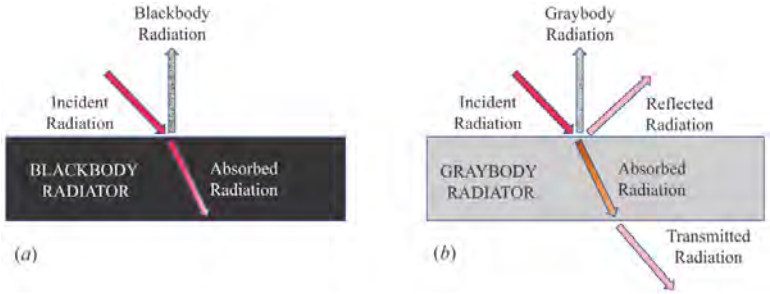


Figure 4.8-1 Radiation flow for a blackbody radiator and a graybody radiator. (a) Blackbody radiation is characterized by a spectral energy density and power per unit surface area that depend only on the thermodynamic temperature of the radiator. Refraction without reflection can be achieved by making use of metamaterials. (b) Graybody radiation has a spectral energy density and power per unit surface area that matches that of blackbody radiation, but with magnitudes that are reduced by the emissivity of the graybody radiator ($\varepsilon < 1$). The light emitted by a heated tungsten-filament incandescent lamp is well approximated by graybody radiation.

Stefan–Boltzmann Law. The radiative properties of a graybody radiator can be characterized by incorporating into the Stefan–Boltzmann law (4.7-15) a dimensionless multiplicative constant called the graybody **emissivity** ε,

$$P/A = \varepsilon\sigma_{\text{SB}}T^4, \tag{4.8-1}$$

Stefan–Boltzmann Law (Graybody Radiation)

where ε is defined as the ratio of the power emitted by the graybody radiator to the power that would be emitted by a blackbody radiator of the same temperature. The absorptivity α $(0 < \alpha < 1)$ of a graybody radiator is equal to its emissivity ε $(0 < \varepsilon < 1)$, and both are defined to be constant over the wavelength region of interest. Good absorbers are therefore good emitters and poor absorbers are poor emitters. The spectral energy density of a graybody radiator is proportional to its blackbody counterpart, as specified

in (4.7-9) and (4.7-10); the emissivity ε serves as the constant of proportionality. Wien's law (4.7-14) remains intact. By definition, $\alpha = \varepsilon = 1$ for blackbody radiators.

Graybody Emissivity. Different materials and objects exhibit different emissivities, as reported in Table 4.8-1.

The emissivity ε is, in general, a function of wavelength, temperature, and angle-of-view, but it is taken to be independent of these parameters in modeling graybody radiation. Highly reflective objects are poor absorbers, and thus poor emitters, so they have low emissivities ($\varepsilon \approx 1 - \mathcal{R}$). Most natural surfaces on earth have emissivities in the range $0.6 \lesssim \varepsilon \lesssim 1.0$; deserts have values that lie toward the lower limit. The emissivities of astronomical bodies are quite close to unity, and the spectral density of the radiation they emit roughly follows (4.7-10). It is straightforward to estimate the effective temperatures of astronomical bodies by making use of the Stefan–Boltzmann law (4.8-1). It has been established, for example, that the effective temperatures of Mars, Earth, and the Sun are roughly 200 K, 300 K, and 5800 K, respectively.

Table 4.8-1 Representative emissivities of materials and objects.

MATERIAL/OBJECT	ε
Silver (Polished)	0.02
Aluminum (Foil)	0.03
Tungsten Filament (Heated)	0.44
Paper (White)	0.84
Aluminum (Anodized)	0.85
Earth (Surface)	0.85
Snow	0.85
Soil	0.92
Glass (Uncoated)	0.95
Vegetation	0.95
Water	0.96
Ice	0.97
Graphite (Powdered)	0.97
Sun	0.99

Universality of Planck's Radiation Law. The Planck spectral energy density provided in (4.7-9) obeys a universal form that characterizes blackbodies, graybodies, and the radiation emitted by physical objects in local thermal equilibrium. As discussed in the foregoing, such objects are ubiquitous, ranging from the iconic oven-and-hole construction; to the earth and the entities that inhabit it, which emit in the middle-infrared region; to the planets, stars, and galaxies; to black holes; and ultimately to the universe itself. The faint cosmic microwave background radiation that fills all of space in this, the stelliferous era, is a remnant of the primordial era of the universe. Its spectrum is that of a near-perfect blackbody with a peak wavelength $\lambda_p \approx 1.063$ mm, corresponding to a temperature ≈ 2.725 K in accordance with Wien's law (4.7-14).

Incandescent Light

Incandescent light is generated by the transitions of free and valence electrons in hot solid materials. The term derives from the Latin verb *incandescere*, which means to "glow." First observed by Sir Humphry Davy in 1802 using a strip of platinum, incandescence provided the first practical means for generating light from an electric current.

The quintessential example of an incandescent source is a glass light bulb containing a thin filament that is ohmically heated by an electric current. Although carbon was initially used by Edison, tungsten is generally the material of choice because it has the lowest vapor pressure (1 Pa at 3477 K) and highest melting point (3695 K) of any metal. The optical reflectance of a heated tungsten filament is $\mathcal{R} \approx 0.55$ and its emissivity is constant at $\varepsilon \approx 1 - \mathcal{R} \approx 0.44$ across the visible region (Table 4.8-1). Hence, a tungsten incandescent lamp radiates as a graybody in the visible with a spectrum that is well-described by the Planck blackbody radiation law (4.7-10). The tungsten-filament lamp is not a perfect graybody when examined over the visible *and* infrared regions, however. While its emissivity $\varepsilon \approx 0.44$ in the visible, it is slightly lower ($\varepsilon \approx 0.33$) in the infrared. Hence, its spectrum does not follow a unique Planckian radiation curve

over this extended spectral region. The incandescence generated by the metallic-oxide gas mantles used in street lighting in the late nineteenth century behaved similarly.

Halogen lamps are incandescent sources whose transparent envelopes contain a small quantity of halogen gas such as bromine. The halogen and tungsten atoms chemically react and, in a process known as the *halogen cycle*, the evaporated tungsten is redeposited on the filament when the halogen cools. This increases the lifespan and efficiency of the lamp, and diminishes the darkening of its envelope.

Incandescent lighting was ubiquitous throughout the twentieth century. It has been highly prized for illumination because its Planckian spectrum is reminiscent of that of sunlight. Aside from ideal color rendering, it has other advantages: simple construction, the ability to operate on AC or DC current, and insensitivity to ambient temperature. However, incandescent lamps are highly inefficient as visible radiators: about 5% of the energy consumed typically emerges as visible light and the remainder is dissipated in the form of infrared radiation. As displayed in Fig. 9.7-1(a), the limited emission in the visible, which lends a reddish tinge to the light, results from the fact that the tungsten filament cannot be heated to a temperature $\gtrsim 3300$ K lest it melt.

Scattered efforts have been made over the years to increase the efficiency of incandescent lamps by making use of techniques that selectively modify the wavelength dependence of the emissivity, or by modifying the filament temperature, fill gas, or bulb reflectance. Nevertheless, as will be discussed in Chapter 10, LED lighting began to make serious inroads in replacing incandescent lighting in about 2010 because of its substantially higher efficiency, along with its environmental friendliness and other manifold merits. Indeed, the worldwide transition from incandescent to LED lighting has continued apace and is approaching its finale.

Thermal Radiation

Many real surfaces have emissivities that are wavelength-dependent. Thermal radiation is sometimes defined as the radiation emitted by bodies in local thermal equilibrium whose emissivity is a function of wavelength ($0 < \varepsilon(\lambda) \leqslant 1$). In accordance with this definition, thermal radiation subsumes blackbody radiation ($\varepsilon = 1$) and graybody radiation ($0 < \varepsilon < 1$) as special cases; both have emissivities that are independent of wavelength. The emissivity of some materials also depends on temperature and angle-of-view, but these dependencies are frequently small and are often ignored.

Thermography. Some of the characteristics of thermal radiation are illustrated by considering a technique called **thermography**, whereby a thermal object is imaged by means of its infrared self-radiation. An image, or **thermograph**, of the temperature distribution across the thermal object or scene is generated by making use of a camera that is sensitive in the wavelength region of the object's peak thermal emissions. Wien's law (4.7-14) dictates that the peak wavelength of a radiating blackbody is inversely proportional to its temperature. Bodies of moderate temperature, including earthly objects such as humans, typically radiate at wavelengths in the mid infrared, whereas cold objects radiate at longer wavelengths that stretch into the far infrared.

The technique makes use of a **thermal camera**, also called an **infrared camera** or **thermographic camera**, that contains an array detector whose elements are sensitive in a particular spectral region of the infrared. Array detectors serving different wavelength regions are fabricated from materials whose detection efficiencies are high in the particular region of interest. Overall, thermography is useful over a broad wavelength range that spans $0.7\ \mu\text{m} \leqslant \lambda_0 \leqslant 300\ \mu\text{m}$, roughly corresponding to temperatures that stretch over the range $10\ \text{K} \leqslant T \leqslant 4000\ \text{K}$.

The photosensitive elements that comprise the array detector in a thermal camera do not resolve the wavelengths of the incident radiation, but rather detect the infrared power radiated by the corresponding pixels of the object. In accordance with the Stefan–Boltzmann law (4.8-1), this power varies as

$$P/A = \varepsilon(\mathbf{r})\,\sigma_{\mathrm{SB}} T^4(\mathbf{r}). \tag{4.8-2}$$

The locations of the pixels are designated by the position vector $\mathbf{r}$ since thermography is designed to garner information about objects and scenes that exhibit spatial temperature variations. Each pixel is assumed to be in local thermal equilibrium (thermal quasi-equilibrium) and is characterized by a unique temperature over the duration of the measurement. Clearly, the higher-temperature pixels in the object radiate more strongly than their lower-temperature counterparts, thereby generating larger responses in the corresponding array-detector elements. Moreover, the different pixels of the object generally consist of different materials, so the emissivity also depends on position. However, (4.8-2) discloses that the dependence of P/A on the emissivity is linear, whereas the dependence of P/A on temperature is far stronger, varying as $T^4(\mathbf{r})$.

Thermography finds use in industrial applications, such as monitoring the overheating of circuit boards and the evolution of oil spills. It is of assistance in search-and-rescue missions for humans and animals. It is also useful in clinical medicine since skin-surface temperature is a diagnostic for blood-flow blockages and tumors. Environmental applications include fire-fighting and forestry. The technique is invaluable in astronomy and cosmology since it allows astronomical objects, such as cooler red stars and red giants, to be imaged in the near infrared; planets, comets, and asteroids to be imaged in the mid-infrared; and central galactic regions, cold dust emissions, and early stars and galaxies to be imaged in the far infrared.

Representative thermographs are presented in Fig. 4.8-2 to illustrate the broad range of temperatures that can be accessed using this technique. The temperature of the object is represented in terms of a false-color palette that spans the visual spectrum. Conventionally, the coldest portions of the image are portrayed as black or violet and the warmest portions as red or white. This is the mapping used in Fig. 4.8-2(a) and (b). However, this convention is not always followed, since the palette is arbitrary, and indeed the opposite color convention is used in Fig. 4.8-2(c).

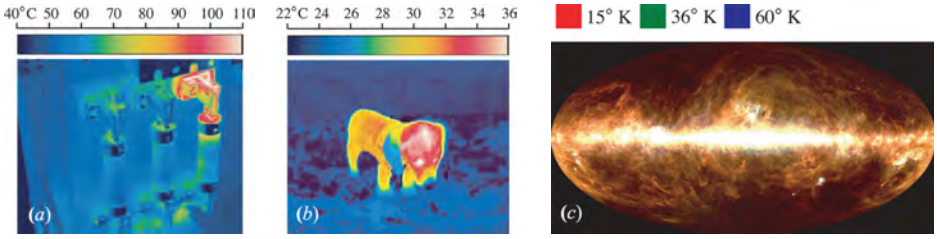


Figure 4.8-2 Representative thermographs in different temperature regions. The scales above the images relate temperature to false color. (a) Thermography in *industrial-systems analysis*. The red and white false colors reveal overheating of the right-most fuse, indicating a poor connection with its holder. (b) Thermography in *search-and-rescue*. The puppy emits radiation in the mid-infrared and can be located at night, even when concealed in foliage. It radiates most strongly where the fur is thinnest and body heat can escape. The eyes, covered by neither fur nor skin, are particularly visible. The nose, ears, and paws, in contrast, are peripheral to the animal's warm central body and therefore exhibit the lowest temperatures. (c) Thermography in *astronomy and cosmology*. A full-sky view collected by NASA's Cosmic Background Explorer (COBE) satellite, represented as a composite of images collected at three far-infrared wavelengths: $\lambda_0 = 240$ μm ($T = 15$ K, red); 100 μm ($T = 36$ K, green); and 60 μm ($T = 60$ K, blue). The prominent white horizontal band is interstellar dust in the plane of the Milky Way galaxy, the center of which is located at the center of the image.

While thermographs display temperature across an array of pixels, as illustrated in Fig. 4.8-2, a technique known as **hyperspectral thermography** simultaneously captures the spectrum of each pixel across a range of wavelengths.

Photon-Number Statistics for a Beam of Thermal Light

In Sec. 4.2, we investigated the statistical behavior of a collection of photons in thermal equilibrium in a cavity. The single-mode and multimode cases were characterized by Bose–Einstein and negative-binomial photon-number distributions, respectively.

Thermal light in the form of a beam is more conveniently analyzed as a doubly stochastic photon-number distribution, which is discussed in Sec. 3.6. It is well-established that thermal light may be modeled as a random wave with real and imaginary complex-field amplitude components that are Gaussian, independent, and identically distributed. The integrated-intensity probability density function for the single-mode case, considered in Example 3.6-1, is applicable for partially coherent light whose spectral width is sufficiently narrow such that $T/\tau_c \ll 1$, where T is the counting time and τ_c is the coherence time. Its Poisson transform is the Bose–Einstein photon-number distribution (4.2-8).

The integrated-intensity probability density function for the multimode case, with $\mathcal{M}$ identical thermal modes, is established by carrying out an $\mathcal{M}$-fold self-convolution of the exponential density provided in (3.6-14). This yields the chi-square (gamma) density function

$$p(w) = \frac{1}{(\mathcal{M}-1)!}\left(\frac{\langle w\rangle}{\mathcal{M}}\right)^{-\mathcal{M}} w^{\mathcal{M}-1}\exp\left(-\frac{w}{\langle w\rangle/\mathcal{M}}\right), \qquad w \geqslant 0, \tag{4.8-3}$$

with *overall* mean $\langle w\rangle$ and *overall* variance $\langle w\rangle^2/\mathcal{M}$. Equation (4.8-3) turns out to provide an excellent approximation for the integrated intensity for arbitrary values of T/τ_c, provided that $\mathcal{M}$ is appropriately chosen. For the single-mode case, $T/\tau_c \ll 1$ and $\mathcal{M} = 1$, as indicated above. When the number of modes is large, we have $T/\tau_c \gg 1$ and $\mathcal{M} \approx T/\tau_c$. The specific functional form of $\mathcal{M}$ for intermediate values of T/τ_c depends on the spectrum of the light.

The photon-number distribution is established by carrying out the Poisson transform of (4.8-3) via (3.6-11) or, alternatively, by performing an $\mathcal{M}$-fold self-convolution of the Poisson transform of the single-mode density function set forth in (3.6-14), which is the Bose–Einstein photon-number distribution provided in (4.2-8). Either way, the result is the negative-binomial distribution specified in (4.2-16). The mean and variance of this distribution, provided in (4.2-17) and (4.2-18), respectively, can also be obtained by making use of the chi-square mean and variance provided above, and using the general formulas (3.6-12) and (3.6-13) for the statistics of doubly stochastic photon-number distributions.

BIBLIOGRAPHY

Thermal and Statistical Physics

N. Sator, N. Pavloff, and L. Couëdel, *Statistical Physics*, CRC Press/Taylor & Francis, 2024.

C. Heissenberg and A. Sagnotti, *Classical and Quantum Statistical Physics: Fundamentals and Advanced Topics*, Cambridge University Press, 2022.

S. Sharma, *Thermal and Statistical Physics: Concepts and Applications*, Springer, 2022.

D. V. Schroeder, *An Introduction to Thermal Physics*, Oxford University Press, 2021.

H. Gould and J. Tobochnik, *Statistical & Thermal Physics: With Computer Applications*, Princeton University Press, 2nd ed. 2021.

R. K. Pathria and P. D. Beale, *Statistical Mechanics*, Academic/Elsevier, 4th ed. 2021.

L. E. Reichl, *A Modern Course in Statistical Physics*, Wiley–VCH, 4th ed. 2016.

J. P. Casquilho and P. I. C. Teixera, *Introduction to Statistical Physics*, Cambridge University Press, 2015.

H. J. W. Müller-Kirsten, *Basics of Statistical Physics*, World Scientific, 2nd ed. 2013.

M. J. R. Hoch, *Statistical and Thermal Physics: An Introduction*, CRC Press/Taylor & Francis, 2011.

A. L. Wasserman, *Thermal Physics: Concepts and Practice*, Cambridge University Press, 2011.

S. J. Blundell and K. M. Blundell, *Concepts in Thermal Physics*, Oxford University Press, 2nd ed. 2010.

C. Kittel, *Elementary Statistical Physics*, Wiley, 1958; Dover, reissued 2004.

Quantum Mechanics, Atomic Physics, and Solid-State Physics

A. J. Larkoski, *Quantum Mechanics: A Mathematical Introduction*, Cambridge University Press, 2023.

M. J. Everitt, K. N. Bjergstrom, and S. N. A. Duffus, *Quantum Mechanics: From Classical Analytical Mechanics to Quantum Mechanics, Simulation, Foundations & Engineering*, Wiley, 2023.

H. Năstase, *Quantum Mechanics: A Graduate Course*, Cambridge University Press, 2023.

B. Zwiebach, *Mastering Quantum Mechanics: Essentials, Theory, and Applications*, MIT Press, 2022.

P. Hofmann, *Solid State Physics: An Introduction*, Wiley–VCH, 3rd ed. 2022.

S. Hunklinger and C. Enss, *Solid State Physics*, Walter de Gruyter, 2022.

P. R. Berman, *Introductory Quantum Mechanics: A Traditional Approach Emphasizing Connections with Classical Physics*, Springer, 2018.

M. Dresselhaus, G. Dresselhaus, S. B. Cronin, and A. G. S. Filho, *Solid State Properties: From Bulk to Nano*, Springer, 2018.

M. G. Raymer, *Quantum Physics: What Everyone Needs to Know*, Oxford University Press, 2017.

H. Friedrich, *Theoretical Atomic Physics*, Springer, 4th ed. 2017.

J. Greensite, *An Introduction to Quantum Theory*, IOP Publishing, 2017.

J. B. Ketterson, *The Physics of Solids*, Oxford University Press, 2016.

S. Weinberg, *Lectures on Quantum Mechanics*, Cambridge University Press, 2nd ed. 2015.

L. E. Ballentine, *Quantum Mechanics: A Modern Development*, World Scientific, 2nd ed. 2015.

L. Susskind and A. Friedman, *Quantum Mechanics: The Theoretical Minimum*, Basic/Perseus, 2014.

E. D. Commins, *Quantum Mechanics: An Experimentalist's Approach*, Cambridge University Press, 2014.

C. Kittel, *Introduction to Solid State Physics*, Wiley, 8th ed. 2012.

D. A. B. Miller, *Quantum Mechanics for Scientists and Engineers*, Cambridge University Press, 2008.

M. Schwoerer and H. C. Wolf, *Organic Molecular Solids*, Wiley–VCH, 2007.

W. Demtröder, *Molecular Physics: Theoretical Principles and Experimental Methods*, Wiley, 2005.

M. Born, *Atomic Physics*, Blackie & Son, 1935, 8th ed. 1969; Dover, reissued 1989.

D. ter Haar, *The Old Quantum Theory*, Pergamon, 1967 [contains English translations of key early papers by Planck, Einstein, Rutherford, and Bohr].

R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Volume 3, *Quantum Mechanics*, 1965 and Volume 1, *Mainly Mechanics, Radiation, and Heat*, 1963, Addison–Wesley, 2nd ed. 2005.

L. D. Landau and E. M. Lifshitz, *Quantum Mechanics*, Addison–Wesley, 1958.

P. A. M. Dirac, *The Principles of Quantum Mechanics*, Oxford University Press, 4th ed. 1958.

E. U. Condon and G. H. Shortley, *The Theory of Atomic Spectra*, Cambridge University Press, 1935.

Interaction of Radiation and Matter

O. Stenzel, *Light–Matter Interaction: A Crash Course for Students of Optics, Photonics and Materials Science*, Springer, 2022.

B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, 3rd ed. 2019, Chaps. 14–18.

P. van der Straten and H. J. Metcalf, *Atoms and Molecules Interacting with Light: Atomic Physics for the Laser Era*, Cambridge University Press, 2016.

G. Grynberg, A. Aspect, and C. Fabre, *Introduction to Quantum Optics: From the Semi-Classical Approach to Quantized Light*, Cambridge University Press, 2010.

J. H. van Vleck and D. L. Huber, Absorption, Emission, and Linebreadths: A Semihistorical Perspective, *Reviews of Modern Physics*, vol. 49, pp. 939–959, 1977.

Blackbody Radiation

D. S. Lemons, W. R. Shanahan, and L. Buchholtz, *On the Trail of Blackbody Radiation: Max Planck and the Physics of his Era*, MIT Press, 2022.

V. Sapritsky and A. Prokhorov, *Blackbody Radiometry. Volume 1: Fundamentals*, Springer, 2020.

B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, 3rd ed. 2019, Chap. 14.

S. M. Stewart and R. B. Johnson, *Blackbody Radiation: A History of Thermal Radiation Computational Aids and Numerical Methods*, CRC Press/Taylor & Francis, 2016.

M. Planck, *Planck's Columbia Lectures: Abridged and Unabridged Versions*, with commentary by W. Vlasak, Adaptive Enterprises, 2005.

H.-G. Schöpf, *Von Kirchhoff bis Planck: Theorie der Wärmestrahlung in historisch-kritischer Darstellung*, Akademie-Verlag (Berlin)/Springer-Fachmedien (Wiesbaden), 1978.

H. P. Baltes, On the Validity of Kirchhoff's Law of Heat Radiation for a Body in a Nonequilibrium Environment, in E. Wolf, ed., *Progress in Optics*, Elsevier/North-Holland, vol. 13, pp. 1–25, 1976.

A. A. Penzias and R. W. Wilson, A Measurement of Excess Antenna Temperature at 4080 Mc/s, *Astrophysical Journal*, vol. 142, pp. 419–421, 1965.

R. H. Dicke, P. J. E. Peebles, P. J. Roll, and D. T. Wilkinson, Cosmic Black-Body Radiation, *Astrophysical Journal*, vol. 142, pp. 414–419, 1965.

H. O. McMahon, Thermal Radiation from Partially Transparent Reflecting Bodies, *Journal of the Optical Society of America*, vol. 40, pp. 376–380, 1950.

A. Einstein, Zur Quantentheorie der Strahlung, *Physikalische Zeitschrift*, vol. 18, pp. 121–128, 1917 [Translation: On the Quantum Theory of Radiation, in D. ter Haar, *The Old Quantum Theory*, Pergamon, 1967].

M. Planck, *Vorlesungen über die Theorie der Wärmestrahlung*, Johann Ambrosius Barth Verlag (Leipzig), 2nd ed. 1913 [Translation: *The Theory of Heat Radiation*, P. Blakiston's Son & Co. (Philadelphia), 1914].

G. Kirchhoff, Über das Verhältniss zwischen dem Emissionsvermögen und dem Absorptionsvermögen der Körper für Wärme and Licht, *Poggendorff's Annalen*, vol. 109, pp. 275–301, 1860 [Translation: On the Relation between the Radiating and Absorbing Powers of Different Bodies for Light and Heat, *London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, Ser. 4, vol. 20(130), 1860].

Incandescent Light

M. F. Gendre, Incandescent Lamps, in R. Karlicek, C.-C. Sun, G. Zissis, and R. Ma, eds., *Handbook of Advanced Lighting Technology*, Springer Nature, pp. 1013–1064, 2017.

O. Ilic, P. Bermel, G. Chen, J. D. Joannopoulos, I. Celanovic, and M. Soljačić, Tailoring High-Temperature Radiation and the Resurrection of the Incandescent Source, *Nature Nanotechnology*, vol. 11, pp. 320–324, 2016.

M. Carlà, Stefan-–Boltzmann Law for the Tungsten Filament of a Light Bulb: Revisiting the Experiment, *American Journal of Physics*, vol. 81, pp. 512–517, 2013.

T. Matsumoto and M. Tomita, Modified Blackbody Radiation Spectrum of a Selective Emitter with Application to Incandescent Light Source Design, *Optics Express*, vol. 18, no. S2, pp. A192–A200, 2010.

V. Roberts, Incoherent Sources: Lamps, in R. D. Guenther, ed., *Encyclopedia of Modern Optics*, Elsevier, pp. 208–217, 2005.

M. R. Vukcevich, *The Science of Incandescence*, Nela Press, 1992.

W. Elenbaas, *Light Sources*, Crane, Russak, 1972.

E. Kauer, Generating Light with Selective Thermal Radiators, *Philips Technical Review*, vol. 26, no. 2, pp. 33–47, 1965.

J. C. De Vos, A New Determination of the Emissivity of Tungsten Ribbon, *Physica*, vol. 20, pp. 115–131, 1954.

W. W. Coblentz, Emissivity of Straight and Helical Filaments of Tungsten, *Bulletin of the Bureau of Standards*, vol. 14, pp. 690–714, 1918.

I. Langmuir, The Characteristics of Tungsten Filaments as Functions of Temperature, *Physical Review*, vol 7, pp. 302–330, 1916.

Infrared Detectors and Thermography

J. S. Campbell and M. N. Mead, *Human Medical Thermography*, CRC Press/Taylor & Francis, 2023.

M. Vollmer and K.-P. Möllmann, *Infrared Thermal Imaging: Fundamentals, Research and Applications*, Wiley–VCH, 2nd ed. 2018.

A. Daniels, *Field Guide to Infrared Systems, Detectors, and FPAs*, SPIE Optical Engineering Press, 3rd ed. 2018.

O. Breitenstein, W. Warta, and M. C. Schubert, *Lock-in Thermography: Basics and Use for Evaluating Electronic Devices and Materials*, Springer, 3rd ed. 2018.

M. A. Kinch, *State-of-the-Art Infrared Detector Technology*, SPIE Optical Engineering Press, 2014.

T. Kuroda, *Essential Principles of Image Sensors*, CRC Press/Taylor & Francis, 2014.

G. C. Holst and T. S. Lomheim, *CMOS/CCD Sensors and Camera Systems*, SPIE Optical Engineering Press, 2nd ed. 2011.

A. Rogalski, *Infrared Detectors*, CRC Press/Taylor & Francis, 2nd ed. 2011.

C. Jagadish, ed., *Semiconductors and Semimetals*, S. Gunapala and D. Rhiger, eds., Volume 84, *Advances in Infrared Photodetectors*, Academic/Elsevier, 2011.

H. Schneider and H. C. Liu, *Quantum Well Infrared Photodetectors: Physics and Applications*, Springer, 2007.

A. Rogalski, ed., *Selected Papers on Infrared Detectors: Developments*, SPIE Optical Engineering Press (Milestone Series Volume 179), 2004.

A. Rogalski, ed., *Selected Papers on Semiconductor Infrared Detectors*, SPIE Optical Engineering Press (Milestone Series Volume 66), 1992.

N. Sclar, Properties of Doped Silicon and Germanium Infrared Detectors, *Progress in Quantum Electronics*, vol. 9, pp. 149–257, 1984.

R. J. Keyes, ed., *Optical and Infrared Detectors*, Volume 19, *Topics in Applied Physics*, Springer, 2nd ed. 1980.

R. K. Willardson and A. C. Beer, eds., *Semiconductors and Semimetals*, *Infrared Detectors II*, Academic Press, vol. 12, 1977.

R. D. Hudson, Jr. and J. W. Hudson, eds., *Benchmark Papers in Optics / 2: Infrared Detectors*, Dowden, Hutchinson & Ross, 1975.

CHAPTER

5

SEMICONDUCTOR PHYSICS

5.1 ENERGY BANDS 127

5.2 CHARGE CARRIERS 131

5.3 SEMICONDUCTOR MATERIALS 134

5.4 CARRIER CONCENTRATIONS 141

5.5 GENERATION, RECOMBINATION, AND INJECTION 148

5.6 JUNCTIONS AND HETEROJUNCTIONS 152

5.7 QUANTUM WELLS AND MULTIQUANTUM WELLS 155

5.8 QUANTUM DOTS 161

5.9 ORGANIC AND PEROVSKITE SEMICONDUCTORS 164

William Shockley (1910–1989), seated, **John Bardeen (1908–1991)**, center, and **Walter Brattain (1902–1987)**, right, shared the Nobel Prize in 1956 for their research on semiconductors and the invention of the transistor. Semiconductor materials lie at the heart of LED technology.

LED Lighting: Devices and Colorimetry. Malvin Carl Teich.
Google Books. Published 2024.

Photonics, the technology of controlling the flow of photons, and electronics, the technology of controlling the flow of charge carriers, come together in a natural way in the domain of semiconductors. Photons generate mobile charge carriers and charge carriers generate photons. Following the invention of the point-contact transistor in 1947 (p. 125), which initiated the era of solid-state electronics, semiconductor devices began to enjoy widespread use. The invention of the p–n junction LED in 1962 (p. 198) opened the door to the era of photonics in a similar way, and semiconductor photonic devices became ubiquitous.

Semiconductors have a number of unique features and characteristics:

- Because the atoms that comprise the semiconductor crystal lattice are in close proximity with each other, these materials are best viewed not in terms of the individual energy levels of the constituent atoms, but rather in terms of energy levels that describe the system as a whole.
- Collections of closely spaced energy levels meld to form energy bands. At absolute zero, in the absence of external excitation, these bands are either fully occupied by electrons or totally unoccupied. The highest-lying fully occupied energy band is known as the valence band while the lowest-lying unoccupied energy band is known as the conduction band. The two bands are separated by what is called the forbidden band, which is characterized by the bandgap energy E_g and is devoid of both electrons and holes.
- An external energy source (whether thermal, optical, or electronic) can impart energy to an electron in the valence band, thereby causing it to jump across the forbidden band and enter the conduction band, where it is mobile. This transition leaves behind a vacancy (hole) in the valence band. The inverse process, called electron–hole recombination, entails an electron decaying from the conduction band to fill a vacancy in the valence band (provided that one is accessible), which generates a photon and/or phonons in the process. Photons therefore couple with electrons and holes.

This chapter provides an introduction to the physical principles that underlie the properties and operation of semiconductors and semiconductor devices. We begin by considering the formation of energy bands and bandgaps in bulk semiconductors (Sec. 5.1) and the motion of charge carriers (electrons and holes) in direct- and indirect-bandgap materials (Sec. 5.2). We then survey the vast landscape of elemental and compound semiconductors, including doped and 2D semiconductor materials (Sec. 5.3). Modern semiconductor photonic devices usually rely on III–V ternary or quaternary compounds such as InGaAsP, AlInGaP, InGaN, or AlInGaN, though they increasingly make use of organic and perovskite semiconductors, as well as compounds forged wholly from elements residing in group-IV of the periodic table, particularly C, Si, Ge, and Sn. (Electronic semiconductor devices are principally fabricated from Si.)

The rules that dictate how carriers fill the available energy levels in semiconductors, both near and far from thermal equilibrium, are examined in Sec. 5.4. Inasmuch as electrons and holes are indistinguishable quantum particles that obey Fermi–Dirac statistics, their energy levels and occupancy conditions differ markedly from those prescribed by the Boltzmann statistics, as discussed in Chapter 4. The generation, recombination, and injection of carriers is considered in Sec. 5.5, while junctions and heterojunctions are the topics of Sec. 5.6. A discussion of the energy levels, bands, and bandgaps of quantum-confined materials, such as quantum wells and multiquantum wells, as well as quantum wires, is provided in Secs. 5.7. The fabrication and energy levels of quantum dots are detailed in 5.8. Finally, Sec. 5.9 provides a brief introduction to organic and perovskite semiconductors.

5.1 ENERGY BANDS

The atoms (or molecules) of solids lie in close proximity to each other and typically coalesce into a periodic arrangement comprising a **crystal lattice**. The strength of the forces that hold the atoms together is roughly of the same magnitude as the forces that bind atoms into molecules. Consequently, the energy levels of solids are determined not only by the potentials associated with the individual atoms, but also by the potentials associated with neighboring lattice atoms. Noncrystalline solids such as glasses and plastics have orderly structures similar to those of crystals, but they extend only over a limited range.

Four principal types of bonding occur in ordinary solids: ionic, covalent, metallic, and molecular. **Ionic solids** (such as CaF_2) comprise a crystalline array of positive and negative ions held together by electrostatic attraction. Since there are no free electrons to carry current, these materials are insulators. They are generally transparent in the visible since their bandgaps usually lie in the ultraviolet. **Covalent solids**, like covalently bound molecules, consist of atoms bound by shared valence electrons. They are often insulators and can be transparent (such as diamond) or opaque (such as graphite) in the visible region. Covalent solids can also behave as semiconductors (e.g., GaAs) that are opaque in the visible and transparent in the infrared. **Metallic solids** have delocalized valence electrons that are collectively shared by all of the positive ions and move in their combined potential. The ability of the electrons to wander through metallic crystals is responsible for their high electrical conductivities. Metals strongly reflect light and are thus opaque in the visible. **Molecular solids** (also called **van der Waals solids**) contain small, non-polar covalent molecules held together by van der Waals forces, which are far weaker than those involved in other forms of binding.

Formation of Energy Bands

It is instructive to examine how the energy levels of an isolated atom are modified as it comes into close contact with neighboring atoms in the course of forming a crystal lattice. Isolated atoms and molecules (e.g., those in gases) exhibit discrete energy levels. Each individual atom in a collection of such identical isolated atoms has an identical set of such energy levels. As the atoms are brought into proximity to form a solid, exchange interactions (arising from the quantum requirement of indistinguishability for identical particles), along with the presence of fields of varying strengths from neighboring atoms, play an increasingly important role. The initially sharp energy levels associated with the valence electrons of the isolated atoms gradually broaden into collections of multiple densely spaced energy levels that form energy bands. This process is illustrated in Fig. 5.1-1, where electron energy levels are illustrated schematically for: (a) two isolated atoms; (b) a molecule containing two such atoms; and (c) a rudimentary one-dimensional (1D) lattice comprising five such atoms.

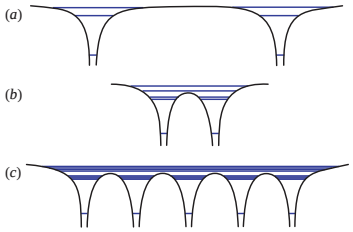


Figure 5.1-1 Schematic energy levels for: (a) two isolated atoms; (b) the same two atoms after having been brought into close contact and forming a diatomic molecule; and (c) five identical atoms in close proximity having formed a rudimentary one-dimensional (1D) crystal.

The lowest-lying energy levels remain sharp because the electrons in the inner subshells are shielded from the influence of nearby atoms, but the initially sharp energy levels

associated with the outer atomic electrons become bands as the atoms enter into close proximity and degeneracies are removed by Stark splitting.

This picture is elaborated in Fig. 5.1-2, where we schematically compare the energy levels of an isolated atom and three different kinds of solids that comprise lattices of such atoms: a metal, a semiconductor, and an insulator. The lowest-lying energy levels of these solids, denoted in this sketch by the electron configurations $1s$, $2s$, and $2p$, resemble those of the isolated atom because the inner electrons are shielded from interatomic forces. In contrast, the discrete higher energy levels of the atomic valence electrons, denoted $3s$ and $3p$, are split into densely packed energy bands in the solids. The lowest-lying unoccupied, or partially occupied, energy band is called the **conduction band** while the highest-lying fully occupied energy band is known as the **valence band**. These two bands are separated by the **forbidden band**, whose energy extent E_g is the **bandgap energy**. As with electrons in individual atoms, the Pauli exclusion principle applies to the electrons in solids so that the lowest-lying energy bands are occupied first. Typical values for the room-temperature conductivity σ for metals, semiconductors, and insulators are 10^8 $(\Omega\text{-m})^{-1}$, 10^{-4}–10^5 $(\Omega\text{-m})^{-1}$, and 10^{-10} $(\Omega\text{-m})^{-1}$, respectively.

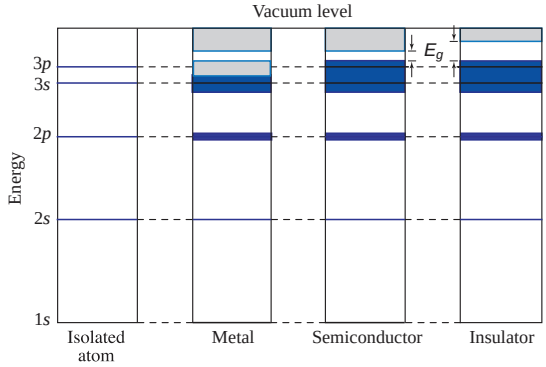


Figure 5.1-2 Broadening of the discrete energy levels of an isolated atom into energy bands when atoms in close proximity form a solid. Fully occupied bands are darkly shaded, unoccupied bands are lightly shaded, and partially occupied bands are both lightly and darkly shaded.

Energy Bands in Metals, Semiconductors, and Insulators

Metals comprise the greatest preponderance of elements in the periodic table. They have a partially occupied conduction band at all temperatures (lightly and darkly shaded region in Fig. 5.1-2). The availability of many unoccupied states in this band is responsible for the high electrical conductivity of metals. Semimetals, in contrast, have overlapping valence and conduction bands.

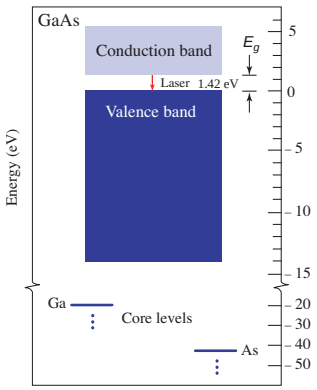
At $T = 0$ K, intrinsic semiconductors have an occupied valence band (dark shading in Fig. 5.1-2) and an unoccupied conduction band (light shading). Since there are no available free states in the valence band, and there are no electrons in the conduction band, the conductivity of an ideal intrinsic semiconductor at $T = 0$ K is zero. As the temperature of the semiconductor rises above absolute zero, however, an increasing number of electrons from the valence band gain sufficient thermal energy to access the conduction band and thereby contribute to the conductivity of the material.

Insulators also have a fully occupied valence band (dark shading in Fig. 5.1-2) and an unoccupied conduction band (light shading). They are typically distinguished from semiconductors by virtue of a bandgap energy $\gtrsim 3$ eV. As an example, the bandgap energy for silicon (a semiconductor) is $E_g \approx 1.1$ eV whereas that for diamond (an insulator) is $E_g \approx 5.5$ eV. Above absolute zero, fewer electrons in insulators acquire the requisite thermal energy to surmount the bandgap energy and contribute to the conductivity of the material. It should be pointed out, however, that the degree of band overlap is also instrumental in determining whether a material is classified as a metal, a semiconductor, or an insulator.

Energy Bands in Bulk Semiconductors

A semiconductor is a crystalline or amorphous solid whose electrical conductivity is typically intermediate between that of a metal and that of an insulator. Its conductivity can be significantly altered by modifying the temperature or doping concentration of the material, or by illuminating it with light. The band structure of semiconductors, and the ability to form junctions and heterostructures, offer unique opportunities.

The binary semiconductor GaAs was discovered early on to be useful in photonics. This material takes the form of a zincblende structure comprising two face-centered-cubic lattices, one of Ga atoms and the other of As atoms, displaced from each other by $1/4$ the length of a body diagonal (Fig. 5.1-3). The conventional unit cell is a cube. Each atom is surrounded by four atoms of the opposite type, which are equally spaced and located at the corners of a regular tetrahedron.



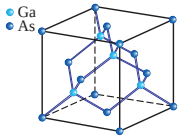


Figure 5.1-3 The semiconductor GaAs takes the form of a zincblende crystal structure comprising two face-centered-cubic lattices, one of Ga and the other of As. The higher energy levels are closely spaced and form bands. The zero of energy is (arbitrarily) defined at the top edge of the valence band. The GaAs light-emitting diode (and laser diode) operate on the electron transition between the bottom of the conduction band and the top of the valence band, in the near-infrared region of the spectrum.

Semiconductors have many closely spaced allowed electron energy levels that take the form of bands, as displayed in Fig. 5.1-3 for GaAs. The bandgap energy E_g, which is the energy separating the valence and conduction bands, is 1.42 eV at room temperature. The Ga and As ($3d$) core levels are quite sharp, as displayed in Fig. 5.1-3. The valence band of GaAs is formed from the $4s$ and $4p$ levels (in analogy with the schematic in Fig. 5.1-2).

Origin of the Energy Bandgap

The atoms comprising solid-state materials have sufficiently strong interatomic interactions that they cannot be treated as individual entities, as discussed at the beginning of this section. Their conduction electrons are not bound to individual atoms; rather, they belong to the collection of atoms as a whole. As illustrated in Fig. 5.1-2, each band contains a large number of densely packed discrete energy levels that is well approximated as a continuum. The solution of the Schrödinger equation for the electron energy, in the periodic potential created by the collection of atoms in the crystal lattice [Fig. 5.1-4(a)], results in a splitting of the atomic energy levels and the formation of energy bands.

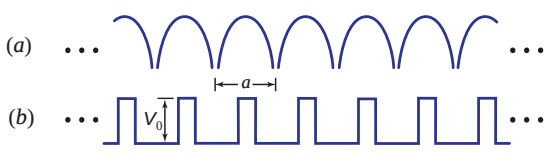


Figure 5.1-4 (a) Crystal-lattice potential for an ideal, 1D, infinite collection of atoms of lattice constant a. (b) Idealized rectangular-barrier potential (height V_0) that describes the Kronig–Penney model.

From a mathematical perspective, the origin of the bandgap may be illustrated in essence via the **Kronig–Penney model**. In this simple theory, the crystal-lattice potential, a one-dimensional version of which is displayed in Fig. 5.1-4(a), is approximated by a 1D periodic rectangular-barrier potential, as depicted in Fig. 5.1-4(b). The solution of the associated Schrödinger equation for this potential yields allowed energy bands with traveling-wave solutions that are separated by forbidden bands with exponentially decaying solutions. It can be shown that the results are general and carry over to three dimensions. The traveling-wave eigenfunctions, known as **Bloch modes**, assume the periodicity of the crystal lattice.

Bandgap Energy and Bandgap Wavelength

As discussed above, the valence and conduction bands of a semiconductor material are separated by a forbidden band with an energy extent known as the **bandgap energy** E_g. These bands and bandgap are illustrated in Fig. 5.1-5 for the iconic semiconductors Si and GaAs.

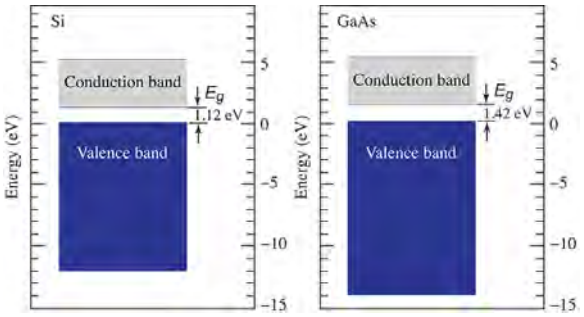


Figure 5.1-5 Energy bands in Si and GaAs. The bandgap energy E_g, which separates the valence and conduction bands, is 1.12 eV for Si and 1.42 eV for GaAs at $T = 300$ K.

The **bandgap wavelength** λ_g is related to the bandgap energy E_g via

$$\lambda_g = \frac{hc_0}{E_g}. \tag{5.1-1}$$

Bandgap Wavelength

When the bandgap wavelength is expressed in μm and the bandgap energy is expressed in eV, the following approximate formula may be used:

$$\lambda_g \approx \frac{1.24}{E_g}. \tag{5.1-2}$$

Bandgap Wavelength
λ_g (μm); E_g (eV)

The bandgap wavelength and bandgap energy are fully equivalent quantities but are inversely related: a decrease in λ_g corresponds to an increase in E_g, and *vice versa*. The bandgap energy/wavelength is a key parameter for characterizing the electrical and optical properties of semiconductor materials, as well as for the operation of LEDs fabricated from these materials, as will be explained in Sec. 6.4 and discussed in Sec. 7.3.

It will become apparent in the sequel that the bandgap wavelength of a semiconductor material λ_g is a key determinant of the wavelength λ_0 of the light emitted by an LED fabricated from that material.

5.2 CHARGE CARRIERS

Electrons and Holes

As is understood from the description provided in Sec. 5.1, the wavefunctions of the electrons in a semiconductor overlap. Since the **Pauli exclusion principle** applies, no two electrons may occupy the same quantum state and the lowest available energy levels fill first. Elemental semiconductors such as Si and Ge have four valence electrons per atom that form covalent bonds. At $T = 0$ K, the number of quantum states that can be accommodated in the valence band is such that it is completely filled while the conduction band is completely empty. The material cannot conduct electricity under these conditions.

As the temperature increases, however, some electrons can be thermally excited from the valence band into the empty conduction band, where unoccupied states are abundant (Fig. 5.2-1). These electrons then behave as mobile carriers that drift through the crystal lattice and generate an electric current in the presence of an applied electric field. Moreover, the electrons that depart the valence band leave behind unoccupied quantum states, which in turn allow the remaining electrons in the valence band to move by exchanging places with each other under the influence of an applied field. The motion of these electrons can just as well be regarded as the motion, in the opposite direction, of the holes left behind by the electrons that ascended to the valence band. A hole thus behaves as a particle with positive charge $+e$.

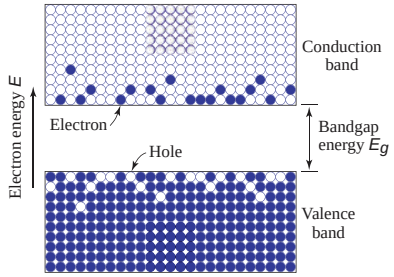


Figure 5.2-1 Electrons in the conduction band and holes in the valence band of a semiconductor material at $T > 0$ K.

The net result is that each electron excitation creates a free electron in the conduction band and a free hole in the valence band. The two charge carriers are free to drift under the effect of an applied electric field, thereby generating an electric current and a hole current. The material behaves as a semiconductor whose conductivity increases sharply with increasing temperature as a consequence of the increasing number of thermally generated mobile carriers.

Energy–Momentum Relations

In accordance with Schrödinger wave mechanics, the energy E and momentum $\mathbf{p}$ of an electron in a region of constant potential, such as free space, are related by $E = p^2/2m_0 = \hbar^2 k^2/2m_0$, where p is the magnitude of the momentum, k is the magnitude of the wavevector $\mathbf{k} = \mathbf{p}/\hbar$, and m_0 is the free-electron mass ($\approx 9.1 \times 10^{-31}$ kg). The E–k relation for a free electron is thus a simple parabola.

☐ **Energy–Momentum Relations for a Free Electron and a Free Photon.**

(a) The energy–momentum relation for a *free electron* of mass m_0 is established by solving the one-dimensional, time-independent, nonrelativistic Schrödinger equation,

$$-\frac{\hbar^2}{2m_0}\frac{d^2\psi(x)}{dx^2} + V(x)\psi(x) = E\psi(x)\,, \tag{5.2-1}$$

Time-Independent Schrödinger Equation

where $\psi(x)$ is the position wavefunction, $V(x) = 0$ is the potential energy for a free particle, and E is the electron energy. Substituting a plane-wave trial solution of the form $\psi(x) = A\exp(-jkx)$, where A is a constant, results in $(-\hbar^2/2m_0)\,(-jk)^2\,e^{-jkx} = Ee^{-jkx}$, which leads to the *quadratic* energy–momentum relation (the energy is not quantized):

$$E = \hbar^2 k^2/2m_0\,. \tag{5.2-2}$$

An alternative route to this result makes use of the *relativistic* energy–momentum relation for the total energy $\widehat{E}$ of a free particle of mass m_0:

$$\widehat{E}^2 = p^2c^2 + m_0^2c^4. \tag{5.2-3}$$

In the nonrelativistic limit, we carry out a Taylor-series expansion for the total energy, and retain the first term. Recalling that $\sqrt{1+x} \approx 1 + x/2$ for $x \ll 1$, we obtain

$$\begin{aligned}\widehat{E} &= \sqrt{p^2c^2 + m_0^2c^4} = \sqrt{m_0^2c^4\,(1 + p^2c^2/m_0^2c^4)} \\ &\approx m_0c^2\,\left(1 + p^2c^2/2m_0^2c^4\right) = m_0c^2 + p^2/2m_0\,.\end{aligned} \tag{5.2-4}$$

The term m_0c^2 represents the rest energy of the electron (≈ 0.511 MeV), so its nonrelativistic kinetic energy is $E = \widehat{E} - m_0c^2 = p^2/2m_0$. Using $p = \hbar k$ for the electron momentum, we arrive at $E = \hbar^2k^2/2m_0$, in accord with (5.2-2).

(b) The energy–momentum relation for a *free photon*, which travels at the speed of light c in a medium, is obtained by making use of (5.2-3) and recognizing that the rest mass of the photon is zero. Employing the relation $p = \hbar k$ from (3.3-11), we arrive at the *linear* energy–momentum relation

$$E = pc = c\hbar k. \tag{5.2-5}$$

■

The motion of an electron in a semiconductor material is similarly governed by the Schrödinger equation, but with a potential generated by the charges in the periodic crystal lattice of the material. As discussed earlier, this gives rise to allowed energy bands separated by forbidden bands, as exemplified by the Kronig–Penney model (Fig. 5.1-4). The ensuing E–k relations for electrons and holes, in the conduction and valence bands respectively, are illustrated in Fig. 5.2-2 for Si and GaAs. The energy E is a periodic function of the components (k_1, k_2, k_3) of the wavevector $\mathbf{k}$, with periodicities $(\pi/a_1, \pi/a_2, \pi/a_3)$, where a_1, a_2, a_3 are the crystal lattice constants. Figure 5.2-2 displays cross sections of this relation along two particular directions of the wavevector $\mathbf{k}$. The range of k values in the interval $[-\pi/a, \pi/a]$ defines the first **Brillouin zone**. The energy of an electron in the conduction band thus depends not only on the magnitude of its momentum, but also on the direction in which it is traveling in the crystal.

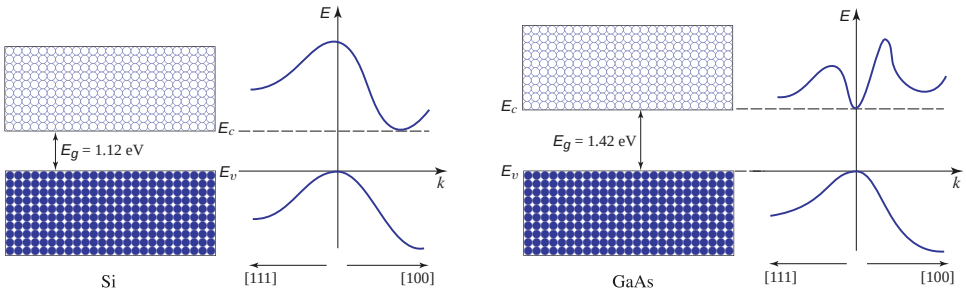


Figure 5.2-2 Cross sections of the E–k relations for Si and GaAs along two crystal directions: [111] toward the left and [100] toward the right.

Effective Mass

It is apparent from Fig. 5.2-2 that near the bottom of the conduction band, for both Si and GaAs, the E–k relation may be approximated by a parabola,

$$E = E_c + \frac{\hbar^2 k^2}{2m_c}, \tag{5.2-6}$$

where E_c is the energy at the bottom of the conduction band and k is measured from the value of the wavenumber where the minimum occurs. This parabolic behavior is highlighted in Fig. 5.2-3.

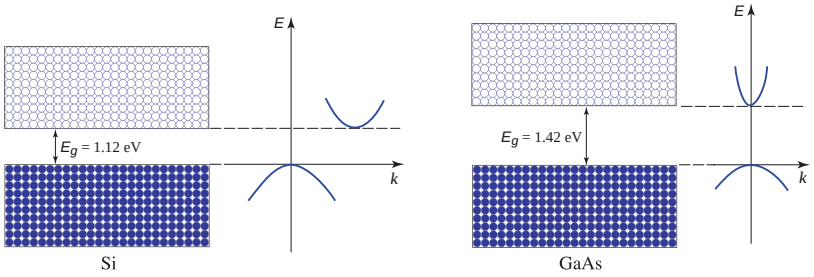


Figure 5.2-3 The E–k relation is well-approximated by parabolas at the bottom of the conduction band and at the top of the valence band, for both Si and GaAs.

The parabolic relation represented in (5.2-6) suggests that a conduction-band electron behaves in a manner analogous to that of a free electron, but with a mass m_c, called the conduction-band effective mass or the **electron effective mass**, that differs from that of the free-electron mass m_0. This is because the electron effective mass accommodates the influence of the lattice ions on the motion of the conduction-band electron.

Similarly, near the top of the valence band, we may write

$$E = E_v - \frac{\hbar^2 k^2}{2m_v}, \tag{5.2-7}$$

where $E_v = E_c - E_g$ is the energy at the top of the valence band and m_v is the valence-band effective mass or **hole effective mass**, as illustrated in Fig. 5.2-3. The influence of the lattice ions on the motion of a valence-band hole is captured by its effective mass.

The effective mass depends on the crystal structure of the material and on the direction of travel of the carrier with respect to the lattice since the interatomic spacing varies with crystallographic direction. It also depends on the particular band under consideration; indeed, several parabolas of different curvature sometimes coexist near the top of the valence band, corresponding to so-called **heavy holes**, **light holes**, and **split-off-band holes**. Selected averaged values of the effective masses for several semiconductor materials, normalized to the free-electron mass m_0, are presented in Table 5.2-1.

Table 5.2-1 Typical averaged values of the normalized electron and hole effective masses for selected semiconductor materials.

SEMICONDUCTOR	m_c/m_0	m_v/m_0
Si	0.98	0.49
Ge	0.34	0.29
GaAs	0.07	0.50
GaN	0.20	0.80

Direct- and Indirect-Bandgap Semiconductors

Semiconductors for which the conduction-band minimum energy and the valence-band maximum energy correspond to the same value of the wavenumber k (same momentum) are called **direct-bandgap** materials. Semiconductors for which this is not the case are known as **indirect-bandgap** materials. As is evident in Fig. 5.2-2, GaAs is a direct-bandgap semiconductor while Si is an indirect-bandgap semiconductor.

The distinction is important because a transition from the bottom of the conduction band to the top of the valence band in an indirect-bandgap semiconductor must accommodate a substantial change in the momentum of the electron to take place. Since a third body, such as a phonon, must participate in the interaction to absorb the excess momentum, the efficiency of photon emission is depressed. Under ordinary circumstances, therefore, as will be elucidated in Sec. 6.4:

Direct-bandgap semiconductors such as GaAs can emit light efficiently whereas indirect-bandgap semiconductors such as Si cannot.

5.3 SEMICONDUCTOR MATERIALS

Figure 5.3-1 displays the portion of the periodic table of the elements that relates to semiconductors. The elements in column IV, along with compound semiconductors formed from various combinations of elements in selected other columns, are the materials that underlie semiconductor physics and photonics. Each *column* of the table, which is designated by both a circled arabic numeral and a roman numeral, is referred to as a *group*. All elements in a group exhibit similar physical and chemical properties since they have the same number of valence electrons. The arabic numeral that labels each row at the left of the table represents the principal quantum number for the elements in that row.

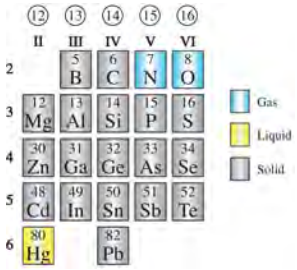


Figure 5.3-1 Element abbreviations and atomic numbers for the section of the periodic table that relates to semiconductors. Each column of the table, referred to as a group, is labeled by both a circled arabic numeral and a roman numeral. The latter, which is the older designation, remains prevalent in semiconductor technology. The arabic numerals at the left represent the principal quantum numbers for elements in that row. Elements depicted as blue, yellow, and silver take the form of gases, liquids, and solids at room temperature, respectively.

We proceed to consider the properties of several classes of elemental and compound semiconductors in the context of their constituent elements and the organization of the periodic table. The bandgap energies and wavelengths of these semiconductors are of particular importance since they determine the colors of the light that are generated by LEDs fabricated from these materials, as discussed in Sec. 7.3.

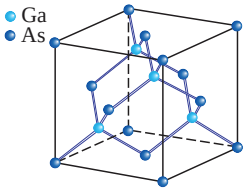
Elemental and Compound Semiconductors

Elemental Semiconductors:

The elemental semiconductors carbon (C), silicon (Si), and germanium (Ge) are located in group IV of the periodic table, and some of their basic properties are tabulated in Table 5.3-1. Although these elements can be coaxed into emitting light under special

conditions, they are typically not suitable for use as light emitters because of their indirect bandgaps, as discussed in the previous section. Group-IV elements can also be mixed to form compound semiconductors. An example of historical significance in the annals of LED technology is the indirect-bandgap, binary semiconductor silicon carbide (SiC). Also known as carborundum, this material serendipitously served as the first light-emitting Schottky-barrier diode in 1907 (p. 169 and Sec. 6.4). Silicon carbide plays a limited role in photonics today, as a substrate for III–nitride photon emitters. Using combinations of group-IV elements for photonic applications is an emerging area of technology known as group-IV photonics.

DIAMOND AND ZINCBLENDE:



WURTZITE:

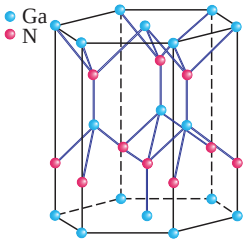


Table 5.3-1 Crystal structure, bandgap type, bandgap energy, and bandgap wavelength of selected elemental and binary III–V semiconductor materials.

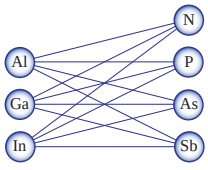
Semiconductor Material	Crystal Structure[a] (D/W/Z)	Bandgap Type (Indirect/Direct)	Bandgap Energy[b,c] E_g (eV)	Bandgap Wavelength[b,c] λ_g (μm)
Si	Diamond	Indirect	1.12	1.11
Ge	Diamond	Indirect	0.66	1.88
AlN	Wurtzite	Direct	6.02	0.206
AlP	Zincblende	Indirect	2.45	0.506
AlAs	Zincblende	Indirect	2.16	0.574
AlSb	Zincblende	Indirect	1.58	0.785
GaN	Wurtzite	Direct	3.39	0.366
GaP	Zincblende	Indirect	2.26	0.549
GaAs	Zincblende	Direct	1.42	0.873
GaSb	Zincblende	Direct	0.73	1.70
InN	Wurtzite	Direct	0.65	1.91
InP	Zincblende	Direct	1.35	0.919
InAs	Zincblende	Direct	0.36	3.44
InSb	Zincblende	Direct	0.17	7.29

[a] The crystal structure indicated represents the most commonly used form of the material. The zincblende structure comprises two interpenetrating face-centered-cubic (fcc) lattices, one for each element, displaced from each other by $1/4$ of the body diagonal. The diamond lattice is the same as zincblende except that all atoms are identical. The wurtzite structure consists of two hexagonal close-packed lattices, one for each element, displaced from each other along the three-fold c axis by $3/8$ of its length. All atoms are tetrahedrally bonded with their neighbors.

[b] At $T = 300$ K.

[c] The free-space bandgap wavelength λ_g and bandgap energy E_g are related by (5.1-1). When the bandgap energy is expressed in eV and the bandgap wavelength is expressed in μm, (5.1-2) proves convenient for calculation.

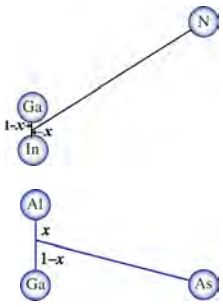
Binary III–V Semiconductors:



Compound semiconductors can be formed from an element in column III of the periodic table, such as aluminum (Al), gallium (Ga), or indium (In), with an element in column V, such as nitrogen (N), phosphorus (P), arsenic (As), or antimony (Sb) (the elements in these two columns are known as **icosagens** and **pnictogens**, respectively). Photonics was revolutionized in the 1950s by the growth of single-crystal III–V semiconductors such as these, which do not occur in nature and often have direct bandgaps. Selected properties of these twelve III–V compounds, including their crystal structure (zincblende or wurtzite), bandgap type

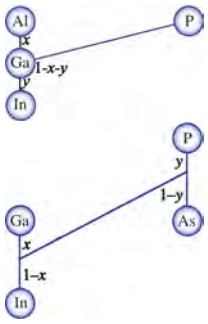
(direct or indirect), bandgap energy E_g, and bandgap wavelength λ_g are provided in Table 5.3-1 and Fig. 5.3-2. Most of them can be used to fabricate light-emitting diodes (and laser diodes). Indeed, the first practical III–V semiconductor LED was fabricated from GaAs in 1962 (see p. 198 and footnotes on p. 199). Thirty years later, in the early 1990s, the binary compound GaN moved to center stage when it was discovered that it could be used as a springboard for the development of InGaN, which enabled the fabrication of efficient blue LEDs and thence white light sources (see p. 306 and footnote on p. 307).

Ternary III–V Semiconductors:



Ternary III–V semiconductors are compounds formed from two elements in column III and one in column V (or one element in column III and two in column V). Although they are more complex to fabricate, ternary compounds offer more flexibility than binary ones because of their additional degree of freedom. An eminently important ternary semiconductor for photonics applications is $In_xGa_{1-x}N$, a direct-bandgap material whose bandgap wavelength can be compositionally tuned to span the green, blue, violet, and near-ultraviolet regions of the spectrum, as shown in Fig. 5.3-2(b). This material has properties that interpolate between those of GaN an InN, as determined by the compositional mixing ratio x (the fraction of Ga atoms in GaN that are replaced by In atoms). The bandgap energy E_g for this material varies between 3.39 eV for GaN and 0.65 eV for InN (Table 5.3-1), as x varies between 0 and 1 along the curve that connects these compounds in Fig. 5.3-2(b). Viable substrates for the III–nitrides include sapphire, SiC, and Si. Another widely used ternary semiconductor is $Al_xGa_{1-x}As$, whose properties interpolate between those of GaAs and AlAs, as specified by the compositional mixing ratio x (the fraction of Ga atoms in GaAs that are replaced by Al atoms). The bandgap energy E_g for this material varies between 1.42 eV (GaAs) and 2.16 eV (AlAs), as x varies between 0 and 1 along the line connecting these materials [Fig. 5.3-2(a)]. Bandgap energies for the ternary semiconductor $GaAs_{1-x}P_x$ (see p. 338 and footnote on p. 339) are also displayed in Fig. 5.3-2(a).

Quaternary III–V Semiconductors:



Quaternary semiconductors are formed by mixing three elements from column III with one from column V (or two from column III with two from column V). In general, these semiconductors offer greater design flexibility than ternary compounds by virtue of the additional degree of freedom. $Al_xIn_yGa_{1-x-y}P$ is a particularly important quaternary semiconductor compound that usually makes use of GaAs as a substrate (x and y represent the fraction of Ga atoms in the GaP that are replaced by Al and In atoms, respectively). This material is widely used for fabricating bright LEDs that operate in the red, orange, and yellow regions of the spectrum [shaded region in Fig. 5.3-2(a)]. Another quaternary material, the III–nitride compound $Al_xIn_yGa_{1-x-y}N$ [Fig. 5.3-2(b)], is used in the ultraviolet. Finally, we mention $In_{1-x}Ga_xAs_{1-y}P_y$, an important material in the near infrared that encompasses the 1550-nm optical fiber telecommunications band. The bandgap energy of this material extends from 0.36 eV (InAs) to 2.26 eV (GaP) for compositional mixing ratios x and y that vary between 0 and 1. The stippled area in Fig. 5.3-2(a) highlights the range of bandgap energies and lattice constants spanned by this compound.

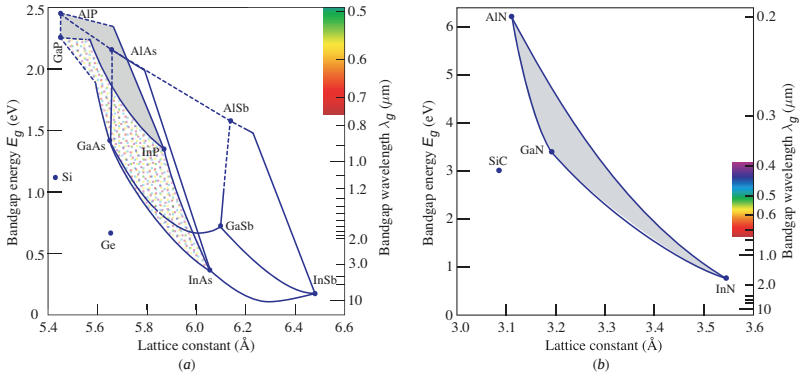


Figure 5.3-2 Dots represent bandgap energies, bandgap wavelengths, and lattice constants for Si, Ge, 6H-SiC, and 12 binary III–V compounds. Solid and dashed curves represent direct-bandgap and indirect-bandgap compositions, respectively. A material may have a direct bandgap for one mixing ratio and an indirect bandgap for a different mixing ratio. Ternary materials are represented along the curve that joins the constituent binary compounds. Quaternary materials are represented within the interior of the geometrical object defined by the binary components at its vertices. (*a*) The ternary compound $Al_xGa_{1-x}As$ is represented by points along the line connecting GaAs and AlAs. As x varies from 0 to 1, the point moves along this line from GaAs toward AlAs. Since the line is nearly vertical, $Al_xGa_{1-x}As$ is lattice matched to GaAs. The quaternary compound $Al_xIn_yGa_{1-x-y}P$ is represented by the shaded area with vertices at AlP, InP, and GaP, while $In_{1-x}Ga_xAs_{1-y}P_y$ is represented by the stippled area with vertices at InP, InAs, GaAs, and GaP. Both are important quaternary compounds, the former in the visible and the latter in the near infrared. (*b*) Although the ternary III–nitride compound $In_xGa_{1-x}N$ can, in principle, be compositionally tuned to accommodate the entire visible spectrum, the material becomes more difficult to grow as the In composition increases. $In_xGa_{1-x}N$ is therefore principally used in the green, blue, and violet spectral regions. $Al_xGa_{1-x}N$ and $Al_xIn_yGa_{1-x-y}N$ primarily serve the ultraviolet region. All compositions of these III–nitride compounds are direct-bandgap materials.

Binary and Ternary II–VI Semiconductors:

Binary II–VI materials, i.e., compound semiconductors formed from an element in column II, such as zinc (Zn), cadmium (Cd), or mercury (Hg), with an element in column VI, such as sulfur (S), selenium (Se), or tellurium (Te), are also important photonic materials (the elements in these two columns are known as **divalent metals** and **chalcogenides**, respectively). All of these materials have a zincblende structure and are direct-bandgap semiconductors, except for HgSe and HgTe, which are semimetals with small negative bandgap energies. Figure 5.3-3 displays the bandgap energies, bandgap wavelengths, and lattice constants of several II–VI compounds: ZnS, ZnSe, ZnTe, CdS, CdSe, CdTe, HgSe, and HgTe. The binary materials CdTe and HgTe are nearly lattice-matched, as is evidenced by the vertical line that connects them in Fig. 5.3-3, so that $Hg_xCd_{1-x}Te$ can be grown without strain on a CdTe substrate; indeed, this ternary semiconductor is widely used in mid-infrared photonics. Although the range of bandgap wavelengths available with II–VI materials encompasses the entire visible region, these *bulk* semiconductors are rarely used for fabricating photon sources because they suffer from limited lifespans as a result of material defects. On the other hand, *quantum dots* fabricated from binary and ternary II–VI materials, such as CdSe and ZnCdS, are essentially unaffected by material defects (Sec. 7.5). Indeed, such quantum dots serve as efficient generators of photoluminescence and electroluminescence that can be tuned over a wide range of visible wavelengths by simply modifying the dot size [Fig. 5.8-1(*a*)]. In fact, II–VI compounds are ubiquitous in nature and can be readily fashioned into colloidal quantum dots, whereas III–V compounds are not found in nature and it is especially challenging to forge Ga-based III–V materials into colloidal quantum dots.

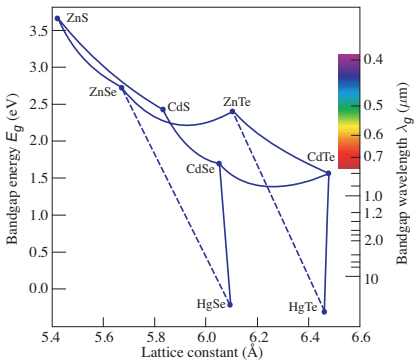


Figure 5.3-3 Bandgap energies, bandgap wavelengths, and lattice constants for several II–VI semiconductor compounds (HgSe and HgTe are semimetals with small negative bandgap energies). Although most of these bulk materials have direct bandgaps, they are rarely used for fabricating photon sources because of their limited lifespans. However, binary and ternary II–VI semiconductors such as CdSe and ZnCdS can be readily fashioned into colloidal quantum dots that efficiently emit luminescence with a wavelength that can be tuned by modifying the dot size [Fig. 5.8-1(a)].

Doped Semiconductors

Having described elemental and compound semiconductors, we now consider the properties of doped semiconductors. The electrical and optical properties of semiconductors can be substantially modified by introducing into the material small quantities of specially chosen impurities called **dopants**; the presence of such impurities can alter the concentration of mobile charge carriers by many orders of magnitude. Dopants with excess valence electrons, called **donors**, that replace a small proportion of the normal atoms in the crystal lattice (and are of similar size) create a predominance of mobile electrons and the material is then said to be an ***n*-type** semiconductor. Examples are atoms from column V of the periodic table, such as P or As, replacing a fraction of the column-IV atoms in an elemental semiconductor such as Si or Ge; and atoms from column VI, such as Se or Te, replacing a small fraction of the column-V atoms in a III–V binary semiconductor containing As or Sb.

Similarly, a ***p*-type** semiconductor is created by employing dopants with a deficiency of valence electrons, called **acceptors**, to replace a small proportion of the normal lattice atoms; the result is then a predominance of mobile holes. Examples are a small proportion of column-III atoms such as B or In replacing column-IV atoms in an elemental semiconductor; and a small proportion of column-II atoms such as Zn or Cd replacing column-III atoms in a III–V binary semiconductor containing Ga or In. Because column-IV atoms act as donors for column III, and as acceptors for column V, they can be used to create an excess of both electrons and holes in III–V materials. The introduction of dopants does not alter the charge neutrality of a material.

Doped materials are called **extrinsic semiconductors** while undoped materials (i.e., semiconductors that are devoid of intentional doping) are referred to as **intrinsic semiconductors**. In an intrinsic semiconductor, the concentrations of mobile electrons and holes are equal, i.e., $\mathtt{n} = \mathtt{p} = \mathtt{n}_\mathrm{i}$, and the intrinsic concentration $\mathtt{n}_\mathrm{i}$ grows exponentially with increasing temperature. In contrast, the concentration of mobile electrons in an n-type semiconductor (**majority carriers**) is far greater than the concentration of holes (**minority carriers**), so that $\mathtt{n} \gg \mathtt{p}$. The opposite is true in a p-type semiconductor, where holes are the majority carriers, and $\mathtt{p} \gg \mathtt{n}$. At room temperature, doped semiconductors typically have a majority-carrier concentration that is approximately equal to the doping concentration.

As semiconductor devices shrink in scale, there are ever smaller numbers of dopant atoms that are randomly distributed in position; hence, there may be only a handful of dopants, on average, at the nanoscale. However, techniques such as single-ion implantation can be used to fabricate semiconductors in which the number of dopant atoms, and their positions, are precisely determined. Indeed, semiconductor materials can be grown with sufficient purity so that nanodevices are devoid of impurities and solitary dopants can be inserted at specified positions to create designer dopant arrays.

EXAMPLE 5.3-1. ***Donor-Electron Ionization Energy.*** Consider a germanium crystal of relative permittivity $\epsilon/\epsilon_0 = n^2 = 16$ (Table 6.7-1) doped with As donor atoms. The electron effective mass of Ge is $m_c = 0.34\, m_0$ (Table 5.2-1), where m_0 is the free electron mass. The donor electron moves in the field of the singly charged arsenic ion (As^+), and has energy levels similar to those of an electron in the hydrogen atom. Hence, we choose the atomic number $Z = 1$ and the principal quantum number $\mathrm{n} = 1$ in the formula for the energy levels of a hydrogen-like atom, represented in (5.3-1), and replace the electric permittivity of free space ϵ_0 by ϵ and the reduced mass M_r by m_c to accommodate the polarization density and the crystal lattice of the semiconductor material, respectively, whereupon the energy of the donor electron becomes

$$E_D = -\left(\frac{1}{4\pi\epsilon_0}\right)^2 \left(\frac{Z^2}{\mathrm{n}^2}\right) \frac{M_r e^4}{2\hbar^2} = -\left(\frac{1}{4\pi\epsilon}\right)^2 \frac{m_c e^4}{2\hbar^2}. \tag{5.3-1}$$

Since the energy of the electron in the ground state of hydrogen ($Z = 1$) is $E_{\mathrm{n}=1} \approx -13.6$ eV with respect to the vacuum level (i.e., it is 13.6 eV below the ionization energy), the energy of the arsenic donor electron is $E_D = -(m_c/m_0)(\epsilon_0/\epsilon)^2 \times 13.6$ eV ≈ -0.018 eV. The donor electron thus resides in the Ge forbidden band, at a level ≈ 0.018 eV below the conduction band edge. Since the thermal energy $kT \approx 0.026$ eV at $T = 300$ K, however, essentially all of the donors are ionized at room temperature and the donor electrons are elevated to the conduction band. The material thus has a conduction-band donor-electron concentration that is roughly equal to the dopant concentration.

Graphene and 2D Materials

As indicated earlier in this section, the elements that reside in group-IV of the periodic table are of increasing interest in photonics. Carbon (C), silicon (Si), germanium (Ge), and tin (Sn) are of particular importance. Group-IV elements exist in various structural forms, known as **allotropes**, which exhibit distinct properties and have different applications. Although others exist, the most widely known allotropes of these four elements are:

- C: Diamond, graphite, carbon nanotubes, carbon dots, and graphene.
- Si: Crystalline and amorphous silicon, as well as silicene (analog of graphene).
- Ge: α-Ge, β-Ge, and germanene (analog of graphene).
- Sn: α-Sn (gray tin), β-Sn (white tin), and stanene (analog of graphene).

Graphene, a material comprising a one-atom-thick carbon honeycomb lattice, has come to the fore in recent years because of its unique properties and because it can be fashioned into a variety of photonic devices. Graphene and its analogs indicated above all take the form of hexagonal-lattice 2D atomic sheets that are usually denoted h-C, h-Si, h-Ge, and h-Sn, respectively, where the designation 'h' represents 'hexagonal.'

The relatively new fields of **graphene photonics** and **2D-material photonics** fall within the rubric of **group-IV photonics**.

Graphene. Graphene is a 2D material comprising a single, 0.33-nm-thick layer of graphite with atoms arranged in a hexagonal honeycomb structure (Fig. 5.3-4). Graphene was first extracted from graphite in 2004 by Andre Geim and Konstantin Novoselov, an achievement for which they were awarded the 2010 Nobel Prize in physics. Graphene is endowed with a collection of exceptional properties that make it useful in many photonics applications:

- It is an excellent conductor of electricity and has an optical transmittance near unity so it can be used as a transparent electrode. Its optical absorbance is nearly constant at $\xi = \pi e^2/\hbar c \rightsquigarrow 2.3\%$ over a broad wavelength band that stretches from 0.7 to 25 μm; its intensity reflectance is a negligible $\mathcal{R} \approx 1.3 \times 10^{-4}$; and its intensity transmittance at normal incidence is $\mathcal{T} \approx 97.7\%$. Its current-carrying capacity is also substantial ($i \approx 10^8$ A/cm^2 on SiO_2).

- It is a semimetal with zero bandgap that can interact with radiation over a broad spectral range stretching from the THz to the ultraviolet. Its absorption coefficient $\alpha \approx 7 \times 10^5$ cm^{-1} is an order of magnitude greater than that of Si or GaAs. It is readily doped, so that its electronic properties can be altered.
- It has an unusually high electron mobility. When deposited on SiO_2, its mobility is $\mu \approx 1.5 \times 10^4$ cm^2/V·s so that the drift velocity of carriers is an order of magnitude greater than that in Si. It therefore has an inordinately fast response and is suitable for use in ultrafast photodetectors. Its high area-to-volume ratio makes it highly effective for applications involving sensing.
- It is chemically stable, refractory to high temperatures, and resilient in high humidity. It has high thermal conductivity and excellent mechanical strength, yet is elastic and therefore bendable.
- It exhibits fast and strong absorption saturation, rendering it suitable for use as a saturable absorber for mode-locked lasers and as a broadband modulator.

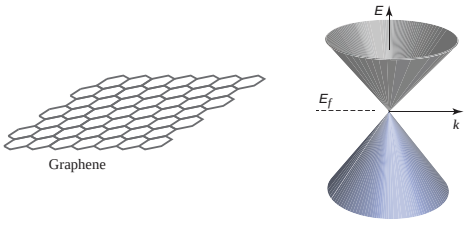


Figure 5.3-4 Graphene, also referred to as h-C, is a single layer of carbon atoms arranged in a hexagonal honeycomb lattice. Its E–k diagram is conical rather than parabolic (compare with Fig. 5.2-3). Graphene behaves as a semimetal with zero bandgap since its conduction- and valence-band cones meet at points that define the Fermi level E_f.

Because of its particular 2D symmetry, the band structure for carriers in graphene takes the form of cones (Fig. 5.3-4), rather than the parabolas that are characteristic of traditional semiconductors (Fig. 5.2-3). The E–k diagram is therefore linear rather than parabolic; it is similar to that for photons and is characterized by (5.2-5) rather than by (5.2-2). As with photons, the electronic excitations (called **Dirac fermions**) behave as if they were massless; this leads to an unusually large Fermi velocity, $v \approx c/300$, that underlies graphene's fast response. Furthermore, the conduction- and valence-band cones meet at single points (called Dirac points) that define the Fermi level, so that graphene behaves as a semimetal with zero bandgap. A number of other 2D materials also host massless Dirac fermions and behave as semimetals (e.g., silicene, germanene, stanene, and β_{12}-borophene), but most have approximately parabolic, rather than conical, band structures. Although Dirac fermions have been most widely studied in 2D materials, they are also hosted by some 3D materials, such as compressively strained α-Sn (gray tin) and Na_3Bi.

While the interaction of light with graphene is strong on a per-unit-distance basis ($\alpha \approx 7 \times 10^5$ cm^{-1}), devices that rely on single-pass operation encounter an insignificant thickness of material (0.33 nm). Building an effective 2D-based device therefore generally requires that the interaction be enhanced, which may be achieved by specialized doping or siting, by coupling to a photonic waveguide or cavity, or by coupling to plasmons, phonons, or excitons. Significant enhancement of the light–matter interaction can be attained by making use of traveling surface plasmon polaritons.

Other 2D Materials. By virtue of its semimetallic nature, graphene is a poor emitter of light. However, a number of other 2D materials, including various **transition-metal dichalcogenides (TMDs)** such as molybdenum disulfide, behave as direct-bandgap semiconductors with bandgap energies E_g that lie between 0.5 and 3 eV. As with 3D semiconductors, the bandgap energy can be tuned via chemistry, composition, and/or quantum confinement. These materials can serve as light emitters or reflectors dominated by excitonic transitions.

Single-layer TMDs such as MoS_2 and WSe_2 consist of a sublayer of transition metal sandwiched between two sublayers of chalcogen. MoS_2, for example, has an overall layer thickness of 0.65 nm and a bandgap energy of 1.8 eV. In their 3D configurations, some of these materials (e.g., graphite, MoS_2) serve as industrial lubricants. This is because consecutive atomic layers are bound only by weak van der Waals forces and easily slide over each other, a property that made it relatively easy for Geim and Novoselov to peel off a single graphene layer from graphite. Indeed, such 2D materials are often called **van der Waals materials**. Other 3D precursors (e.g., silicon, germanium) form tight bonds in all three dimensions so that their 2D versions, when extracted, tend to buckle.

The number of possible TMDs that can be formed is substantial since there are tens of transition metals and at least three chalcogens (S, Se, and Te; the elements O, Po, and Lv are sometimes also included in this category). Some single-layer materials behave as insulators (e.g., hexagonal BN, with $E_g \approx 6$ eV) and others behave as metals. 2D materials can be used in isolation, or combined in layers of various compositions, to create atomically thin heterostructures that serve as planar photonic devices. Hybrid structures comprising layers of graphene and TMDs have been shown to exhibit unique properties that offer promise for efficient light sources, modulators, and photodetectors.

5.4 CARRIER CONCENTRATIONS

Determining the concentration of carriers (electrons and holes) in a semiconductor as a function of energy requires knowledge of two features that we calculate in turn:

- The density of allowed energy levels (density of states).
- The probability that each of these levels is occupied (occupancy probability).

We conclude this section by analyzing semiconductor carrier concentrations in thermal equilibrium and quasi-equilibrium.

Density of States

The quantum state of an electron in a conventional bulk semiconductor material is characterized by its energy E, its wavevector $\mathbf{k}$ [the magnitude of which is approximately related to E by (5.2-6) or (5.2-7)], and its spin. The state is described by a wavefunction that satisfies certain boundary conditions.

An electron near the conduction-band edge may be approximately described as a particle of mass m_c confined to a three-dimensional cubic box of dimension d with perfectly reflecting walls, i.e., a three-dimensional infinite rectangular potential well. The standing-wave solutions require that the components of the vector $\mathbf{k} = (k_x, k_y, k_z)$ assume the discrete values $\mathbf{k} = (q_1\pi/d,\ q_2\pi/d,\ q_3\pi/d)$, where the respective mode numbers (q_1, q_2, q_3) are positive integers. This result is a three-dimensional generalization of the one-dimensional infinite square well (Example 5.7-1). The tip of the vector $\mathbf{k}$ must lie on the points of a lattice whose cubic unit cell has dimension π/d. There are therefore $(d/\pi)^3$ points per unit volume in $\mathbf{k}$-space. The number of states whose vectors $\mathbf{k}$ have magnitudes between 0 and k is determined by counting the number of points lying within the positive octant of a sphere of radius k [with volume $\approx (\frac{1}{8})4\pi k^3/3 = \pi k^3/6$]. Because of the two possible values of the electron spin, each point in $\mathbf{k}$-space corresponds to two states. There are therefore approximately $2(\pi k^3/6)/(\pi/d)^3 = (k^3/3\pi^2)d^3$ such points in the volume d^3 and hence $(k^3/3\pi^2)$ points per unit volume. It follows that the number of states with electron wavenumbers between k and $k + \Delta k$, per unit volume, is $\varrho(k)\Delta k = [(d/dk)(k^3/3\pi^2)]\Delta k = (k^2/\pi^2)\Delta k$, which leads to a density of states given by

$$\varrho(k) = \frac{k^2}{\pi^2} \,. \tag{5.4-1}$$

Density of States

This derivation is identical to that used for counting the number of modes that can be supported in a three-dimensional electromagnetic cavity (Sec. 4.4). In the case of electromagnetic modes there are two degrees of freedom associated with the field polarization (i.e., two photon spin values), whereas in the semiconductor case there are two spin values associated with the electron state. In resonator optics the allowed electromagnetic solutions for $\mathbf{k}$ were converted into allowed frequencies via the linear frequency–wavenumber relation $\nu = ck/2\pi$. In semiconductor physics, in contrast, the allowed solutions for $\mathbf{k}$ are converted into allowed energies via the quadratic energy–wavenumber relations specified in (5.2-6) and (5.2-7) and displayed in Fig. 5.4-1(a).

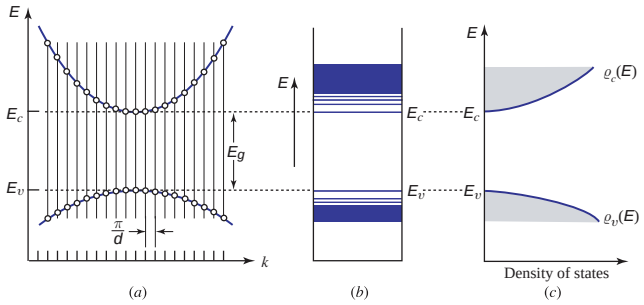


Figure 5.4-1 (a) Cross section of the E–k diagram (e.g., in the direction of the k_1 component, with k_2 and k_3 fixed). (b) Allowed energy levels (at all $\mathbf{k}$). (c) Density of states near the edges of the conduction and valence bands. The quantity $\varrho_c(E)\,dE$ is the number of quantum states with energy between E and $E+dE$, per unit volume, in the conduction band. The quantity $\varrho_v(E)$ has an analogous interpretation for the valence band.

If $\varrho_c(E)\,\Delta E$ represents the number of conduction-band energy levels (per unit volume) lying between E and $E + \Delta E$, then, because of the one-to-one correspondence between E and k specified in (5.2-6), the densities $\varrho_c(E)$ and $\varrho(k)$ must be related by $\varrho_c(E)\,dE = \varrho(k)\,dk$. Thus, the density of allowed energies in the conduction band is $\varrho_c(E) = \varrho(k)/(dE/dk)$. Similarly, the density of allowed energies in the valence band is $\varrho_v(E) = \varrho(k)/(dE/dk)$, where E is given by (5.2-7). The approximate quadratic E–k relations (5.2-6) and (5.2-7), which are valid near the edges of the conduction band and valence band, respectively, are used to evaluate the derivative dE/dk for each band, and the result is

$$\varrho_c(E) = \frac{(2m_c)^{3/2}}{2\pi^2\hbar^3}\sqrt{E - E_c}\,, \quad E \geqslant E_c \tag{5.4-2}$$

$$\varrho_v(E) = \frac{(2m_v)^{3/2}}{2\pi^2\hbar^3}\sqrt{E_v - E}\,, \quad E \leqslant E_v. \tag{5.4-3}$$

Density of States Near Band Edges

The square-root relation in (5.4-3) is a result of the quadratic energy–wavenumber formulas for electrons and holes near the band edges. The dependence of the density of states on energy is illustrated in Fig. 5.4-1(c). It is zero at the band edge, and increases away from it at a rate that depends on the effective masses of the electrons and holes. The values of m_c and m_v provided in Table 5.2-1 are averaged values suitable for calculating the density of states.

Occupancy Probabilities

As discussed in Sec. 4.2, the laws of statistical mechanics dictate that under conditions of thermal equilibrium at temperature T, the probability that a given state of energy E is occupied by an electron is determined by the **Fermi function**

$$f(E) = \frac{1}{\exp\left[(E - E_f)/kT\right] + 1}, \tag{5.4-4}$$

Fermi Function

where k is Boltzmann's constant and E_f is the **Fermi level**. Also called the **Fermi–Dirac distribution**, this result was initially set forth in (4.2-4) and plotted in Fig. 4.2-2.

The function $f(E)$ is not itself a probability distribution, and it does not integrate to unity; rather, it is a sequence of occupation probabilities for successive energy levels. Each energy level E is either occupied [with probability $f(E)$], or unoccupied [with probability $1 - f(E)$]. The Fermi function is applicable for indistinguishable particles when the electron density is nonnegligible. Unlike the valence electrons in an atom, the electron density in a semiconductor is large so the Boltzmann approximation to the Fermi function displayed in Fig. 4.2-2 is not applicable.

The Fermi function is displayed in Fig. 5.4-2 in relation to the conduction and valence bands of an intrinsic semiconductor at $T > 0$ K and $T = 0$ K. Because (5.4-4) dictates that $f(E_f) = 1/2$, whatever the temperature T, the Fermi level is that particular energy at which the probability of occupancy is $1/2$ (when there is an allowed energy state at E_f). By virtue of the symmetry of the Fermi function, and the fact that the number of electrons and number of holes are equal in an intrinsic semiconductor, the Fermi level falls in the middle of the forbidden band, as illustrated in Fig. 5.4-2.

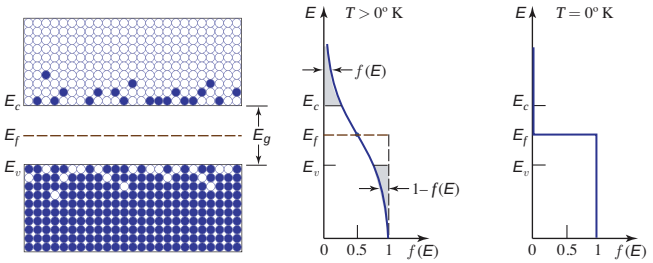


Figure 5.4-2 The Fermi function $f(E)$ is the probability that energy level E in the conduction band is filled with an electron, while $1 - f(E)$ is the probability that energy level E in the valence band is occupied by a hole. Results are sketched for $T > 0$ K and $T = 0$ K.

In the absence of thermal excitation, at $T = 0$ K, all electrons occupy the lowest possible energy levels, subject to the Pauli exclusion principle, so that $f(E) = 1$ for $E \leqslant E_f$, and $f(E) = 0$ for $E > E_f$. There are then no holes in the valence band (it is completely filled with electrons) and no electrons in the conduction band (it is completely empty). At $T = 0$ K, the Fermi level E_f marks the division between the occupied and unoccupied energy levels.

When the temperature is increased, so that $T > 0$ K, thermal excitations raise some electrons from the valence band to the conduction band, leaving behind empty states in the valence band (holes). The Fermi function $f(E)$ then represents the probability that energy level E in the conduction band is filled with an electron and $1 - f(E)$ is the probability that it is empty. Analogously, $1 - f(E)$ is the probability that energy level E in the valence band is occupied by a hole and $f(E)$ is the probability that it is not:

$$f(E) = \text{probability of occupancy by an electron in the conduction band.} \tag{5.4-5}$$

$$1 - f(E) = \text{probability of occupancy by a hole in the valence band.} \tag{5.4-6}$$

When $E - E_f \gg kT$, $f(E) \approx \exp[-(E-E_f)/kT]$ so that the high-energy tail of the Fermi function in the conduction band decreases exponentially with increasing energy. The Fermi function is then proportional to the Boltzmann distribution, which describes the exponential energy dependence of the fraction of a population of atoms excited to a given energy level (Sec. 4.2). By symmetry, when $E < E_f$ and $E_f - E \gg kT$, $1 - f(E) \approx \exp[-(E_f - E)/kT]$; the probability of occupancy by holes in the valence band then decreases exponentially as the energy decreases well below the Fermi level.

Thermal-Equilibrium Carrier Concentrations

Let $\mathtt{n}(E)\,\Delta E$ and $\mathtt{p}(E)\,\Delta E$ be the number of electrons and holes per unit volume, respectively, with energy lying between E and $E + \Delta E$. The densities $\mathtt{n}(E)$ and $\mathtt{p}(E)$ are obtained by multiplying the densities of states at energy level E provided in (5.4-2) and (5.4-3), respectively, by the occupancy probabilities of electrons and holes at that level as specified in (5.4-5) and (5.4-6), so that

$$\mathtt{n}(E) = \varrho_c(E) f(E), \qquad \mathtt{p}(E) = \varrho_v(E)[1 - f(E)]. \tag{5.4-7}$$

The concentrations (populations per unit volume) of electrons and holes, $\mathtt{n}$ and $\mathtt{p}$, are then obtained via the integrals

$$\mathtt{n} = \int_{E_c}^{\infty} \mathtt{n}(E)\, dE, \qquad \mathtt{p} = \int_{-\infty}^{E_v} \mathtt{p}(E)\, dE. \tag{5.4-8}$$

In an intrinsic (pure) semiconductor at any temperature, $\mathtt{n} = \mathtt{p}$ because thermal excitations always create electrons and holes in pairs. The Fermi level must therefore be placed at an energy value such that $\mathtt{n} = \mathtt{p}$. In materials for which $m_v = m_c$, the functions $\mathtt{n}(E)$ and $\mathtt{p}(E)$ are also symmetric, in which case E_f must lie precisely in the middle of the forbidden band (Fig. 5.4-3). In most intrinsic semiconductors, the Fermi level does indeed lie near the middle of the bandgap.

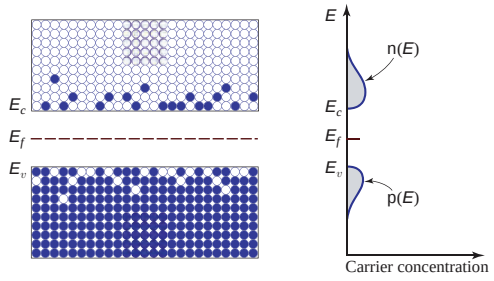


Figure 5.4-3 The electron and hole concentrations, $\mathtt{n}(E)$ and $\mathtt{p}(E)$, respectively, as a function of energy E, for an intrinsic semiconductor in thermal equilibrium. The total electron and hole concentrations are denoted $\mathtt{n}$ and $\mathtt{p}$, respectively.

The energy-band diagrams, Fermi functions, and concentrations of electrons and holes for n-type and p-type doped semiconductors in thermal equilibrium are displayed in Figs. 5.4-4 and 5.4-5, respectively. Donor electrons occupy an energy E_D that is slightly below the conduction-band edge, so that they are easily raised to the conduction band. If $E_D = 0.01$ eV, for example, most donor electrons at room temperature ($kT = 0.026$ eV) will be thermally excited into the conduction band (Example 5.3-1). As a result, the Fermi level [the energy at which $f(E_f) = 1/2$] lies above the middle of the bandgap.

For a p-type semiconductor, the acceptor energy level lies at an energy E_A just above the valence-band edge so that the Fermi level lies below the middle of the bandgap. Our attention has been directed to the mobile carriers in doped semiconductors; however, it must be kept in mind that these materials are electrically neutral, as is assured by the fixed donor and acceptor ions. Hence, $\mathtt{n} + N_A = \mathtt{p} + N_D$, where N_A and N_D are the number of ionized acceptors and donors per unit volume, respectively.

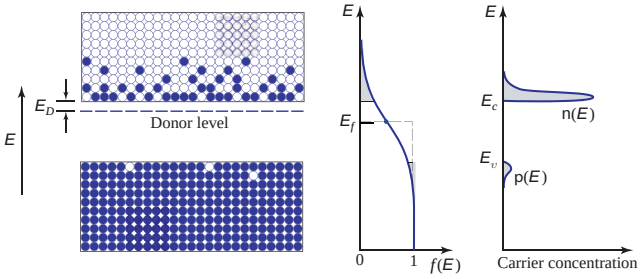


Figure 5.4-4 Energy-band diagram, Fermi function $f(E)$, and concentrations of mobile electrons and holes, $\mathtt{n}(E)$ and $\mathtt{p}(E)$, respectively, for an n-type semiconductor in thermal equilibrium.

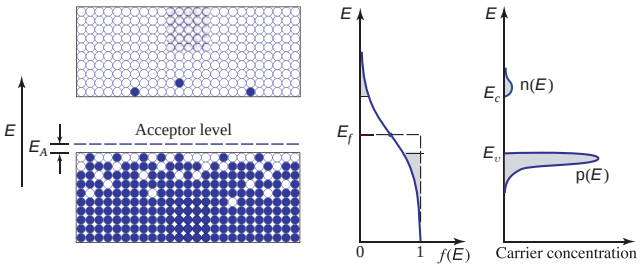


Figure 5.4-5 Energy-band diagram, Fermi function $f(E)$, and concentrations of mobile electrons and holes, $\mathtt{n}(E)$ and $\mathtt{p}(E)$, respectively, for a p-type semiconductor in thermal equilibrium.

☐ **Exponential Approximation of the Fermi Function.** The Fermi function $f(E)$ given in (5.4-4) may be approximated by the exponential function $f(E) \approx \exp[-(E - E_f)/kT]$ when $E - E_f \gg kT$. Similarly, $1 - f(E)$ may be approximated by an exponential function when $E_f - E \gg kT$. These approximations are applicable when the Fermi level lies within the bandgap, but away from the band edges by an energy of at least several times kT (at room temperature $kT \approx 0.026$ eV while $E_g = 1.12$ eV in Si and 1.42 eV in GaAs).

Substituting these exponential approximations into (5.4-8), and making use of (5.4-2) and (5.4-7), leads to $\mathtt{n} = \int_{E_c}^{\infty} A(E - E_c)^{1/2} \exp[-(E - E_f)/kT]\, dE$, where $A = (2m_c)^{3/2}/2\pi^2\hbar^3$ is a constant. To carry out the integration, we use the transformation $u = (E - E_c)/kT$, with $du = dE/kT$, so that $\exp[-(E - E_f)/kT] = \exp(-u)\exp[-(E_c - E_f)/kT]$, whereupon the integral becomes

$$\mathtt{n} = A(kT)^{3/2} \exp\left(-\frac{E_c - E_f}{kT}\right) \int_0^{\infty} u^{1/2} \exp(-u)\, du$$

$$= \frac{4\pi(2m_c kT)^{3/2}}{h^3} \sqrt{\frac{\pi}{4}} \exp\left(-\frac{E_c - E_f}{kT}\right), \tag{5.4-9}$$

from which we obtain (5.4-10). A similar analysis leads to (5.4-11), while (5.4-12) follows by multiplication of the two results:

$$\mathtt{n} = N_c \exp\left(-\frac{E_c - E_f}{kT}\right) \tag{5.4-10}$$

$$\mathtt{p} = N_v \exp\left(-\frac{E_f - E_v}{kT}\right) \tag{5.4-11}$$

$$\mathtt{np} = N_c N_v \exp\left(-\frac{E_g}{kT}\right). \tag{5.4-12}$$

Here $N_c = 2(2\pi m_c kT/h^2)^{3/2}$ and $N_v = 2(2\pi m_v kT/h^2)^{3/2}$ represent constants associated with the conduction and valence bands, respectively. These approximations are applicable for both intrinsic and doped semiconductors.

If $m_v = m_c$, then $N_c = N_v$ whereupon (5.4-10) and (5.4-11) give rise to the ratio $\mathtt{n}/\mathtt{p} = \exp[+(E_f - E_v)/kT - (E_c - E_f)/kT]$. Hence, if $E_c - E_f < E_f - E_v$, the argument of the exponent is positive and then so too is $\mathtt{n}/\mathtt{p}$. This indicates that if E_f is closer to the conduction band than to the valence band, then $\mathtt{n} > \mathtt{p}$, and *vice versa*. ■

Law of Mass Action

Equation (5.4-12), which relies on the validity of exponential approximations to the Fermi functions, reveals that, in thermal equilibrium, the product

$$\mathtt{np} = 4\left(\frac{2\pi kT}{h^2}\right)^3 (m_c m_v)^{3/2} \exp\left(-\frac{E_g}{kT}\right) \tag{5.4-13}$$

is independent of both the locations of the Fermi levels E_f within the forbidden band and the semiconductor doping levels. The constancy of the this product of concentrations is known as the **law of mass action**.

For an intrinsic semiconductor, $\mathtt{n} = \mathtt{p} \equiv \mathtt{n}_\mathrm{i}$, which, when combined with (5.4-12), yields

$$\boxed{\mathtt{n}_\mathrm{i} \approx \sqrt{N_c N_v}\, \exp\left(-\frac{E_g}{2kT}\right),} \tag{5.4-14}$$

Intrinsic Carrier Concentration

revealing that the intrinsic concentration of electrons and holes increases with increasing temperature T at an exponential rate. The law of mass action may then be written in the form

$$\boxed{\mathtt{np} = \mathtt{n}_\mathrm{i}^2.} \tag{5.4-15}$$

Law of Mass Action

The values of $\mathtt{n}_\mathrm{i}$ for various materials differ because of differences in the bandgap energies and effective masses. The room-temperature intrinsic carrier concentrations for several semiconductors are provided in Table 5.4-1.

Table 5.4-1 Intrinsic semiconductor carrier concentrations at $T = 300$ K.[a]

SEMICONDUCTOR	$\mathtt{n}_\mathrm{i}$ (cm^{-3})
Si	1.5×10^{10}
Ge	2.5×10^{13}
GaAs	1.8×10^{6}
GaN	1.9×10^{-10}

[a] Substitution into (5.4-14) of the values of m_c and m_v provided in Table 5.2-1, along with the values of E_g given in Table 5.3-1, fails to yield the numerical values of $\mathtt{n}_\mathrm{i}$ listed in the table because of the sensitivity of (5.4-14) to the precise values of the parameters.

The law of mass action may also be used to determine the concentrations of electrons and holes in doped semiconductors. A moderately doped n-type material, for example, has a concentration of electrons $\mathtt{n}$ that is essentially equal to the donor concentration N_D, so that the hole concentration is $\mathtt{p} = \mathtt{n}_\mathrm{i}^2/N_D$. Knowledge of $\mathtt{n}$ and $\mathtt{p}$ then allows the Fermi level to be determined via (5.4-8). When the Fermi level lies within the forbidden band, at an energy greater than several times kT from its edges, the approximate relations in (5.4-10) and (5.4-11) can be used to determine it directly.

On the other hand, if the Fermi level lies inside the conduction (or valence) band, the exponential approximations of the Fermi function are invalid so that $\mathtt{np} \neq \mathtt{n}_\mathrm{i}^2$ and the carrier concentrations must be obtained numerically. The material is then referred to as a **degenerate semiconductor**. For very heavy doping, the donor (acceptor) impurity band merges with the conduction (valence) band to become what is known as the **band tail**, which effectively results in a decrease of the bandgap.

Thermal Quasi-Equilibrium Carrier Concentrations

The occupancy probabilities and carrier concentrations considered previously are applicable only for semiconductors in thermal equilibrium. Another condition, known as **thermal quasi-equilibrium**, prevails when the relaxation (decay) times for transitions within the conduction and valence bands are severally much shorter than the relaxation time between the two bands. Indeed, the intraband relaxation time is typically far shorter ($< 10^{-12}$ s) than the radiative electron–hole recombination time ($\approx 10^{-9}$ s). Under these conditions, the conduction-band electrons achieve thermal equilibrium among themselves, as do the valence-band holes, but the electrons and holes are not in mutual thermal equilibrium. A state of quasi-equilibrium is created, for example, when an external electric current or photon flux applied to a semiconductor induces band-to-band transitions at a rate greater than that allowing interband equilibrium to be attained.

Under these circumstances, it is reasonable to use a distinct Fermi function for each band. The associated Fermi levels, known as **quasi-Fermi levels**, are denoted E_{fc} and E_{fv} for the conduction and valence bands, respectively (Fig. 5.4-6). Given the electron and hole concentrations, analytical expressions for E_{fc} and E_{fv} may be readily obtained (see below). When the quasi-Fermi levels lie well inside the conduction and valence bands, respectively, the concentrations of *both* electrons and holes can be quite large.

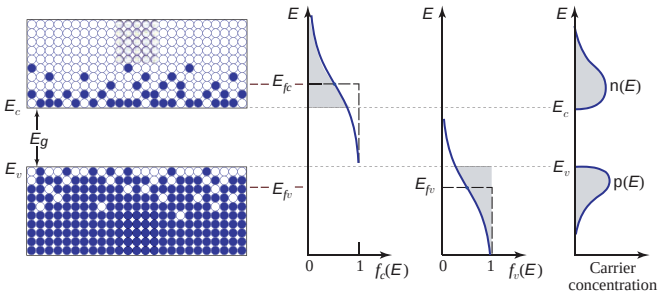


Figure 5.4-6 Carriers in thermal quasi-equilibrium. The conduction- and valence-band Fermi functions $f_c(E)$ and $f_v(E)$ have Fermi levels E_{fc} and E_{fv}, respectively. The electron and hole concentrations, denoted $\mathrm{n}(E)$ and $\mathrm{p}(E)$, respectively, can then both be large.

☐ **Determination of the Quasi-Fermi Levels Given the Electron and Hole Concentrations.**

(a) At $T = 0$ K, the Fermi function is expressed as $f_c(E) = 1$ for $E < E_{fc}$ and 0 otherwise. Using this expression in conjunction with (5.4-2) and (5.4-7) leads to the following integral for (5.4-8): $\mathrm{n} = \int_{E_c}^{E_{fc}} A\,(E - E_c)^{1/2}\,dE = \frac{2}{3}A\,(E_{fc} - E_c)^{3/2}$, where $A = (2m_c)^{3/2}/2\pi^2\hbar^3$ is a constant. It follows that $E_{fc} - E_c = (3\mathrm{n}/2A)^{2/3}$, which leads to the following quasi-Fermi levels (the derivation for E_{fv} parallels that for E_{fc}):

$$E_{fc} = E_c + (3\pi^2)^{2/3}\frac{\hbar^2}{2m_c}\,\mathrm{n}^{2/3}, \tag{5.4-16a}$$

$$E_{fv} = E_v - (3\pi^2)^{2/3}\frac{\hbar^2}{2m_v}\,\mathrm{p}^{2/3}. \tag{5.4-16b}$$

(b) For $T > 0$ K, (5.4-16a) and (5.4-16b) remain approximately valid, provided that the quasi-Fermi levels lie within the conduction and valence bands, away from the band edges, i.e., when n and p are sufficiently large such that $E_{fc} - E_c \gg kT$ and $E_v - E_{fv} \gg kT$. The function $f_c(E)$ then undergoes a smooth, rather than sharp, transition from unity to zero at E_{fc}, as illustrated by the solid curve in the middle panel of Fig. 5.4-7. The product $\varrho_c(E)f_c(E)$, depicted as the solid curve in the right-hand panel of Fig. 5.4-7, is then also a smooth curve. Nevertheless, the area under this curve, which represents the concentration n, does not deviate substantially from the area under the dashed curve, which is applicable for $T = 0$ K. Consequently, (5.4-16a) remains approximately applicable for $T > 0$ K, as does (5.4-16b) via a parallel argument.

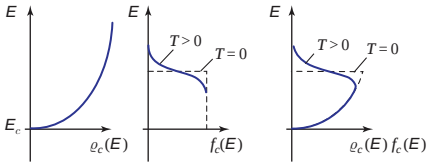


Figure 5.4-7 Plotted along the abscissas, as functions of E on the ordinates, are the density of states $\varrho_c(E)$, the Fermi function $f_c(E)$, and their product $\varrho_c(E)f_c(E)$.

■

5.5 GENERATION, RECOMBINATION, AND INJECTION

Generation and Recombination in Thermal Equilibrium

The thermal excitation of electrons from the valence band to the conduction band results in **electron–hole generation** (Fig. 5.5-1). Thermal equilibrium requires that this generation process be accompanied by a simultaneous reverse process of de-excitation. Called **electron–hole recombination**, this latter process occurs when an electron decays from the conduction band to fill a hole in the valence band (Fig. 5.5-1). The energy released in this process may take the form of an emitted photon, in which case the process is known as **radiative recombination** and the emitted light is referred to as **recombination radiation**.

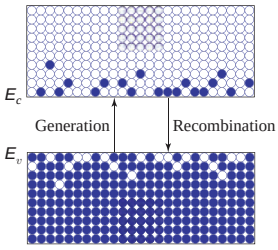


Figure 5.5-1 Electron–hole generation (upward arrow) and recombination (downward arrow).

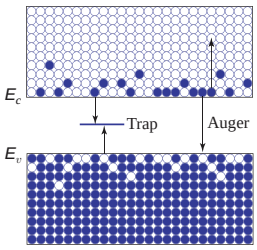


Figure 5.5-2 Electron–hole recombination via a trap and via Auger recombination.

Nonradiative recombination, an alternate process, can occur via a number of independent and competing processes. These include the transfer of energy to lattice vibrations, which creates one or more phonons, or to another free electron via **Auger recombination**, a three-particle interaction that can occur when the carrier density is sufficiently high (Fig. 5.5-2). Recombination can also take place at surfaces, and at impurity or defect centers located at grain boundaries, dislocations, or other lattice imperfections. A defect center whose energy lies within the forbidden band can facilitate recombination if it is capable of trapping both an electron and a hole (Fig. 5.5-2). Impurity-assisted recombination may be radiative or nonradiative.

Because both an electron *and* a hole are required for a recombination to occur, the rate of recombination is proportional to the product of their concentrations, i.e.,

$$\text{rate of recombination} = \mathtt{r}np. \tag{5.5-1}$$

The **recombination coefficient** $\mathtt{r}$ (cm^3/s) in (5.5-1) is dependent on both the characteristics of the material, including its composition and defect density, and its temperature. It also depends on the doping level, although relatively weakly.

The concentrations of electrons and holes in steady state, n_0 and p_0, respectively, are established when the generation and recombination rates are in balance. If G_0 is the rate of electron–hole generation in thermal equilibrium at a given temperature, we then have

$$G_0 = \mathrm{r}\mathrm{n}_0\mathrm{p}_0. \tag{5.5-2}$$

The product of the electron and hole concentrations $\mathrm{n}_0\mathrm{p}_0 = G_0/\mathrm{r}$ is approximately the same whether the material is n-type, p-type, or intrinsic. Hence, $\mathrm{n}_\mathrm{i}^2 \approx G_0/\mathrm{r}$, which leads to the law of mass action $\mathrm{n}_0\mathrm{p}_0 = \mathrm{n}_\mathrm{i}^2$. This highlights the fact that this law follows from the balance between generation and recombination in thermal equilibrium.

Electron–Hole Injection

A semiconductor in thermal equilibrium with carrier concentrations n_0 and p_0 has equal rates of generation and recombination, $G_0 = \mathrm{r}\mathrm{n}_0\mathrm{p}_0$. We now consider a configuration in which additional electron–hole pairs are generated at a steady rate R (pairs per unit volume per unit time) by means of an external (nonthermal) injection mechanism, such as light falling on the material. A new steady state will be reached in which the carrier concentrations are $\mathrm{n} = \mathrm{n}_0 + \Delta\mathrm{n}$ and $\mathrm{p} = \mathrm{p}_0 + \Delta\mathrm{p}$. It remains clear, however, that $\Delta\mathrm{n} = \Delta\mathrm{p}$ since the electrons and holes are created in pairs.

Equating the new rates of generation and recombination, we have

$$G_0 + R = \mathrm{r}\mathrm{n}\mathrm{p}. \tag{5.5-3}$$

Substituting $G_0 = \mathrm{r}\mathrm{n}_0\mathrm{p}_0$ from (5.5-2) into (5.5-3) then leads to

$$R = \mathrm{r}(\mathrm{n}\mathrm{p} - \mathrm{n}_0\mathrm{p}_0) = \mathrm{r}\left(\mathrm{n}_0\Delta\mathrm{n} + \mathrm{p}_0\Delta\mathrm{n} + \Delta\mathrm{n}^2\right) = \mathrm{r}\Delta\mathrm{n}(\mathrm{n}_0 + \mathrm{p}_0 + \Delta\mathrm{n}), \tag{5.5-4}$$

which we write in the form

$$R = \frac{\Delta\mathrm{n}}{\tau}, \tag{5.5-5}$$

where the recombination lifetime is

$$\tau = \frac{1}{\mathrm{r}\left[(\mathrm{n}_0 + \mathrm{p}_0) + \Delta\mathrm{n}\right]}. \tag{5.5-6}$$

For weak injection such that $\Delta\mathrm{n} \ll \mathrm{n}_0 + \mathrm{p}_0$, (5.5-6) reduces to

$$\boxed{\tau \approx \frac{1}{\mathrm{r}(\mathrm{n}_0 + \mathrm{p}_0)}.} \tag{5.5-7}$$

Excess-Carrier Recombination Lifetime

Hence, in an n-type material, where $\mathrm{n}_0 \gg \mathrm{p}_0$, the recombination lifetime $\tau \approx 1/\mathrm{r}\mathrm{n}_0$ is inversely proportional to the electron carrier concentration. Similarly, for a p-type material where $\mathrm{p}_0 \gg \mathrm{n}_0$, we obtain $\tau \approx 1/\mathrm{r}\mathrm{p}_0$. However, this simple formulation is not applicable when traps play a significant role in the process.

The parameter τ expressed in (5.5-6) and (5.5-7) may be regarded as the **electron–hole recombination lifetime** of the injected excess electron–hole pairs. This is readily understood by noting that the injected-carrier concentration is governed by the rate equation

$$\frac{d(\Delta\mathrm{n})}{dt} = R - \frac{\Delta\mathrm{n}}{\tau}. \tag{5.5-8}$$

In steady state, $d(\Delta\mathrm{n})/dt = 0$ and (5.5-5) is recovered. If the source of injection is suddenly removed ($R \to 0$) at time t_0, we obtain $\Delta\mathrm{n}(t) = \Delta\mathrm{n}(t_0)\exp[-(t - t_0)/\tau]$,

indicating that $\Delta \mathrm{n}$ decays exponentially with time constant τ. In the presence of strong injection, however, τ is itself a function of $\Delta \mathrm{n}$, as evident in (5.5-6), so that the rate equation is nonlinear and the decay is no longer exponential.

If the injection rate R is known, the steady-state injected concentration may be determined from

$$\boxed{\Delta \mathrm{n} = R\tau,} \tag{5.5-9}$$

thereby permitting the total concentrations $\mathrm{n} = \mathrm{n}_0 + \Delta \mathrm{n}$ and $\mathrm{p} = \mathrm{p}_0 + \Delta \mathrm{n}$ to be established. Furthermore, if thermal quasi-equilibrium is assumed, (5.4-8) may be used to determine the quasi-Fermi levels. Quasi-equilibrium is not inconsistent with the steady-state balance of generation and recombination assumed in the above analysis; it simply requires that the intraband equilibrium time be short in comparison with the recombination time τ. The same type of analysis is useful in developing the theory of the semiconductor light-emitting diode, which is based on enhancing light emission by means of carrier injection, as will become clear in Sec. 6.3.

EXAMPLE 5.5-1. ***Parameters Associated with Electron–Hole Pair Injection in GaAs.*** Typical parameters are set forth for an n-type sample of GaAs at $T = 300$ K, whose thermal equilibrium concentration of electrons is given by $\mathrm{n}_0 = 10^{16}/\mathrm{cm}^3$. The sample is injected with electron–hole pairs at a rate $R = 10^{23}/\mathrm{cm}^3\text{-s}$ and the recombination coefficient is taken to be $\mathrm{r} = 10^{-11}$ cm^3/s. The material characteristics of the GaAs sample under consideration are: $E_g = 1.42$ eV, $m_c \approx 0.07\, m_0$, and $m_v \approx 0.50\, m_0$ (Tables 5.2-1 and 5.3-1).

(a) *The equilibrium concentration of holes* p_0: Using $\mathrm{n}_\mathrm{i} = 1.8 \times 10^6$ cm^{-3} (Table 5.4-1), together with $\mathrm{n}_0 = 10^{16}$ cm^{-3}, the law of mass action (5.4-15) provides $\mathrm{p}_0 = \mathrm{n}_\mathrm{i}^2/\mathrm{n}_0 = 3.24 \times 10^{-4}$ cm^{-3}. We therefore have $\mathrm{n}_0 \gg \mathrm{p}_0$ for this n-type material.

(b) *The steady-state excess concentration* $\Delta \mathrm{n}$: With an injection rate $R = 10^{23}$ cm^{-3}s^{-1}, the steady-state concentrations are determined from (5.5-4), which provides: $R = \mathrm{r}(\mathrm{np} - \mathrm{n}_0\mathrm{p}_0) = \mathrm{r}\Delta \mathrm{n}(\mathrm{n}_0 + \mathrm{p}_0 + \Delta \mathrm{n}) \approx \mathrm{r}\Delta \mathrm{n}(\mathrm{n}_0 + \Delta \mathrm{n})$, so that $\Delta \mathrm{n}^2 + \mathrm{n}_0 \Delta \mathrm{n} - R/\mathrm{r} = 0$. Solving this quadratic equation for $\Delta \mathrm{n}$ yields $\Delta \mathrm{n} = \frac{1}{2}\left[-\mathrm{n}_0 + (\mathrm{n}_0^2 + 4R/\mathrm{r})^{1/2}\right] = 9.5 \times 10^{16}$ cm^{-3}. We conclude that $\Delta \mathrm{n}$ is a factor of 9.5 greater than n_0.

(c) *The recombination lifetime* τ: Using $\mathrm{n}_0 = 10^{16}$ cm^{-3} and $\Delta \mathrm{n} = 9.5 \times 10^{16}$ cm^{-3} in (5.5-6) yields $\tau \approx 952$ ns.

(d) *The separation between the quasi-Fermi levels,* $E_{fc} - E_{fv}$, *assuming that* $T = 0$ K: This quantity may be determined by subtracting (5.4-16b) from (5.4-16a), which provides $E_{fc} - E_{fv} = E_g + (3\pi^2)^{2/3}(\hbar^2/2)[\mathrm{n}^{2/3}/m_c + \mathrm{p}^{2/3}/m_v]$. Following the conversion of the values for $\mathrm{n} = \mathrm{n}_0 + \Delta \mathrm{n}$ and $\mathrm{p} = \mathrm{p}_0 + \Delta \mathrm{n} \approx \Delta \mathrm{n}$ obtained above from cm^{-3} to m^{-3} by multiplying them by 10^6, and dividing by the electronic charge e to convert J to eV, substitution in this equation yields

$$\begin{aligned} E_{fc} - E_{fv} &= E_g + \frac{(3\pi^2)^{2/3}}{2}\frac{\hbar^2}{m_0 e}\left[\frac{(\mathrm{n} \times 10^6)^{2/3}}{0.07} + \frac{(\mathrm{p} \times 10^6)^{2/3}}{0.5}\right] \\ &= E_g + 4.785 \cdot \frac{43.8 \times 10^{-68}}{5.74 \times 10^{-48}}\left[\frac{22.3 \times 10^{14}}{0.07} + \frac{20.8 \times 10^{14}}{0.5}\right] = E_g + 0.013 \text{ eV}. \end{aligned}$$

The result indicates that $E_{fc} - E_{fv}$ is greater than the bandgap energy E_g by 0.013 eV so that $E_{fc} - E_{fv} = 1.433$ eV. Using (5.4-16a) and (5.4-16b) individually then provides that $E_{fc} - E_c \approx 0.011$ eV and $E_v - E_{fv} \approx 0.002$ eV, revealing that both quasi-Fermi levels lie within, but very close to the edges of, the conduction and valence bands. However, since neither $E_{fc} - E_c$ nor $E_v - E_{fv}$ are $\gg kT = 0.026$ eV at $T = 300$ K, (5.4-16a) and (5.4-16b) should not be used for calculating the carrier concentration at $T = 300$ K, thereby clarifying why $T = 0$ K was expressly specified for this part of the example.

Internal Quantum Efficiency

The recombination coefficient r set forth in (5.5-1) is generally split into a sum of radiative and nonradiative parts, $\mathrm{r_r}$ and $\mathrm{r_{nr}}$, respectively, so that $\mathrm{r} = \mathrm{r_r} + \mathrm{r_{nr}}$. The **internal quantum efficiency (IQE)** of a semiconductor material is defined as the ratio of the radiative electron–hole recombination coefficient to the overall (radiative plus nonradiative) coefficient, i.e.,

$$\eta_{\mathtt{IQE}} = \frac{\mathrm{r_r}}{\mathrm{r}} = \frac{\mathrm{r_r}}{\mathrm{r_r} + \mathrm{r_{nr}}} \,. \tag{5.5-10}$$

The internal quantum efficiency is important because it establishes the efficiency of light generation internal to a semiconductor material. It is maximized by simultaneously making the radiative and nonradiative recombination coefficients as large and as small as possible, respectively. Radiative recombination is facilitated by spatially confining the electrons and holes in a common location.

Given that (5.5-6) specifies that the recombination lifetime τ is inversely proportional to r, the internal quantum efficiency can alternatively be written in terms of τ. Defining the radiative and nonradiative lifetimes as τ_r and τ_nr, respectively, leads to

$$1/\tau = 1/\tau_\mathrm{r} + 1/\tau_\mathrm{nr} \,. \tag{5.5-11}$$

Since (5.5-10) provides that $\eta_{\mathtt{IQE}} = \mathrm{r_r}/\mathrm{r} = (1/\tau_\mathrm{r})/(1/\tau)$, we arrive at

$$\boxed{\eta_{\mathtt{IQE}} = \frac{\tau}{\tau_\mathrm{r}} = \frac{\tau_\mathrm{nr}}{\tau_\mathrm{r} + \tau_\mathrm{nr}} \,.} \tag{5.5-12}$$

Internal Quantum Efficiency

The radiative recombination lifetime τ_r is a particularly important parameter in semiconductor photonics since it governs the rates of photon emission and photon absorption, as will be discussed in Sec. 6.3.

For low to moderate injection rates, the radiative version of (5.5-7) is

$$\tau_\mathrm{r} \approx \frac{1}{\mathrm{r_r}(\mathrm{n_0} + \mathrm{p_0})} \,, \tag{5.5-13}$$

which demonstrates that the radiative lifetime depends on both the material parameter $\mathrm{r_r}$ and the carrier concentrations. The nonradiative lifetime obeys a similar equation, but if defect centers in the forbidden band contribute to the recombination, the lifetime is generally more sensitive to the defect-center concentrations than to the electron and hole concentrations. Table 5.5-1 lists representative recombination parameters for Si, GaAs, GaN, and InGaN.

Table 5.5-1 Order-of-magnitude values for the radiative recombination coefficient $\mathrm{r_r}$; the radiative, nonradiative, and overall recombination lifetimes, τ_r, τ_nr, and τ, respectively; and the internal quantum efficiency $\eta_{\mathtt{IQE}}$, for several widely used semiconductor materials.[a]

SEMICONDUCTOR	$\mathrm{r_r}$ ($\mathrm{cm^3/s}$)	τ_r	τ_nr	τ	$\eta_{\mathtt{IQE}}$
Si	10^{-15}	10 ms	100 ns	100 ns	10^{-5}
GaAs	10^{-10}	100 ns	100 ns	50 ns	0.5
GaN[b]	10^{-8}	20 ns	0.1 ns	0.1 ns	0.005

[a] Approximate values are provided for n-type materials with a carrier concentration $\mathrm{n_0} = 10^{17}/\mathrm{cm^3}$ and defect centers with a concentration $10^{15}/\mathrm{cm^3}$, at $T = 300$ K.
[b] InGaN, which is used in practice, has a substantially larger IQE, namely $\eta_{\mathtt{IQE}} \approx 0.3$.

Examination of the entries in Table 5.5-1 reveals that the nonradiative lifetime for bulk (indirect-bandgap) Si is orders of magnitude shorter than its radiative lifetime, and this is responsible for its small internal quantum efficiency. For (direct-bandgap) GaAs and InGaN, on the other hand, the radiative and nonradiative lifetimes are far more similar, which leads to substantial values of the internal quantum efficiency. It will become apparent in the sequel that direct-bandgap materials are widely used for fabricating light-emitting structures that operate via interband transitions, while indirect-bandgap materials are not, except under special conditions.

5.6 JUNCTIONS AND HETEROJUNCTIONS

Juxtapositions of differently doped regions of a single semiconductor material are called **homojunctions**. An important example is the p–n junction, which is discussed in this section. Junctions between different semiconductor materials, known as **heterojunctions**, are discussed subsequently.

The p–n Junction

The p–n junction is a homojunction between a p-type and an n-type semiconductor. It acts as a diode, which can serve in electronics as a rectifier, logic gate, voltage regulator (Zener diode), or tuner (varactor diode); and in photonics as a light-emitting diode (LED), laser diode (LD), photodetector, or solar cell.

A p–n junction consists of a p-type and an n-type section of the same semiconductor materials in metallurgical contact. The p-type region has an abundance of holes (majority carriers) and few mobile electrons (minority carriers); the n-type region has an abundance of mobile electrons and few holes (Fig. 5.6-1). Both charge carriers are in continuous random thermal motion in all directions.

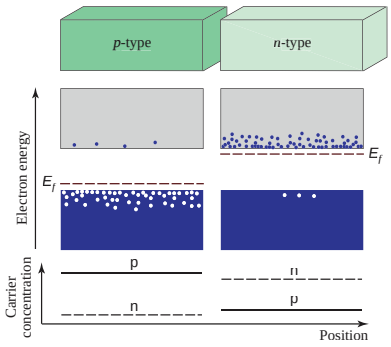


Figure 5.6-1 Energy levels and carrier concentrations for a p-type and an n-type semiconductor before contact.

When the two regions are brought into contact (Fig. 5.6-2), the following sequence of events takes place:

- Electrons and holes diffuse from areas of high concentration toward areas of low concentration. Thus, electrons diffuse from the n-region into the p-region, leaving behind positively charged ionized donor atoms. In the p-region the electrons recombine with the abundant holes. Similarly, holes diffuse from the p-region into the n-region, leaving behind negatively charged ionized acceptor atoms. In the n-region the holes recombine with the abundant mobile electrons. This diffusion process does not continue indefinitely, however, because it causes a disruption of the charge balance in the two regions.

- As a result, a narrow region on both sides of the junction becomes nearly depleted of *mobile* charge carriers. This region is called the **depletion layer**. It contains only the *fixed* charges (positive ions on the n-side and negative ions on the p-side). The thickness of the depletion layer in each region is inversely proportional to the concentration of dopants in the region.
- The fixed charges create an electric field in the depletion layer that points from the n-side toward the p-side of the junction. This **built-in field** obstructs the diffusion of further mobile carriers through the junction region.
- An equilibrium condition is established that results in a net built-in potential difference V_0 between the two sides of the depletion layer, with the n-side exhibiting a higher potential than the p-side.
- The built-in potential provides a lower potential energy for an electron on the n-side relative to the p-side. As a result, the energy bands bend, as illustrated in Fig. 5.6-2. In thermal equilibrium there is only a single Fermi function for the entire structure so that the Fermi levels in the p- and n-regions must align.
- No *net* current flows across the junction. The currents associated with diffusion and built-in field (drift current) cancel for both the electrons and holes.

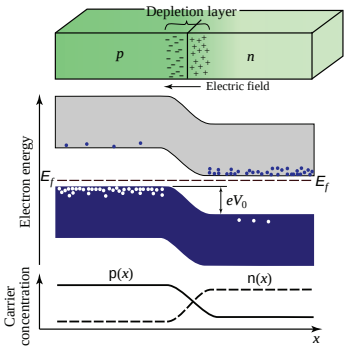


Figure 5.6-2 A p–n junction in thermal equilibrium at $T > 0$ K. The depletion-layer, energy-band diagram, and carrier concentrations (on a logarithmic scale) of mobile electrons $\mathtt{n}(x)$ and holes $\mathtt{p}(x)$ are shown as functions of the position x. The built-in potential difference V_0 corresponds to an energy eV_0, where e is the electron charge.

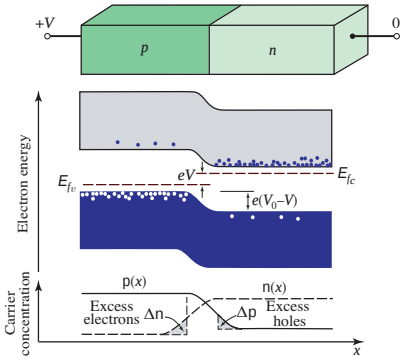


Figure 5.6-3 A forward-biased p–n junction in thermal quasi-equilibrium that sports two quasi-Fermi levels in the depletion layer, E_{fc} and E_{fv}. The energy-band diagram and concentrations $\mathtt{n}(x)$ and holes $\mathtt{p}(x)$ are displayed as functions of the position x. The forward bias reduces the height of the potential-energy hill by eV.

The Biased p–n Junction

An externally applied potential will alter the potential difference between the p- and n-regions. This in turn will modify the flow of majority carriers, so that the junction can be used as a "gate." If the junction is **forward biased** by applying a positive voltage V to the p-region (Fig. 5.6-3), its potential is increased with respect to the n-region, so that an electric field is produced in a direction opposite to that of the built-in field. The presence of the external bias voltage causes a departure from equilibrium and a misalignment of the Fermi levels in the p- and n-regions, as well as in the depletion layer. The presence of two Fermi levels in the depletion layer, E_{fc} and E_{fv}, represents a state of thermal quasi-equilibrium.

The net effect of the forward bias is to reduce the height of the potential-energy hill by an amount eV. The majority carrier current turns out to increase by an exponential

factor $\exp(eV/kT)$ so that the net current becomes $i = i_s \exp(eV/kT) - i_s$, where i_s is a constant. The excess majority carrier holes and electrons that enter the n- and p-regions, respectively, become minority carriers and recombine with the local majority carriers. Their concentration therefore decreases with distance from the junction as shown in Fig. 5.6-3. This process is known as **minority carrier injection**.

If the junction is **reverse biased** by applying a negative voltage V to the p-region, the height of the potential-energy hill is augmented by eV. This impedes the flow of majority carriers. The corresponding current is multiplied by the exponential factor $\exp(eV/kT)$, where V is negative; i.e., it is reduced. The net result for the current is $i = i_s \exp(eV/kT) - i_s$, so that a small current of magnitude $\approx i_s$ flows in the reverse direction when $|V| \gg kT/e$.

A p–n junction therefore acts as a diode with a current–voltage (i–V) characteristic

$$i = i_s \left[\exp\left(\frac{eV}{kT}\right) - 1 \right], \tag{5.6-1}$$

Ideal Diode Characteristic

as illustrated in Fig. 5.6-4. The ideal diode characteristic in (5.6-1) is known as the **Shockley equation**.

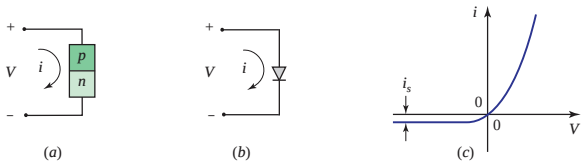


Figure 5.6-4 (a) Voltage and current in a p–n junction. (b) Circuit representation of the p–n junction diode. (c) Current–voltage characteristic of the ideal p–n junction diode.

The response of a p–n junction to a dynamic (AC) applied voltage is determined by solving the set of differential equations governing the processes of electron and hole diffusion, drift (under the influence of the built-in and external electric fields), and recombination. These effects are important for determining the speed at which the diode can be operated. They may be conveniently modeled by two capacitances, a junction capacitance and diffusion capacitance, in parallel with an ideal diode. The **junction capacitance** accounts for the time necessary to change the fixed positive and negative charges stored in the depletion layer when the applied voltage changes. The thickness l of the depletion layer turns out to be proportional to $\sqrt{V_0 - V}$; it therefore increases under reverse-bias conditions (negative V) and decreases under forward-bias conditions (positive V). The junction capacitance $C = \epsilon A/l$ (where A is the area of the junction) is therefore inversely proportional to $\sqrt{V_0 - V}$. The junction capacitance of a reverse-biased diode is smaller (and the RC response time is therefore shorter) than that of a forward-biased diode. The dependence of C on V is used to make voltage-variable capacitors (varactors).

Minority carrier injection in a forward-biased diode is described by the **diffusion capacitance**, which depends on the minority carrier lifetime and the operating current.

Heterojunctions

Junctions between different semiconductor materials are known as heterojunctions. Optical sources and detectors make extensive use of heterojunctions in their designs; they are used not only as active regions but also as contact layers and waveguiding regions.

The electron affinities of the materials determine the alignments of the conduction- and valence-band edges. It is often advantageous to lattice match the semiconductor materials and to make use of graded junctions rather than abrupt ones. The juxtaposition of different semiconductors can have manifold advantages in photonics:

- Junctions between materials of different bandgap create localized jumps in the energy-band diagram, as portrayed in Fig. 5.6-5. A potential-energy discontinuity provides a barrier that can be useful in preventing selected charge carriers from entering regions where they are undesired. This property may be used in a p–n junction, for example, to reduce the proportion of current carried by minority carriers, and thus to increase injection efficiency.

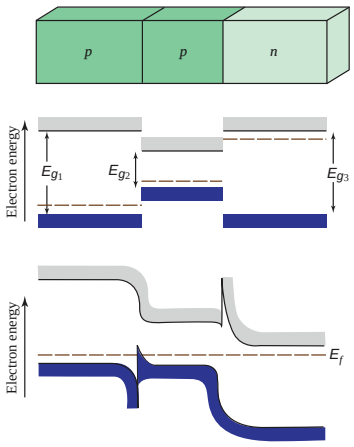


Figure 5.6-5 The p–p–n double heterojunction structure. The middle layer is of narrower bandgap than the outer layers. In equilibrium, the Fermi levels align so that the edge of the conduction band drops sharply at the p–p junction and the edge of the valence band drops sharply at the p–n junction. The conduction- and valence-band discontinuities are known as **band offsets**. When the device is forward biased, these jumps act as barriers that confine the injected minority carriers to the region of lower bandgap. Electrons injected from the n-region, for example, are prevented from diffusing beyond the barrier at the p–p junction. Similarly, holes injected from the p-region are not permitted to diffuse beyond the energy barrier at the p–n junction. This double-heterostructure configuration therefore forces electrons and holes to occupy a narrow common region. This substantially increases the efficiency of light-emitting diodes, as discussed in Chapter 7.

- Discontinuities in the energy-band diagram created by two heterojunctions can be useful for confining charge carriers to a desired region of space. For example, a layer of narrow-bandgap material can be sandwiched between two layers of a wider bandgap material, as shown in the p–p–n structure illustrated in Fig. 5.6-5 (which consists of a p–p heterojunction and a p–n heterojunction). This **double-heterostructure** (DH) configuration is used effectively in the fabrication of LEDs.
- Heterojunctions are useful for creating energy-band discontinuities that accelerate carriers at specific locations. The additional kinetic energy suddenly imparted to a carrier can be useful for selectively enhancing the probability of impact ionization in a multilayer avalanche photodiode.
- Semiconductors of different bandgap type (direct and indirect) can be used in the same device to select regions of the structure where light is emitted. Semiconductors of the direct-bandgap type emit light efficiently (Sec. 6.2).
- Semiconductors of different bandgaps can be used in the same device to select regions of the structure where light is absorbed. A semiconductor material whose bandgap energy is larger than the photon energy of light incident on it will be transparent, acting as a **window layer**.
- Heterojunctions of materials with different refractive indices can be used to create photonic structures and optical waveguides that confine and direct photons.

5.7 QUANTUM WELLS AND MULTIQUANTUM WELLS

Heterostructures of thin layers of semiconductor materials with specially designed band structures can be grown epitaxially (as layers of one semiconductor material over another) by using techniques such as molecular-beam epitaxy (MBE); liquid-phase epitaxy (LPE); and vapor-phase epitaxy (VPE), of which common variants are metalorganic chemical vapor deposition (MOCVD) and hydride vapor-phase epitaxy (HVPE). **Homoepitaxy** is the growth of materials that have the same composition as the substrate whereas **heteroepitaxy** is the growth of materials on a substrate of different composition, whether lattice-matched or not. MBE operates by directing molecular beams of the constituent elements to an appropriately prepared substrate in a high-vacuum environment, LPE uses the cooling of a saturated solution containing the constituents in contact with the substrate, and VPE uses gases in a reactor. The compositions and dopings of the individual layers, which can be made as thin as monolayers, are determined by manipulating the arrival rates of the molecules and the temperature of the substrate surface.

When the thickness of a layer is comparable to, or smaller than, the de Broglie wavelength of a thermalized electron, the energy of an electron resident in the layer is quantized, whereupon the energy–momentum relation suitable for a bulk semiconductor material is no longer applicable. The de Broglie wavelength is expressed as $\lambda_{\mathrm{dB}} = h/p$, where h is Planck's constant and p is the electron momentum ($\lambda_{\mathrm{dB}} \approx 50$ nm for GaAs). Three classes of such **quantum-confined structures** offer substantial advantages and are widely used in photonics: quantum wells and multiquantum wells, quantum wires, and quantum dots. The appropriate energy–momentum relations for these structures are derived below and in Sec. 5.8. The use of quantum-confined structures in optical sources is considered further in Secs. 6.5, 6.6, and 7.3.

A **quantum-well** structure, such as displayed in Fig. 5.7-1, is a double heterostructure consisting of an ultrathin ($\lesssim$ 50 nm) layer of semiconductor material whose bandgap is smaller than that of the surrounding material (e.g., an ultrathin layer of GaAs surrounded by AlGaAs). The sandwich results in one-dimensional conduction- and valence-band rectangular potential wells, within which electrons and holes are confined, respectively.

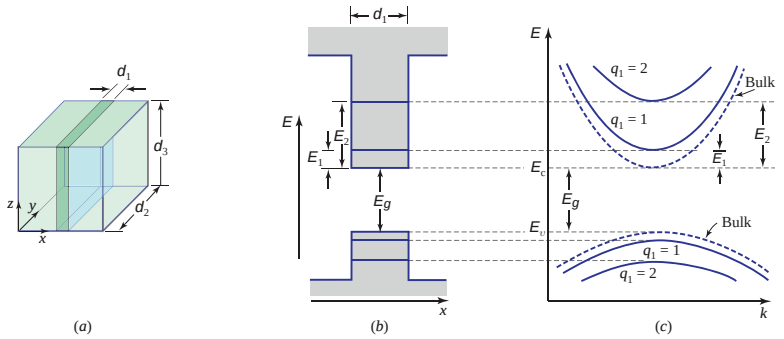


Figure 5.7-1 (a) Geometry of a quantum-well structure (e.g., an ultrathin layer of GaAs surrounded by AlGaAs). (b) Energy-level diagram for electrons and holes in their respective quantum wells. (c) Cross sections of the E–k relation in the direction of k_2 or k_3. The different energy subbands are designated by their quantum number $q_1 = 1, 2, \ldots$. The E–k relation for a bulk semiconductor material is indicated by the dashed curves.

A sufficiently deep potential well can be approximated by the well-known infinite rectangular potential well of quantum mechanics [Fig. 5.7-2 (left)]. The energy levels

E_q of a particle of mass m (m_c for electrons and m_v for holes) confined to a one-dimensional infinite rectangular well of full width d are determined by solving the time-independent Schrödinger equation. As shown in Example 5.7-1, the energy levels turn out to be

$$E_q = \frac{\hbar^2(q\pi/d)^2}{2m}, \qquad q = 1, 2, 3, \ldots . \tag{5.7-1}$$

As an example, the first three allowed energy levels of an electron in an infinitely deep GaAs quantum well ($m_c = 0.07\, m_0$) of width $d = 10$ nm are $E_q = 54$, 216, and 486 meV, respectively (recall that $kT = 26$ meV at $T = 300$ K). The smaller the width of the well, the larger the separation of adjacent energy levels.

EXAMPLE 5.7-1. ***Energy Levels of a Quantum Well.*** The allowed energies of an electron of mass m trapped in an infinitely deep, one-dimensional rectangular potential well are determined by solving the one-dimensional time-independent Schrödinger equation (5.2-1), $(-\hbar^2/2m)\, d^2\psi(x)/dx^2 + V(x)\psi(x) = E\psi(x)$, where $\psi(x)$ is the position wavefunction, E is the electron energy, and the potential energy $V(x) = 0$ for $0 < x < d$ and ∞ otherwise. This equation, which takes the form $d^2\psi(x)/dx^2 + k^2\psi(x) = 0$ with $k^2 = 2mE/\hbar^2$ inside the well, has the general solution $\psi(x) = A\sin(kx) + B\cos(kx)$.

Since the electron is trapped within the infinite walls of the well, $\psi(x) = 0$ at its boundaries, $x = 0$ and $x = d$, so that $B = 0$ and $\sin(kd) = 0$. This leads to $kd = q\pi$, where $q = 1, 2, 3, ...$, which indicates that k is restricted to the values $k_q = q\pi/d$, $q = 1, 2, 3, \ldots$. The electron energy $E = (\hbar^2/2m)k^2$ is thus quantized to the values $E_q = (\hbar^2/2m)(q\pi/d)^2$, $q = 1, 2, 3, \ldots$. The lowest three energy levels are therefore $E_1 = 4.9\hbar^2/md^2$, $E_2 = 19.7\hbar^2/md^2$, and $E_3 = 44.4\hbar^2/md^2$, as shown in Fig. 5.7-2(a).

By comparison, a quantum well of finite energy depth $V_0 = 32\hbar^2/md^2$ has energies: $E_1 = 3.2\hbar^2/md^2$, $E_2 = 11.9\hbar^2/md^2$, and $E_3 = 25.9\hbar^2/md^2$, as illustrated in Fig. 5.7-2(b). The finite well depth is seen to compress the energy-level spacings and to yield a continuum of energy levels above V_0.

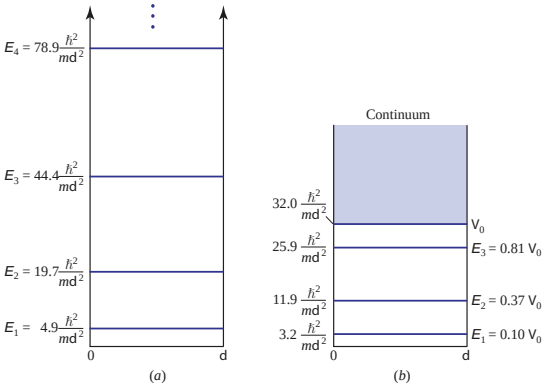


Figure 5.7-2 Energy levels of (a) a one-dimensional (1D) infinite rectangular potential well, and (b) a finite quantum square well with an energy depth $V_0 = 32\hbar^2/md^2$.

In actuality, however, semiconductor quantum wells are three-dimensional constructs. In the quantum-well structure shown in Fig. 5.7-1, electrons and holes are confined in the x direction to within a distance d_1 (the well thickness), but in the plane

of the confining layer they extend over far larger dimensions ($d_2, d_3 \gg d_1$). Hence, in the y–z plane, they behave as if they were in bulk semiconductor.

The electron energy–momentum relation is

$$E = E_c + \frac{\hbar^2 k_1^2}{2m_c} + \frac{\hbar^2 k_2^2}{2m_c} + \frac{\hbar^2 k_3^2}{2m_c}, \tag{5.7-2}$$

where $k_1 = q_1\pi/d_1$, $k_2 = q_2\pi/d_2$, $k_3 = q_3\pi/d_3$, and $q_1, q_2, q_3 = 1, 2, 3, \ldots$. Since $d_1 \ll d_2, d_3$, the parameter k_1 takes on well-separated discrete values, whereas k_2 and k_3 have finely spaced discrete values that may be approximated as a continuum. It follows that the energy–momentum relation for electrons in the conduction band of a quantum well is given by

$$E = E_c + E_{q1} + \frac{\hbar^2 k^2}{2m_c}, \qquad q_1 = 1, 2, 3, \ldots, \tag{5.7-3}$$

where k is the magnitude of a two-dimensional $\mathbf{k} = (k_2, k_3)$ vector in the y–z plane. Each quantum number q_1 corresponds to a **subband** whose lowest energy is $E_c + E_{q1}$. Similar relations apply for the valence band.

The energy–momentum relation for a bulk semiconductor is given by (5.2-6), where k is the magnitude of a three-dimensional vector $\mathbf{k} = (k_1, k_2, k_3)$. The key distinction is that for the quantum well, k_1 takes on well-separated, discrete values. As a result, the density of states associated with a quantum-well structure differs from that associated with bulk material, for which the density of states is determined from the magnitude of the three-dimensional vector with components $k_1 = q_1\pi/d$, $k_2 = q_2\pi/d$, and $k_3 = q_3\pi/d$ for $d_1 = d_2 = d_3 = d$. The result is $\varrho(k) = k^2/\pi^2$ per unit volume as per (5.4-1), which yields the density of conduction-band states [see (5.4-2) and Fig. 5.4-1]

$$\varrho_c(E) = \frac{\sqrt{2}\, m_c^{3/2}}{\pi^2\hbar^3}\sqrt{E - E_c}, \qquad E > 0. \tag{5.7-4}$$

In a quantum-well structure the density of states is derived from the magnitude of the *two-dimensional* vector (k_2, k_3). For each quantum number q_1 the density of states is therefore $\varrho(k) = k/\pi$ states per unit area in the y–z plane, and therefore $k/\pi d_1$ per unit volume. The densities $\varrho_c(E)$ and $\varrho(k)$ are related by $\varrho_c(E)\, dE = \varrho(k)\, dk = (k/\pi d_1)\, dk$. Finally, using the E–k relation (5.7-3) we obtain $dE/dk = \hbar^2 k/m_c$, from which

$$\varrho_c(E) = \begin{cases} \dfrac{m_c}{\pi\hbar^2 d_1}, & E > E_c + E_{q1} \\ 0, & E < E_c + E_{q1}, \end{cases} \qquad q_1 = 1, 2, 3, \ldots. \tag{5.7-5}$$

Thus, for each quantum number q_1, the density of states per unit volume is constant when $E > E_c + E_{q1}$. The overall density of states is the sum of the densities for all values of q_1, so that it exhibits the staircase distribution shown in Fig. 5.7-3. Each step of the staircase corresponds to a different quantum number q_1 and may be regarded as a subband within the conduction band (Fig. 5.7-1). The bottoms of these subbands move progressively higher for higher quantum numbers. It can be shown by substituting $E = E_c + E_{q1}$ in (5.7-4), and by using (5.7-1), that at $E = E_c + E_{q1}$ the quantum-well

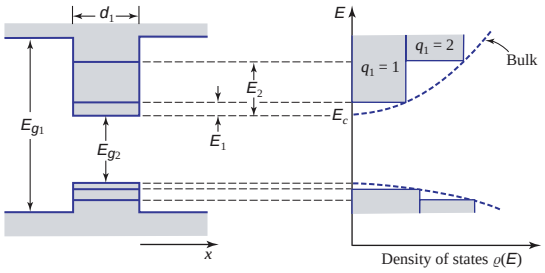


Figure 5.7-3 Density of states for a quantum-well structure (solid curve) and for a bulk semiconductor (dashed curve).

density of states is the same as that for the bulk material. The density of states in the valence band has a similar staircase distribution.

In distinction to bulk semiconductors, quantum-well structures exhibit substantial densities of states at the lowest allowed conduction-band energy level and at the highest allowed valence-band energy level. This property has a significant influence on the optical characteristics of the material.

Multiquantum Wells and Superlattices

Multiquantum Wells. The energy bandgap in a semiconductor quantum-well structure can be engineered to vary with position in any manner desired, which gives rise to materials with unique electronic and optical properties. Such structures can take many forms, including **multiquantum-well (MQW)** structures fabricated from alternating layers of materials of different bandgaps, as illustrated in Fig. 5.7-4 (left). The number of layers can stretch from just a few to hundreds, and still remain thin. For example, a MQW structure with 100 layers, each of thickness $\approx$ 10 nm and containing some 40 atomic planes, results in an overall thickness $\approx 1\ \mu$m.

Alternating layers of AlGaAs and GaAs are often used to fabricate MQW structures because these materials can be lattice matched over a broad range of compositions [Fig. 5.3-2(a)], thereby minimizing the strain between the two lattices. Moreover, the large difference in bandgap energies [Table 5.3-1] provides substantial carrier confinement. Other combinations of MQW materials commonly used in photonics include AlInAsSb/GaSb, AlInAs/InGaAs, AlInGaP/InGaP, GaN/InGaN, and $Al_xGa_{1-x}N/Al_yGa_{1-y}N$.

The hypothetical MQW structure portrayed in Fig. 5.7-4 (right) comprises a collection of ultrathin (2- to 15-nm-thick) layers of GaAs alternating with thin (20-nm-thick) layers of AlGaAs (well widths can be periodic or arbitrary). Since bandgap of GaAs is smaller than that of AlGaAs, the shapes of the conduction and valence bands are similar to the finite square-well potentials encountered in elementary quantum mechanics. Hence, for carrier motion perpendicular to the GaAs layers, the allowed energy levels for electrons in the conduction band, as well as for holes in the valence band, are discrete and well separated (see Example 5.7-1). The lowest energy levels are schematically indicated in the quantum wells displayed in Fig. 5.7-4.

Superlattices. If the AlGaAs barrier regions in Fig. 5.7-4 are made sufficiently thin ($<$ 1 nm), the electrons in adjacent wells can couple with each other via quantum-mechanical tunneling, whereupon the discrete energy levels displayed in Fig. 5.7-4 broaden into miniature bands called **minibands** separated by **minigaps**. The material is then referred to as a **superlattice structure** because the associated lattice is "super to" (i.e., larger than) the atomic crystal lattice, resulting in minibands rather than the full-size energy bands arising from the latter. Hence, the transition from MQW subbands to superlattice minibands is analogous to the transition from discrete energy levels in an atom to energy bands in a solid resulting from atomic interactions when the atoms are

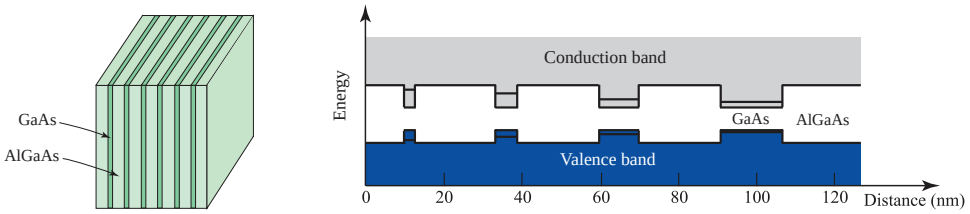


Figure 5.7-4 *Left*: Multiquantum-well (MQW) structure fabricated from alternating layers of two materials with different bandgaps, in this case AlGaAs and GaAs. The number of layers can stretch from several to hundreds. *Right*: Quantized energy levels in a hypothetical AlGaAs/GaAs single-crystal MQW structure.

brought into close proximity, as displayed in Figs. 5.1-1 and 5.1-2. Quantum wells and superlattices can also be created by spatially modulating the doping of a material, thus creating space-charge fields that form potential barriers.

Biased Multiquantum-Well and Superlattice Structures. The energy-band diagrams of unbiased and biased multiquantum-well and superlattice structures are sketched in Fig. 5.7-5. The electric field resulting from the applied bias causes the wells to become canted and modifies the energy levels. In superlattice structures, the discrete energy levels smear into minibands and minigaps. Multiquantum-well structures are widely used in the fabrication of light-emitting diodes (Sec. 7.3), as well as in other photonic devices.

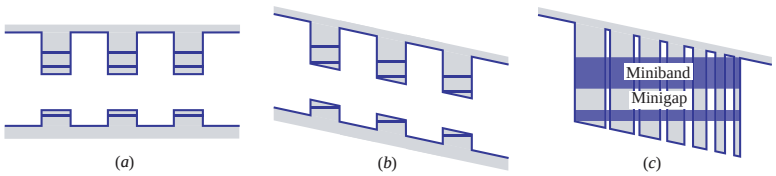


Figure 5.7-5 Energy-band diagrams of MQW and superlattice structures fabricated from alternating layers of materials with different bandgaps, such as AlGaAs and GaAs. (*a*) Unbiased MQW structure. (*b*) Biased MQW structure. (*c*) Biased superlattice structure with minibands and a minigap.

Quantum Wires

A semiconductor material that takes the form of a thin wire surrounded by a material of wider bandgap is known as a **quantum wire** structure (Fig. 5.7-6). The wire acts as a potential well that narrowly confines electrons (and holes) in the two lateral directions, x and y, but not in the direction along the axis of the wire. Quantum wires are readily made from III–V and II–VI semiconductors such as InP and CdSe, respectively; they usually have rectangular or circular cross section. Nanotubes and nanowires fabricated from a broad variety of materials can behave as quantum wires.

Carbon nanotubes, for example, are cylindrical carbon molecules with diameters of one or a few nm in which the carbon molecules organize themselves into thin hollow ropes held together by van der Waals forces. Single- or multiwalled nanotubes exhibit unique optical, mechanical, and electrical properties. They can behave as semiconductors or highly conductive metals, depending on the details of their structure. There are a multitude of uses for carbon nanotubes in photonics, including filaments for incandescent light sources.

Assuming that the quantum wire has a rectangular cross section of area $d_1 d_2$, the energy–momentum relation in the conduction band is

$$E = E_c + E_{q1} + E_{q2} + \frac{\hbar^2 k^2}{2m_c}, \tag{5.7-6}$$

where

$$E_{q1} = \frac{\hbar^2 (q_1 \pi / d_1)^2}{2m_c}, \qquad E_{q2} = \frac{\hbar^2 (q_2 \pi / d_2)^2}{2m_c}, \qquad q_1, q_2 = 1, 2, 3, \ldots, \tag{5.7-7}$$

and k is the vector component in the z direction (along the axis of the wire).

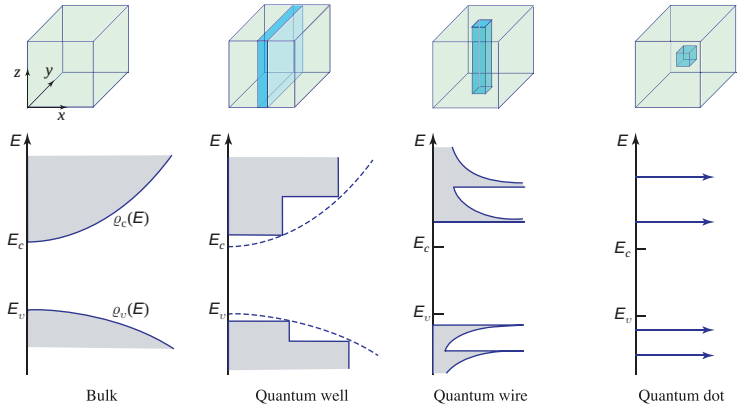


Figure 5.7-6 The density of states in different confinement configurations. The conduction and valence bands split into overlapping subbands that become successively narrower as the electron motion is restricted in a greater number of dimensions.

Each pair of quantum numbers (q_1, q_2) is associated with an energy subband that has a density of states $\varrho(k) = 1/\pi$ per unit length of the wire and therefore $1/\pi d_1 d_2$ per unit volume. The corresponding quantum-wire density of states (per unit volume), as a function of energy, is

$$\varrho_c(E) = \begin{cases} \dfrac{(1/d_1 d_2)\left(\sqrt{m_c}/\sqrt{2}\,\pi\hbar\right)}{\sqrt{E - E_c - E_{q1} - E_{q2}}}, & E > E_c + E_{q1} + E_{q2} \\ 0, & \text{otherwise,} \end{cases}$$
$$q_1, q_2 = 1, 2, 3, \ldots. \tag{5.7-8}$$

These are decreasing functions of energy, as illustrated in Fig. 5.7-6.

5.8 QUANTUM DOTS

Quantum dots (QDs) are semiconductor particles whose dimensions typically fall in the range 1–50 nm. Also known as **nanocrystals (NCs)** and **quantum boxes**, they can be fabricated from many materials and in a broad range of geometrical shapes, including

cubes, spheres, hemispheres, disks, and pyramids, depending on the growth mechanism and conditions. The 2023 Nobel Prize in Chemistry was awarded to Moungi Bawendi, Louis Brus, and Alexei Ekimov for the discovery and synthesis of quantum dots.

Fabrication

While QDs can be prepared in a variety of ways, three principal strategies are currently used for fabricating them:

1. High-temperature, vacuum-based approaches for fabricating highly crystalline **epitaxial quantum dots (eQDs)**.
2. Wet-chemical methods that enable the fabrication of **colloidal quantum dots (cQDs)** under mild conditions, and thereby large-area manufacturing.
3. **Electron-beam lithography** that allows nm-size patterns to be etched onto semiconductor chips, on which conducting metal can be deposited.

Innovative techniques for growing quantum dots continue to be developed.

Using wet chemistry, self-assembled cQDs with typical dimensions in the range 1–50 nm are formed from colloidal nanocrystals provided in liquid suspension or dispersed in a plastic composite. Chemical synthesis yields near-perfect crystalline clusters that range from several hundred to tens-of-thousands of atoms and that assume various shapes, again depending on growth conditions. cQDs are typically deposited onto substrates, or incorporated directly into devices designed to accommodate them, using liquid-phase processes such as spray coating, spin coating, microcontact printing, and inkjet printing, which are simple and cost-effective. Self-assembly can also be achieved by means of epitaxial synthesis, which can yield strained quantum-dot layers designed to improve device characteristics.

Colloidal quantum dots can be grown from II–VI, IV–VI, III–V, and group-IV semiconductors, as well as from organic compounds and perovskites.

Energy Levels

The number of atoms contained in a quantum dot, as well as its size, varies over a broad range; a 10-nm cube of GaAs, for example, contains some 40000 atoms. All electrons belong to the dot as a whole; the number of electrons can be as small as just a few or as large as millions.

As with atoms, a series of sharp energy levels results from tight electron confinement; quantum dots are in fact often called **artificial atoms**. Unlike atoms, however, a quantum dot fabricated from a given material has the property that its energy levels are strongly dependent on its size. Much as with the energy levels of an electron in a quantum well (Example 5.7-1), tight confinement in a small quantum dot corresponds to large energy-level differences, so that the wavelength of an emitted photon decreases as the size of the quantum dot diminishes. Since the electrons within the quantum dot are narrowly confined in all three dimensions, it can be modeled as a box of volume $d_1 d_2 d_3$ (Fig. 5.7-6). In accordance with (5.7-2), therefore, the energy is quantized to

$$E = E_c + E_{q1} + E_{q2} + E_{q3}, \tag{5.8-1}$$

where

$$E_{q1} = \frac{\hbar^2 (q_1 \pi / d_1)^2}{2m_c}, \qquad E_{q2} = \frac{\hbar^2 (q_2 \pi / d_2)^2}{2m_c}, \qquad E_{q3} = \frac{\hbar^2 (q_3 \pi / d_3)^2}{2m_c},$$
$$q_1, q_2, q_3 = 1, 2, 3, \ldots. \tag{5.8-2}$$

The allowed energy levels are discrete and well separated so that, as illustrated in Fig. 5.7-6, the density of states is represented by a sequence of delta functions at the

allowed energies. Although a quantum dot contains a vast number of strongly interacting natural atoms, its discrete energy levels can, in principle, be designed at will.

The energy levels of a quantum dot are those of its excitons, the electron–hole pairs strongly bound by Coulomb attraction and spin-exchange coupling that are generated within, and confined to, the dot.

Dot Size and Photoluminescence Wavelength

Figure 5.8-1(a) illustrates the color of the photoluminescence elicited from (II–VI) CdSe quantum dots as the dot size is gradually tuned from a diameter of 5 nm for red to 1.5 nm for violet. The photoexcitation wavelength is not a critical parameter, as long as it is shorter than the photoluminescence emission wavelength. An analogous illustration for $CsPbX_3$ metal-halide perovskite quantum dots is shown in Fig. 5.8-1(b). Photoexcited Si quantum dots can also emit light over a broad spectral range (Example 6.6-2).

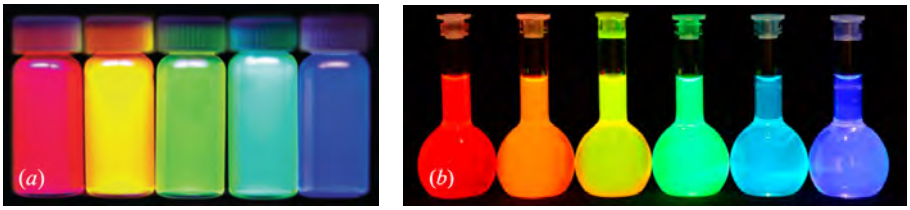


Figure 5.8-1 As a consequence of quantum confinement, the wavelength of the photoluminescence emission from quantum dots is governed by quantum-size effects as well as by composition. The photon energy increases, and the light wavelength decreases, from left to right in both panels. (a) Photoluminescence from II–VI CdSe colloidal quantum dots (with oleylamine surface capping molecules) dispersed in n-hexane. Photoexcitation is provided at $\lambda_0 = 365$ nm in the ultraviolet. (Courtesy of Don Seo, Arizona State University.) (b) Photoluminescence emission from $CsPbX_3$ metal-halide perovskite colloidal quantum dots dispersed in toluene, where X = I (red), Br (green), and Cl (violet). Photoexcitation is again provided at $\lambda_0 = 365$ nm. (Adapted from L. Protesescu, S. Yakunin, M. I. Bodnarchuk, F. Krieg, R. Caputo, C. H. Hendon, R. X. Yang, A. Walsh, and M. V. Kovalenko, Nanocrystals of Cesium Lead Halide Perovskites ($CsPbX_3$, X = Cl, Br, and I): Novel Optoelectronic Materials Showing Bright Emission with Wide Color Gamut, *Nano Letters*, vol. 15, pp. 3692–3696, 2015.)

Variations on the Theme

Quantum dots overcoated with semiconductor materials of higher bandgap are known as **core–shell quantum dots**. The presence of the shell reduces surface defects and provides increased luminescence, tunability, and lifespan. Quantum dots overcoated with multiple semiconductors of alternating higher and lower bandgaps are called **quantum well–quantum dots**. Quantum dots can also be embedded in semiconductor materials with larger bandgaps or in various solvents, inks, glasses, polymers, and matrices, including solid metal-halide perovskites. Arrays and self-assembled stacks of quantum dots are also readily fabricated. Ordered arrangements of quantum dots, called **quantum-dot solids**, can be grown by any number of methods, including the self-assembly into close-packed configurations. Quantum-dot solids called **nanocrystal superlattices** can support tunneling, much as multiquantum-well superlattices do (Fig. 5.7-5).

Quantum-dot structures brought into contact with electrodes behave as miniature photonic devices, examples being single-photon emitters (Sec. 6.6) and quantum-dot

light-emitting diodes (Sec. 7.5). As discussed above, quantum dots of different sizes provide operation over different wavelength ranges. QDs are also used for photonic applications such as displays, backlights, memory elements, spectral tags, and absorbers.

Quantum dots are not subject to the same material-quality limitations as single crystals, whose proper functioning generally depends on the absence of defects.

5.9 ORGANIC AND PEROVSKITE SEMICONDUCTORS

A brief introduction to organic and perovskite semiconductors is provided since both classes of materials are used to fabricate LEDs (Secs. 7.6 and 7.7, respectively).

Organic Semiconductors

Organic semiconductors are widely used in photonics for the fabrication of light-emitting diodes (Sec. 7.6), high-quality organic light-emitting displays, and photovoltaic devices. Organic materials typically have slower responses than their inorganic cousins, but they can be deposited via either solution processing or vacuum processing. Organic semiconductors can be inexpensively printed on thin plastic substrates using inkjet technology and can be engineered to suit specific requirements, such as mechanical flexibility.

The organic semiconductors used in photonics exist in two principal variants:

1. **Small organic molecules** such as pentacene, which comprises five linearly joined benzene rings [Fig. 5.9-1(a)].
2. **Conjugated polymer chains** such as polyacetylene, which comprises hundreds or thousands of carbon atoms [Fig. 5.9-1(b)].

Figure 5.9-1 Organic semiconductors for photonics exist in two main variants: (a) small organic molecules (e.g., pentacene), and (b) conjugated polymer chains (e.g., polyacetylene). (c) Doping polyacetylene with Na^+ donors yields an n-type material, whereas doping with I^- acceptors yields a p-type material. In the representation displayed in the figure, each vertex represents a carbon atom, each line represents a bond between two carbon atoms, and each double line represents a double bond. Hydrogen bonds are omitted for clarity. Various types of organic molecules and polymers find use in photonics.

The alternation of single and double carbon–carbon bonds, indicating *conjugation*, is common in organic semiconductors. While the double-bond electrons represented in Figs. 5.9-1(a) and (b) are portrayed as belonging to particular atoms, in actuality they are delocalized and shared among multiple atoms, or along a segment of the polymer that comprises some ten repeat units. The molecules, or polymer segments, behave as integrated systems in which the allowed electron states form bands.

When undoped, the valence band of a conjugated polymer chain is typically full, and its conduction band empty, so that it behaves as an insulator. Doping the polymer with sodium and iodine ions, which act as donors and acceptors, respectively, yield n-type and p-type variants, as illustrated in Fig. 5.9-1(c). Small organic molecules, on the other hand, are often conductive in their pure state.

Comparison of Organic and Inorganic Semiconductors. A number of fundamental features distinguish organic semiconductors from their inorganic counterparts:

- *Binding energies.* The constituent molecules are bound by weak van der Waals forces (with bond energies ≈ 0.1 eV) whereas the atoms in inorganic semiconductors are bound by strong covalent bonds (with bond energies ≈ 3 eV).
- *Mechanical flexibility.* Weak intermolecular bonds offer mechanical flexibility whereas inorganic semiconductors are rigid.
- *Energy bands.* Narrow energy bands derive from localized behavior at the molecular level while the broad energy bands of inorganic semiconductors derive from distributed behavior across the entire collection of atoms.
- *Energy levels.* Key energy levels are the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) whereas in inorganic semiconductors the conduction and valence bands are paramount.
- *Effective mass and carrier mobility.* Charge carriers are endowed with large effective mass ($m_c/m_0 \approx 500$) and small mobility ($\mu \approx 10^{-3}$ cm^2/V·s) whereas charge carriers in inorganic semiconductors have small effective mass ($m_c/m_0 < 1$) and large mobility ($\mu \approx 10^3$ cm^2/V·s).
- *Electron transport.* Intermolecular electron transfer occurs via hopping (phonon-assisted tunneling) whereas drift and diffusion characterize electron transport in inorganic semiconductors.
- *Conductivity.* The electrical conductivity is usually lower than that in inorganic semiconductors.
- *Moisture.* Organic materials are generally sensitive to moisture whereas inorganic semiconductors are not.

Organic Quantum Dots. Organic quantum dots, generally comprised of small organic molecules or conjugated polymer chains, can be synthesized via solution processing. Many of their features are similar to those of inorganic chalcogenide quantum dots (Sec. 5.8), but they generally exhibit inferior quantum efficiency, luminance, and stability and are therefore rarely used for fabricating LEDs.

Perovskite Semiconductors

Perovskites are compounds that are isostructural to calcium titanate ($CaTiO_3$), which is the mineral **perovskite**. Named in honor of the mineralogist Lev von Perovski, these compounds are identified by the chemical formula ABX_3 (A = Ca, B = Ti, and X = O for $CaTiO_3$). In their cubic (α) phase, perovskites exhibit the unit cell sketched in Fig. 5.9-2. Depending on the details of their composition and preparation, perovskites can exhibit photovoltaic, pyroelectric, piezoelectric, ferroelectric, photorefractive, and superconducting properties.

Perovskites have a dizzying array of variants that include organic, hybrid organic-inorganic, and fully inorganic versions. They are available as single crystals, polycrystalline films, and collections of quantum dots (nanocrystals). Some perovskites exhibit large absorption coefficients, high carrier mobilities, and long carrier lifetimes, rendering them efficient absorbers and emitters of light and endowing them with superior charge-transport properties. Perovskites cast in the form of thin amorphous films or quantum dots are increasingly used in photonics for the fabrication of light-emitting diodes (Sec. 7.7). They are also used for displays and photovoltaic devices.

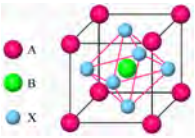


Figure 5.9-2 Single unit cell of the cubic (α) phase of a perovskite crystal with the chemical formula ABX_3. Inorganic metal-halide perovskites such as cesium lead triiodide ($CsPbI_3$), for which A = Cs, B = Pb, and X = I, are useful for fabricating high-efficiency perovskite light-emitting diodes (PeLEDs).

Perovskite Quantum Dots. Many of the features of perovskite quantum dots, including their amenability to synthesis by solution processing, resemble those of chalcogenide quantum dots (Sec. 5.8) and organic quantum dots (Sec. 5.9). Like organic quantum dots, their performance in terms of quantum efficiency, luminance, and stability is inferior to that of chalcogenide quantum dots. On the other hand, their performance is markedly superior to that of organic quantum dots and they are commonly used for fabricating LEDs.

BIBLIOGRAPHY

Semiconductor Materials, Growth, and Characterization

S. Pizzini, *Chemistry of Semiconductors*, Royal Society of Chemistry, 2024.

J. E. Ayers, T. Kujofsa, P. Rago, and J. Raphael, *Heteroepitaxy of Semiconductors: Theory, Growth, and Characterization*, CRC Press/Taylor & Francis, 2nd ed. 2017.

J. Orton and T. Foxon, *Molecular Beam Epitaxy: A Short History*, Oxford University Press, 2015.

D. K. Schroder, *Semiconductor Material and Device Characterization*, Wiley–IEEE, 3rd ed. 2015.

O. Oda, *Compound Semiconductor Bulk Materials and Characterizations*, Volume 2, World Scientific, 2012.

P. M. Koenraad and M. E. Flatté, Single Dopants in Semiconductors, *Nature Materials*, vol. 10, pp. 91–100, 2011.

M. Razeghi, *The MOCVD Challenge: A Survey of GaInAsP-InP and GaInAsP-GaAs for Photonic and Electronic Device Applications*, CRC Press/Taylor & Francis, 2010.

Graphene and 2D-Material Photonics

A. Wee, X. Yin, and C. S. Tang, *Two-Dimensional Transition-Metal Dichalcogenides: Phase Engineering and Applications in Electronics and Optoelectronics*, Wiley–VCH, 2024.

T. Zhang, *Graphene: From Theory to Applications*, Springer, 2022.

A. Diebold and T. Hofmann, *Optical and Electrical Properties of Nanoscale Materials*, Springer, 2022.

Y. Al-Douri, ed., *Graphene, Nanotubes and Quantum Dots-Based Nanotechnology: Fundamentals and Applications*, Woodhead/Elsevier, 2022.

Z. Zhang, Z. Kang, Q. Liao, and Y. Zhang, eds., *Van der Walls Heterostructures: Fabrications, Properties, and Applications*, Wiley–VCH, 2022.

S. Cahangirov, H. Sahin, G. Le Lay, and A. Rubio, *Introduction to the Physics of Silicene and Other 2D Materials*, Springer, 2017.

P. Avouris, T. F. Heinz, and T. Low, eds., *2D Materials: Properties and Devices*, Cambridge University Press, 2017.

P. Ajayan, P. Kim, and K. Banerjee, Two-dimensional van der Waals Materials, *Physics Today*, vol. 69, no. 9, pp. 38–44, 2016.

C. Jagadish and E. R. Weber, eds., *Semiconductors and Semimetals*, F. Iacopi, J. J. Boeckl, and C. Jagadish, eds., Volume 95, *2D Materials*, Academic/Elsevier, 2016.

E. L. Wolf, *Graphene: A New Paradigm in Condensed Matter and Device Physics*, Oxford University Press, 2016.

P. A. D. Gonçalves and N. M. R. Peres, *An Introduction to Graphene Plasmonics*, World Scientific, 2016.

A. C. Ferrari *et al.*, Science and Technology Roadmap for Graphene, Related Two-Dimensional Crystals, and Hybrid Systems, *Nanoscale*, vol. 7, pp. 4598–4810, 2015.

F. Xia, H. Wang, D. Xiao, M. Dubey, and A. Ramasubramaniam, Two-Dimensional Material Nanophotonics, *Nature Photonics*, vol. 8, pp. 899–907, 2014.

A. K. Geim, Nobel Lecture: Random Walk to Graphene, *Reviews of Modern Physics*, vol. 83, pp. 851–862, 2011.

K. S. Novoselov, Nobel Lecture: Graphene: Materials in the Flatland, *Reviews of Modern Physics*, vol. 83, pp. 837–849, 2011.

Semiconductor Physics and Devices

C. Hamaguchi, *Basic Semiconductor Physics*, Springer, 4th ed. 2023.

K. W. Böer and U. W. Pohl, *Semiconductor Physics*, Springer, 2nd ed. 2023.

A. Kitai, *Fundamentals of Semiconductor Materials and Devices*, Wiley, 2023.

E. F. Schubert, *Physical Foundations of Solid-State Devices*, Google Books, 2nd ed. 2022.

D. Jena, *Quantum Physics of Semiconductor Materials and Devices*, Oxford University Press, 2022.

S. M. Sze, Y. Li, and K. K. Ng, *Physics of Semiconductor Devices*, Wiley, 4th ed. 2021.

C. Jagadish and E. R. Weber, eds., *Semiconductors and Semimetals*, Z. Mi and C. Jagadish, eds., Volume 96, *III–Nitride Semiconductor Optoelectronics*, Academic/Elsevier, 2017.

M. Grundmann, *The Physics of Semiconductors: An Introduction Including Nanophysics and Applications*, Springer, 3rd ed. 2016.

M. Rudan, *Physics of Semiconductor Devices*, Springer, 2015.

B. G. Streetman and S. Banerjee, *Solid State Electronic Devices*, Pearson, 7th ed. 2014.

G. Grosso and G. V. Parravicini, *Solid State Physics*, Academic/Elsevier, 2nd ed. 2014.

S. H. Simon, *The Oxford Solid State Basics*, Oxford University Press, paperback ed. 2013.

W. Brütting and C. Adachi, eds., *Physics of Organic Semiconductors*, Wiley–VCH, 2nd ed. 2012.

D. A. Neamen, *Semiconductor Physics and Devices: Basic Principles*, McGraw–Hill, 4th ed. 2011.

J. Chu and A. Sher, *Device Physics of Narrow Gap Semiconductors*, Springer, 2010.

P. Yu and M. Cardona, *Fundamentals of Semiconductors: Physics and Materials Properties*, Springer, 4th ed. 2010.

C. Kittel, *Introduction to Solid State Physics*, Wiley, 8th ed. 2004.

K. K. Ng, *Complete Guide to Semiconductor Devices*, Wiley–IEEE, 2nd ed. 2002.

Quantum-Confined Materials and Nanostructures

W. Chiang, O. Morshed, and T. D. Krauss, *Quantum Confined Semiconductor Nanocrystals*, American Chemical Society, 2023.

R. Ameta, J. P. Bhatt, and S. C. Ameta, eds., *Quantum Dots: Fundamentals, Synthesis and Applications*, Woodhead/Elsevier, 2023.

S. Ganguly, P. Das, and J. Parameswaranpillai, eds., *Quantum Dots and Polymer Nanocomposites: Synthesis, Chemistry, and Applications*, CRC Press/Taylor & Francis, 2023.

P. K. Basu, B. Mukhopadhyay, and R. Basu, *Semiconductor Nanophotonics*, Oxford University Press, 2022.

V. B. Pawade, S. J. Dhoble, and H. C. Swart, *Nanoscale Compound Semiconductors and Their Optoelectronics Applications*, Woodhead/Elsevier, 2022.

F. P. García de Arquer, D. V. Talapin, V. I. Klimov, Y. Arakawa, M. Bayer, and E. H. Sargent, Semiconductor Quantum Dots: Technological Progress and Future Challenges, *Science*, vol. 373, eaaz8541, 2021.

I. Vurgaftman, M. P. Lumb, and J. R. Meyer, *Bands and Photons in III–V Semiconductor Quantum Structures*, Oxford University Press, 2021.

S. Y. Ren, *Electronic States in Crystals of Finite Size: Quantum Confinement of Bloch Waves*, Springer, 2nd ed. 2017.

M. V. Fischetti and W. G. Vandenberghe, *Advanced Physics of Electron Transport in Semiconductors and Nanostructures*, Springer, 2016.

P. Harrison and A. Valavanis, *Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures*, Wiley, 4th ed. 2016.

A. Zhang, G. Zheng, and C. M. Lieber, *Nanowires: Building Blocks for Nanoscience and Nanotechnology*, Springer, 2016.

C. Jagadish and E. R. Weber, eds., *Semiconductors and Semimetals*, S. A. Dayeh, A. Fontcuberta i Morral, and C. Jagadish, eds., Volume 94, *Semiconductor Nanowires II: Properties and Applications*, Academic/Elsevier, 2016.

C. R. Kagan, E. Lifshitz, E. H. Sargent, and D. V. Talapin, Building Devices from Colloidal Quantum Dots, *Science*, vol. 353, aac5523, 2016.

Y. S. Zhao, ed., *Organic Nanophotonics: Fundamentals and Applications*, Springer, 2015.

H. Ünlü and N. J. M. Horing, eds., *Low Dimensional Semiconductor Structures: Characterization, Modeling and Applications*, Springer, 2013.

D. Vollath, *Nanomaterials: An Introduction to Synthesis, Properties and Applications*, Wiley–VCH, 2nd ed. 2013.

G. Cao and Y. Wang, *Nanostructures and Nanomaterials: Synthesis, Properties, and Applications*, World Scientific, 2nd ed. 2011.

F. W. Wise, ed., *Selected Papers on Semiconductor Quantum Dots*, SPIE Optical Engineering Press (Milestone Series Volume 180), 2005.

F. T. Vasko and A. V. Kuznetsov, *Electronic States and Optical Transitions in Semiconductor Heterostructures*, Springer, 1999.

Historical Accounts and Seminal Publications

A. Nathan, S. K. Saha, and R. M. Todi, eds., *75th Anniversary of the Transistor*, Wiley/IEEE Press, 2023.

A. L. Efros and L. E. Brus, Nanocrystal Quantum Dots: From Discovery to Modern Development, *ACS Nano*, vol. 15, pp. 6192–6210, 2021.

C. B. Murray, D. J. Norris, and M. G. Bawendi, Synthesis and Characterization of Nearly Monodisperse CdE (E = Sulfur, Selenium, Tellurium) Semiconductor Nanocrystallites, *Journal of the American Chemical Society*, vol. 115, pp. 8706–8715, 2002.

M. Riordan and L. Hoddeson, The Origins of the pn Junction, *IEEE Spectrum*, vol. 34, no. 6, pp. 46–51, 1997.

L. Esaki, A Bird's-Eye View on the Evolution of Semiconductor Superlattices and Quantum Wells, *IEEE Journal of Quantum Electronics*, vol. QE-22, pp. 1611–1624, 1986.

L. Esaki, Long Journey into Tunneling (Nobel Lecture in Physics, 1973), in S. Lundqvist, ed., *Nobel Lectures in Physics 1971–1980*, World Scientific, 1992.

W. B. Shockley, Transistor Technology Evokes New Physics (Nobel Lecture in Physics, 1956), in *Nobel Lectures in Physics, 1942–1962*, World Scientific, 1998.

J. Bardeen, Semiconductor Research Leading to the Point Contact Transistor (Nobel Lecture in Physics, 1956), in *Nobel Lectures in Physics, 1942–1962*, World Scientific, 1998.

W. H. Brattain, Surface Properties of Semiconductors (Nobel Lecture in Physics, 1956), in *Nobel Lectures in Physics, 1942–1962*, World Scientific, 1998.

L. E. Brus, A Simple Model for The Ionization Potential, Electron Affinity, and Aqueous Redox Potentials of Small Semiconductor Crystallites, *Journal of Chemical Physics*, vol. 79, pp. 5566–5571, 1983.

A. Ekimov and A. A. Onushchenko, Quantum Size Effect in Three-Dimensional Microscopic Semiconductor Crystals, *Journal of Experimental and Theoretical Physics (JETP) Letters*, vol. 34, pp. 345–349, 1981.

W. H. Brattain and J. Bardeen, The Transistor, Nature of the Forward Current in Germanium Point Contacts, *Physical Review*, vol. 74, pp. 231–232, 1948.

J. Bardeen and W. H. Brattain, The Transistor, A Semi-Conductor Triode, *Physical Review*, vol. 74, pp. 230–231, 1948.

CHAPTER

6

SEMICONDUCTOR PHOTONICS

6.1 CARRIER TRANSITIONS IN BULK SEMICONDUCTORS 170

6.2 INTERBAND TRANSITIONS 171

6.3 ABSORPTION, EMISSION, AND GAIN 175

6.4 INJECTION ELECTROLUMINESCENCE 182

6.5 QUANTUM WELLS AND MULTIQUANTUM WELLS 192

6.6 QUANTUM-DOT SINGLE-PHOTON EMITTERS 193

6.7 REFRACTIVE INDEX 194

Henry Round (1881–1966), a British engineer working at Marconi's Wireless Telegraph Company, observed injection electroluminescence from a forward-biased SiC/point-contact Schottky-barrier diode in 1907.

The Russian physicist **Oleg Losev (1903–1942)** carried out extensive experiments on SiC/point-contact and ZnO Schottky-barrier diodes, and advanced the hypothesis that the emitted light was electroluminescence.

LED Lighting: Devices and Colorimetry. Malvin Carl Teich. Google Books. Published 2024.

This chapter provides an introduction to the properties of semiconductor materials and devices in connection with the absorption and emission of photons. Both bulk and quantum-confined semiconductors are considered; the latter offer a number of salutary features that make them useful for constructing photonic devices. This area of study is known as **semiconductor photonics** or **semiconductor optics**.

Photon absorption and emission are essential features of the operation of semiconductor photonic devices including photodetectors and LEDs:

- *The absorption of a photon can create an electron–hole pair.* Mobile charge carriers that result from the absorption of a photon alter the electrical properties of the semiconductor material. This process is the basis of operation of photoconductive photodetectors.
- *The recombination of an electron and a hole can result in the emission of a photon.* This process is responsible for the operation of semiconductor photon sources such as LEDs and LDs. LEDs emit spontaneous recombination radiation while LDs operate on the basis of stimulated recombination radiation.

The absorption and emission of photons occur as electrons and holes execute transitions between allowed energy levels within their prescribed energy bands (Sec. 6.1). A simplified theory of interband transitions that relies on the conservation of energy and momentum forms the theoretical underpinning for understanding absorption, spontaneous emission, and stimulated emission in bulk materials (Sec. 6.2). The basic rules that govern these interactions are set forth in Sec. 6.3. Although these rules are patterned on the approach used to describe the interaction of photons with atoms provided in Secs. 4.3–4.5, they differ in some respects because of the special properties of semiconductors, as discussed in Chapter 5. In Sec. 6.4, we focus on **injection electroluminescence**, the heart of LED functionality. This phenomenon was first observed in 1907 in a rudimentary point-contact diode, a device that is sometimes considered to be the prototypical LED. Finally, photon interactions in quantum wells and quantum dots are considered in Secs. 6.5 and 6.6, respectively, and a brief discussion of the refractive indices for semiconductors follows in Sec. 6.7.

6.1 CARRIER TRANSITIONS IN BULK SEMICONDUCTORS

A number of mechanisms can lead to the absorption and emission of photons in bulk semiconductors. The most important of these are:

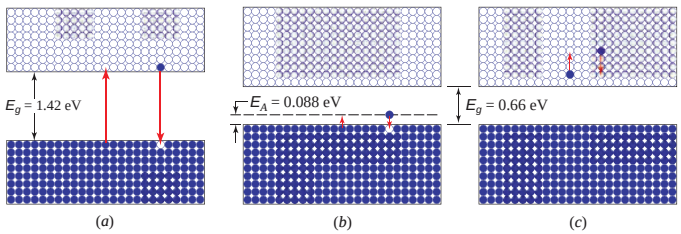


Figure 6.1-1 Examples of absorption and emission of photons in bulk semiconductors. (a) Band-to-band transitions in GaAs can result in the absorption or emission of photons of wavelength $\lambda_0 < \lambda_g = hc_0/E_g = 0.87\ \mu$m. ($b$) The absorption of a photon of wavelength $\lambda_A = hc_0/E_A = 14\ \mu$m results in a valence-band to acceptor-level transition in Hg-doped Ge (Ge:Hg). (c) Free-carrier transitions within the conduction band of Ge.

Band-to-Band (Interband) Transitions. An absorbed photon can result in an electron in the valence band making an upward transition to the conduction band, thereby creating an electron–hole pair [Fig. 6.1-1(a)]. Electron–hole recombination can result in the emission of a photon. Band-to-band transitions may be assisted by one or more phonons. A **phonon** is a quantum of the lattice vibrations associated with molecular or acoustic vibrations of the atoms in a material.

Impurity-to-Band Transitions. An absorbed photon can result in a transition between a donor (or acceptor) level and a band in a doped semiconductor. In a p-type material, for example, a low-energy photon can lift an electron from the valence band to the acceptor level, where it becomes trapped by an acceptor atom [Fig. 6.1-1(b)]. A hole is created in the valence band and the acceptor atom is ionized. Or a hole may be trapped by an ionized acceptor atom; the result is that the electron decays from its acceptor level to recombine with the hole. The energy may be released radiatively (in the form of an emitted photon) or nonradiatively (in the form of phonons). The transition may also be assisted by traps in defect states, as illustrated in Fig. 5.5-2.

Free-Carrier (Intraband) Transitions. An absorbed photon can impart its energy to an electron in a given band, causing it to move higher within that band. An electron in the conduction band, for example, can absorb a photon and move to a higher energy level within the conduction band [Fig. 6.1-1(c)]. This is followed by thermalization, a process whereby the electron relaxes down to the bottom of the conduction band while releasing its energy in the form of phonons. The strength of free-carrier absorption is proportional to the carrier density; it decreases with photon energy as a power-law function.

Phonon Transitions. Long-wavelength photons can release their energy by directly exciting lattice vibrations, i.e., by creating phonons.

Excitonic Transitions. The absorption of a photon in a semiconductor can result in the formation of a free electron in the conduction band and a hole that rises to the top of the valence band, where its energy is minimized. The hole and electron can be bound together by their mutual Coulomb attraction to form an **exciton**; the attractive potential results in a reduction of the total energy of the electron and hole. This entity is much like a hydrogen atom in which a hole plays the role of the proton. Excitons typically have lifetimes that range from hundreds of picoseconds to nanoseconds. A photon may be emitted as a result of the electron and hole recombining, thereby annihilating the exciton.

These transitions all contribute to the overall absorption coefficient, which is displayed in Fig. 6.1-2 for Si and GaAs, and at greater magnification in Fig. 6.1-3 for a number of semiconductor materials. For photon energies greater than the bandgap energy E_g, the absorption is dominated by band-to-band transitions that form the basis of many photonic devices. The spectral region where the material changes from being relatively transparent ($h\nu < E_g$) to strongly absorbing ($h\nu > E_g$) is known as the **absorption edge**. Direct-bandgap semiconductors have a more abrupt absorption edge than indirect-bandgap materials, as is apparent in Figs. 6.1-2 and 6.1-3.

6.2 INTERBAND TRANSITIONS

We proceed to develop a simple theory of direct interband (band-to-band) photon absorption and emission in bulk semiconductors, ignoring the other types of transitions.

Bandgap Wavelength

Direct interband absorption can take place only at light frequencies for which the photon energy $h\nu > E_g$. The minimum frequency at which this is possible is therefore $\nu_g = E_g/h$, which, in accordance with (5.1-1), corresponds to a maximum wavelength $\lambda_g =$

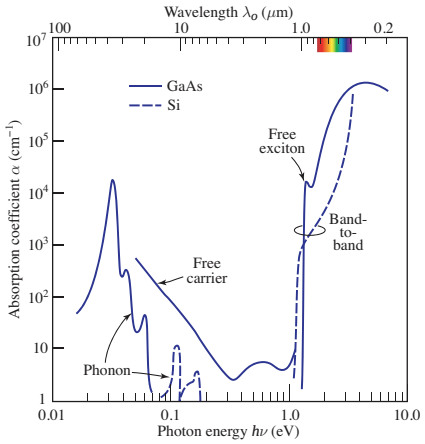


Figure 6.1-2 Observed optical absorption coefficient α versus photon energy and wavelength for Si and GaAs in thermal equilibrium at $T = 300$ K. The bandgap energy E_g is 1.12 eV for Si and 1.42 eV for GaAs. Silicon is relatively transparent in the band $\lambda_0 \approx 1.1$ to 12 μm, whereas intrinsic GaAs is relatively transparent in the band $\lambda_0 \approx 0.87$ to 12 μm.

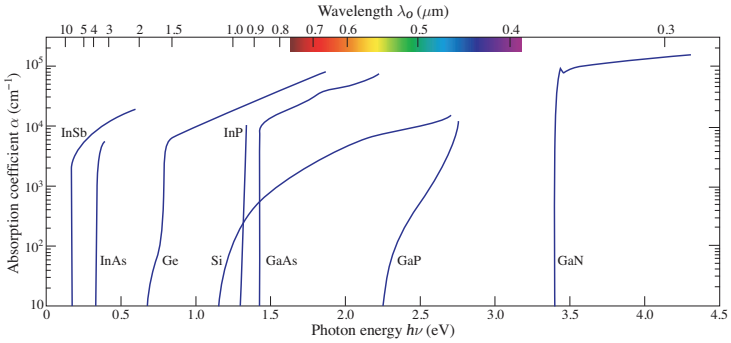


Figure 6.1-3 Absorption coefficient versus photon energy and wavelength for Ge, Si, GaAs, GaN, and several other III–V binary semiconductors at $T = 300$ K, on an expanded scale. Direct- and indirect-bandgap materials follow different functional forms near the band edge.

$c_0/\nu_g = hc_0/E_g$. (The bandgap wavelength λ_g is also called the *long-wavelength limit* in this context.) The approximate relationship provided in (5.1-2) is convenient for relating the bandgap energy and wavelength if they are specified in eV and μm, respectively. Bandgap wavelengths and energies are provided in Table 5.3-1 for selected elemental and binary III-V semiconductors, and can also be extracted from Figs. 5.3-2 and 5.3-3 for various compound semiconductors. III–V ternary and quaternary materials of various compositions span the visible region.

Conditions for Photon Absorption and Emission

Electron excitation from the valence to the conduction band may be induced by the absorption of a photon of appropriate energy ($h\nu > E_g$ or $\lambda < \lambda_g$). An electron–hole pair is generated [Fig. 6.2-1(a)]. This adds to the concentration of mobile charge carriers and increases the conductivity of the material. The material behaves as a photoconductor with a conductivity proportional to the photon flux. This effect is used for the photodetection of light.

Electron de-excitation from the conduction to the valence band (electron–hole recombination) may result in the spontaneous emission of a photon of energy $h\nu > E_g$ [Fig. 6.2-1(b)], or in the stimulated emission of a photon [Fig. 6.2-1(c)] when a photon of energy $h\nu > E_g$ is initially present (Sec. 4.3). Spontaneous emission is

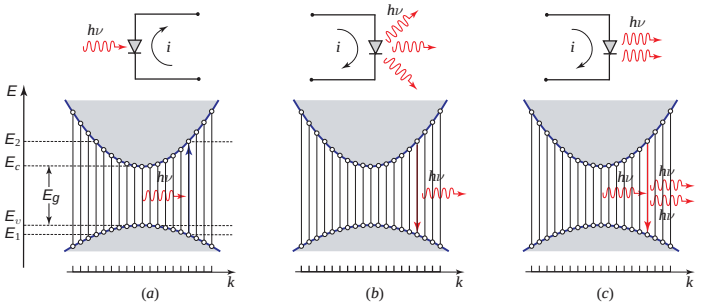


Figure 6.2-1 (a) The absorption of a photon results in the generation of an electron–hole pair. This process is used for the photodetection of light. (b) The recombination of an electron–hole pair results in the spontaneous emission of a photon. Light-emitting diodes (LEDs) operate on this basis. (c) Electron–hole recombination can be induced by a photon. The result is the stimulated emission of an identical photon. This is the underlying process responsible for the operation of semiconductor laser diodes (LDs).

the underlying phenomenon on which the light-emitting diode is based (Chapter 6). Stimulated emission is responsible for the operation of semiconductor laser diodes, as discussed in Sec. 7.8.

The conditions under which interband absorption and emission take place are summarized as follows:

Conservation of Energy. The absorption or emission of a photon of energy $h\nu$ requires that the energies of the two states involved in the interaction (say E_1 and E_2 in the valence band and conduction band, respectively, as depicted in Fig. 6.2-1) be separated by $h\nu$. Thus, for photon emission to occur by electron–hole recombination, for example, an electron occupying an energy level E_2 must interact with a hole occupying an energy level E_1, such that energy is conserved:

$$E_2 - E_1 = h\nu. \tag{6.2-1}$$

Conservation of Momentum. Momentum must also be conserved in the process of photon emission/absorption, so that $p_2 - p_1 = h\nu/c = h/\lambda$, or $k_2 - k_1 = 2\pi/\lambda$. The magnitude of the photon momentum h/λ is, however, very small in comparison with the range of momentum values that electrons and holes can assume. The semiconductor E–k diagram extends to values of k of the order $2\pi/a$, where the lattice constant a is much smaller than the wavelength λ, so that $2\pi/\lambda \ll 2\pi/a$. The momenta of the electron and the hole participating in the interaction must therefore be approximately equal. This condition, $k_2 \approx k_1$, is called the k**-selection rule**. Transitions that obey this rule are represented in the E–k diagram (Fig. 6.2-1) by vertical lines, indicating that the change in k is negligible on the scale of the diagram.

Energies and Momenta of the Electron and Hole with Which a Photon Interacts. As is apparent from Fig. 6.2-1, conservation of both energy and momentum requires that a photon of frequency ν interact with electrons and holes of specific energies and momenta determined by the semiconductor E–k relation. Using (5.2-6) and (5.2-7) to approximate this relation for a direct-bandgap semiconductor by two parabolas, and writing $E_c - E_v = E_g$, (6.2-1) may be written in the form

$$E_2 - E_1 = \frac{\hbar^2 k^2}{2m_v} + E_g + \frac{\hbar^2 k^2}{2m_c} = h\nu, \tag{6.2-2}$$

from which

$$k^2 = \frac{2m_r}{\hbar^2}(h\nu - E_g), \tag{6.2-3}$$

where

$$\frac{1}{m_r} = \frac{1}{m_v} + \frac{1}{m_c}. \tag{6.2-4}$$

Substituting (6.2-3) into (5.2-6) provides the energy levels E_1 and E_2 with which the photon interacts:

$$E_2 = E_c + \frac{m_r}{m_c}(h\nu - E_g), \tag{6.2-5}$$

$$E_1 = E_v - \frac{m_r}{m_v}(h\nu - E_g) = E_2 - h\nu. \tag{6.2-6}$$

In the special case when $m_c = m_v$, we obtain $E_2 = E_c + \frac{1}{2}(h\nu - E_g)$, as required by symmetry.

Optical Joint Density of States. We now determine the density of states $\varrho(\nu)$ with which a photon of energy $h\nu$ interacts under conditions of energy and momentum conservation in a direct-bandgap semiconductor. This quantity incorporates the density of states in both the conduction and valence bands and is known as the **optical joint density of states**. The one-to-one correspondence between E_2 and ν embodied in (6.2-5) permits $\varrho(\nu)$ to be related to the density of states $\varrho_c(E_2)$ in the conduction band by use of the incremental relation $\varrho_c(E_2)\,dE_2 = \varrho(\nu)\,d\nu$, from which $\varrho(\nu) = (dE_2/d\nu)\varrho_c(E_2)$, so that

$$\varrho(\nu) = \frac{hm_r}{m_c}\varrho_c(E_2). \tag{6.2-7}$$

Using (5.4-2) and (6.2-5), we finally obtain the number of interacting states per unit volume per unit frequency,

$$\varrho(\nu) = \frac{(2m_r)^{3/2}}{\pi\hbar^2}\sqrt{h\nu - E_g}, \qquad h\nu \geqslant E_g, \tag{6.2-8}$$

(6.2-8) Optical Joint Density of States

which is sketched in Fig. 6.2-2. The one-to-one correspondence between E_1 and ν in (6.2-6), together with $\varrho_v(E_1)$ from (5.4-3), results in an expression for $\varrho(\nu)$ identical to (6.2-8).

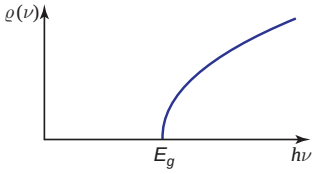


Figure 6.2-2 The density of states with which a photon of energy $h\nu$ interacts increases with $h\nu - E_g$ in accordance with a square-root law.

Photon Absorption is Not Unlikely in an Indirect-Bandgap Semiconductor. The energy and momentum conservation required for photon absorption in an indirect-bandgap semiconductor is readily accommodated by means of a two-step process (Fig. 6.2-3). The electron is first excited to a high energy level within the conduction band by a k-conserving vertical transition. It then quickly relaxes to the bottom of the conduction band by a process called **thermalization**, in which its momentum is transferred to phonons. The generated hole behaves similarly. Since the process occurs sequentially, it does not require the simultaneous presence of three bodies and is thus not unlikely in indirect- bandgap semiconductors. Indeed, Si and Ge are widely used as photodetector materials, as are direct-bandgap semiconductors such as AlGaAs and InGaAs.

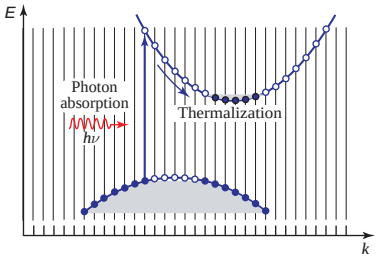


Figure 6.2-3 Photon absorption in an indirect-bandgap semiconductor via a vertical (k-conserving) transition. The photon generates an excited electron in the conduction band, leaving behind a hole in the valence band. The electron and hole then undergo fast transitions — to the lowest and highest available levels in the conduction and valence bands, respectively, releasing their energy in the form of phonons. Since the process is sequential, it is not unlikely.

Photon Emission is Unlikely in an Indirect-Bandgap Semiconductor. Radiative electron–hole recombination is unlikely in an indirect-bandgap semiconductor. This is because a transition from near the bottom of the conduction band to near the top of the valence band (where electrons and holes are most likely to reside, respectively) requires an exchange of momentum that cannot be accommodated by the emitted photon (Fig. 6.2-4). Momentum may be conserved, however, by the participation of phonons in the interaction. Phonons can carry relatively large momenta but typically have small energies ($\approx$ 0.01–0.1 eV; see Fig. 6.1-2), so that their transitions appear horizontal on the E–k diagram as portrayed in Fig. 6.2-4. The net result is that the k-selection rule is violated but momentum is conserved. However, because phonon-assisted emission involves the simultaneous participation of three bodies (electron, photon, and phonon), the probability of its occurrence is substantially reduced. Thus, Si, which is an indirect-bandgap semiconductor, has a substantially lower radiative recombination coefficient than does GaAs, which is a direct-bandgap semiconductor (Table 5.5-1). Silicon therefore does not emit light efficiently via interband transitions, whereas GaAs does. Under some circumstances, incorporating isoelectronic co-dopants such as N, Zn, or O into a material can mitigate momentum-convervation limitations. Because such co-dopants reside at sharply localized positions in the crystal, the uncertainty principle set forth in (A.2-9) of Appendix A endows them with the ability to accommodate substantial momentum changes such as those required for indirect transitions, but this approach frequently suffers from the effects of lattice-constant mismatch. Specific examples of the use of this technique are provided in connection with the discussion of GaAsP in Sec. 7.3.

6.3 ABSORPTION, EMISSION, AND GAIN

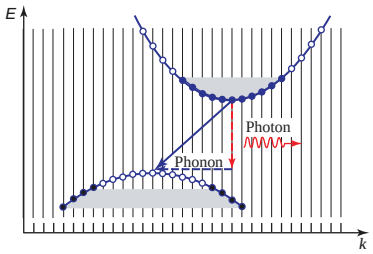


Figure 6.2-4 Photon emission via an interband transition in an indirect-bandgap semiconductor. The recombination of an electron near the bottom of the conduction band with a hole near the top of the valence band requires the exchange of energy and momentum. The energy may be carried off by a photon, but one or more phonons are required to conserve momentum. Such a simultaneous multiparticle interaction has a reduced likelihood of taking place.

We now proceed to determine the probability densities of a photon of energy $h\nu$ being emitted or absorbed by a bulk semiconductor material in a direct interband transition. Conservation of energy and momentum, in the form of (6.2-3), (6.2-5), and (6.2-6), determines the energies E_1 and E_2, and the momentum $\hbar k$, of the electrons and holes with which the photon may interact. Three factors determine these probability densities, as discussed below:

1. Occupancy probabilities
2. Transition probabilities
3. Optical joint density of states

Occupancy Probabilities

The occupancy conditions for photon emission and absorption by means of transitions between the discrete energy levels E_2 and E_1 are stated as follows:

Emission condition: A conduction-band state of energy E_2 is filled (with an electron) *and* a valence-band state of energy E_1 is empty (i.e., filled with a hole).

Absorption condition: A conduction-band state of energy E_2 is empty *and* a valence-band state of energy E_1 is filled.

The probabilities that these occupancy conditions are satisfied for various values of E_2 and E_1 are determined from the appropriate Fermi functions $f_c(E)$ and $f_v(E)$ associated with the conduction and valence bands of a semiconductor in thermal quasi-equilibrium. Thus, the probability $f_e(\nu)$ that the emission condition is satisfied for a photon of energy $h\nu$ is the product of the probabilities that the upper state is filled and that the lower state is empty since these are independent events, i.e.,

$$f_e(\nu) = f_c(E_2)\,[1 - f_v(E_1)]. \tag{6.3-1}$$

The energies E_1 and E_2 are related to ν by (6.2-5) and (6.2-6). Similarly, the probability $f_a(\nu)$ that the absorption condition is satisfied is

$$f_a(\nu) = [1 - f_c(E_2)]\,f_v(E_1). \tag{6.3-2}$$

☐ **Condition for the Photon Emission Rate to Exceed the Photon Absorption Rate.**

(a) For a bulk semiconductor in thermal equilibrium, the rate of photon emission cannot exceed the rate of photon absorption. In thermal equilibrium $E_{fc} = E_{fv} = E_f$ so that, in accordance with (5.4-4), $f(E) = 1/\{\exp[(E - E_f)/kT + 1]\}$. The difference between the emission

and absorption conditions, provided by (6.3-1) and (6.3-2), respectively, is $f_e(\nu) - f_a(\nu) = f_c(E_2) - f_v(E_1)$. Since $f_c(E) = f_v(E) = f(E)$ in thermal equilibrium, we have $f_e(\nu) - f_a(\nu) = f(E_2) - f(E_1)$. Because $f(E)$ is a monotonically decreasing function of E, we therefore obtain $f(E_2) < f(E_1)$ so that $f_e(\nu) - f_a(\nu) < 0$. Hence, $f_e(\nu) < f_a(\nu)$, which reveals that the rate of emission is smaller than the rate of absorption.

(b) For a semiconductor in thermal quasi-equilibrium ($E_{fc} \neq E_{fv}$), with radiative transitions occurring between a conduction-band state of energy E_2 and a valence-band state of energy E_1 with the same value of k, emission is more likely than absorption if the separation between the quasi-Fermi levels is larger than the photon energy, i.e., if

$$\boxed{E_{fc} - E_{fv} > h\nu.} \qquad \text{(6.3-3)}$$

Condition for Net Emission

In thermal quasi-equilibrium, $f_e(\nu) - f_a(\nu) = f_c(E_2) - f_v(E_1) = 1/\{1 + \exp[(E_2 - E_{fc})/kT]\} - 1/\{1 + \exp[(E_1 - E_{fv})/kT]\}$. This is a positive quantity if $\exp[(E_2 - E_{fc})/kT] < \exp[(E_1 - E_{fv})/kT]$, or equivalently if $E_2 - E_{fc} < E_1 - E_{fv}$ or if $E_2 - E_1 < E_{fc} - E_{fv}$. Since $E_2 - E_1 = h\nu$, the emission rate is greater than the absorption rate if $E_{fc} - E_{fv} > h\nu$. This implies that the separation between the two Fermi levels must be greater than the bandgap energy, i.e., that E_{fc} and E_{fv} must lie within the conduction and valence bands, respectively.

■

Transition Probabilities

Satisfying the emission/absorption occupancy condition does not assure that the emission/absorption actually takes place. These processes are governed by the probabilistic laws of interaction between photons and atomic systems examined at length in Sec. 4.3. As they relate to semiconductors, these laws are generally expressed in terms of emission into (or absorption from) a narrow band of frequencies between ν and $\nu + d\nu$:

Summary: Spontaneous & Stimulated Emission, and Absorption

A radiative transition between two discrete energy levels E_1 and E_2 is characterized by a transition cross section $\sigma(\nu) = (\lambda^2/8\pi t_{\rm sp})g(\nu)$, where ν is the frequency, $t_{\rm sp}$ is the effective spontaneous lifetime, and $g(\nu)$ is the lineshape function [centered about the transition frequency $\nu_0 = (E_2 - E_1)/h$, with transition linewidth $\Delta\nu$ and with unity area]. In semiconductors, the radiative electron–hole recombination lifetime $\tau_{\rm r}$, which was discussed in Sec. 5.5, plays the role of $t_{\rm sp}$ so that

$$\sigma(\nu) = \frac{\lambda^2}{8\pi\tau_{\rm r}}\, g(\nu)\,. \tag{6.3-4}$$

- If the occupancy condition for emission is satisfied, the probability density (per unit time) for the spontaneous emission of a photon into any of the available radiation modes in the narrow frequency band between ν and $\nu+d\nu$ is

$$P_{\rm sp}(\nu)\, d\nu = \frac{1}{\tau_{\rm r}}\, g(\nu)\, d\nu. \tag{6.3-5}$$

- If the occupancy condition for emission is satisfied *and* a mean spectral photon-flux density ϕ_ν (photons per unit time per unit area per unit frequency) at frequency ν is present, the probability density (per unit time) for the stimulated emission of one photon into the narrow frequency band between ν and $\nu + d\nu$ is

$$W_i(\nu)\, d\nu = \phi_\nu\, \sigma(\nu)\, d\nu = \phi_\nu \frac{\lambda^2}{8\pi\tau_{\rm r}}\, g(\nu)\, d\nu. \tag{6.3-6}$$

- If the occupancy condition for absorption is satisfied *and* a mean spectral photon-flux density ϕ_ν at frequency ν is present, the probability density for the absorption of one photon from the narrow frequency band between ν and $\nu + d\nu$ is also given by (6.3-6).

Since each transition has a different central frequency ν_0, and since we are considering a collection of such transitions, we explicitly label the central frequency of the transition by writing $g(\nu)$ as $g_{\nu 0}(\nu)$. In semiconductors the homogeneously broadened lineshape function $g_{\nu 0}(\nu)$ associated with a pair of energy levels generally has its origin in electron–phonon collision broadening. It therefore typically exhibits a Lorentzian lineshape [see (4.6-3)] of width $\Delta\nu \approx 1/\pi T_2$, where the electron–phonon collision time T_2 is of the order of picoseconds. If $T_2 = 1$ ps, for example, then $\Delta\nu = 318$ GHz, corresponding to an energy width $h\Delta\nu \approx 1.3$ meV. The radiative lifetime broadening of the levels is negligible in comparison with collisional broadening.

Overall Emission and Absorption Transition Rates

For a pair of energy levels separated by $E_2 - E_1 = h\nu_0$, the rates of spontaneous emission, stimulated emission, and absorption of photons of energy $h\nu$ (in units of photons/s-Hz-cm^3 of the semiconductor material), at the frequency ν, are obtained as follows: The appropriate transition probability density $P_{\rm sp}(\nu)$ or $W_i(\nu)$ [as provided in (6.3-5) or (6.3-6)] is multiplied by the appropriate occupation probability $f_e(\nu_0)$ or $f_a(\nu_0)$ [as given in (6.3-1) or (6.3-2)], and by the density of states that can interact with the photon $\varrho(\nu_0)$ [as set forth in (6.2-8)]. The overall transition rate for all allowed

frequencies is then calculated by integrating over ν_0.

The rate of spontaneous emission at frequency ν, for example, is given by

$$r_{\rm sp}(\nu) = \int [(1/\tau_{\rm r}) g_{\nu 0}(\nu)]\ f_e(\nu_0)\, \varrho(\nu_0)\, d\nu_0. \tag{6.3-7}$$

When the collision-broadened width $\Delta\nu$ is substantially smaller than the width of the product $f_e(\nu_0)\varrho(\nu_0)$, which is the situation that is usually encountered, $g_{\nu 0}(\nu)$ may be approximated by $\delta(\nu - \nu_0)$, whereupon the sifting property of the delta function simplifies (6.3-7) to $r_{\rm sp}(\nu) = (1/\tau_{\rm r})\varrho(\nu) f_e(\nu)$. The rates of stimulated emission and absorption are obtained in a similar manner, and result in the following formulas:

$$r_{\rm sp}(\nu) = \frac{1}{\tau_{\rm r}}\, \varrho(\nu) f_e(\nu) \tag{6.3-8}$$

$$r_{\rm st}(\nu) = \phi_\nu \frac{\lambda^2}{8\pi\tau_{\rm r}}\, \varrho(\nu) f_e(\nu) \tag{6.3-9}$$

$$r_{\rm ab}(\nu) = \phi_\nu \frac{\lambda^2}{8\pi\tau_{\rm r}}\, \varrho(\nu) f_a(\nu). \tag{6.3-10}$$

Emission and Absorption Rates

These equations, together with (6.2-8)–(6.3-2), permit the rates of spontaneous emission, stimulated emission, and absorption arising from direct interband transitions (photons/s-Hz-cm^3) to be calculated in the presence of a mean spectral photon-flux density ϕ_ν (photons/s-Hz-cm^2). The products $\varrho(\nu) f_e(\nu)$ and $\varrho(\nu) f_a(\nu)$ are analogous to the products of the lineshape function and atomic number densities in the upper and lower levels, $g(\nu)N_2$ and $g(\nu)N_1$, respectively, used to study emission and absorption in atomic systems.

The determination of the occupancy probabilities $f_e(\nu)$ and $f_a(\nu)$ requires knowledge of the quasi-Fermi levels E_{fc} and E_{fv}. It is via the control of these two parameters (by the application of an external bias to a p–n junction, for example) that the emission and absorption rates are modified to produce semiconductor photonic devices that carry out different functions. Equation (6.3-8) is the basic result that describes the operation of the light-emitting diode (LED), a semiconductor source based on spontaneous emission (Chapter 7). Equation (6.3-9) is applicable for semiconductor optical amplifiers and laser diodes, which operate on the basis of stimulated emission. Equation (6.3-10) is appropriate for semiconductor detectors that function by means of photon absorption.

Spontaneous-Emission Spectral Density in Thermal Equilibrium

A semiconductor in thermal equilibrium has only a single Fermi function so that (6.3-1) becomes $f_e(\nu) = f(E_2)[1 - f(E_1)]$. If the Fermi level lies within the bandgap, away from the band edges by at least several times kT, use may be made of the exponential approximations to the Fermi functions, $f(E_2) \approx \exp[-(E_2 - E_f)/kT]$ and $1 - f(E_1) \approx \exp[-(E_f - E_1)/kT]$, whereupon $f_e(\nu) \approx \exp[-(E_2 - E_1)/kT]$, i.e.,

$$f_e(\nu) \approx \exp\left(-\frac{h\nu}{kT}\right). \tag{6.3-11}$$

Substituting (6.2-8) for $\varrho(\nu)$, and (6.3-11) for $f_e(\nu)$, into (6.3-8) therefore provides

$$r_{\rm sp}(\nu) \approx D_0 \sqrt{h\nu - E_g}\, \exp\left(-\frac{h\nu - E_g}{kT}\right), \qquad h\nu \geqslant E_g, \tag{6.3-12}$$

where

$$D_0 = \frac{(2m_r)^{3/2}}{\pi\hbar^2\tau_{\rm r}} \exp\left(-\frac{E_g}{kT}\right) \tag{6.3-13}$$

is a parameter that increases with temperature at an exponential rate.

The spontaneous emission rate (6.3-12), which is sketched in Fig. 6.3-1 as a function of $h\nu$, has the functional form indicated by virtue of two factors: 1) a function associated with the density of states, which increases as the square-root of $h\nu - E_g$; and 2) an exponentially decreasing function of $h\nu - E_g$, which is associated with the Fermi function. Clearly, the spontaneous emission rate can be increased by augmenting $f_e(\nu)$. As provided in (6.3-1), this can be achieved by making $f_c(E_2)$ and $f_v(E_1)$ large and small, respectively, thereby causing the material to depart from thermal equilibrium. This in turn assures that a sufficient number of *both* electrons and holes is available in the junction region to attain the condition required for LED operation, as detailed in Chapter 7.

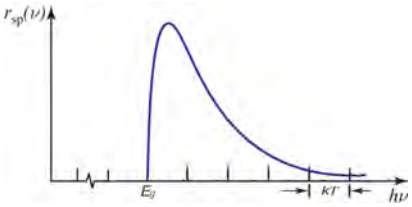


Figure 6.3-1 Direct interband spontaneous-emission spectral density $r_{\rm sp}(\nu)$ (photons/s-Hz-cm^3) for a semiconductor material in thermal equilibrium, plotted as a function of the photon energy $h\nu$. Emission is present for photon energies greater than the bandgap energy E_g, over a range of photon energies of approximate width $1.8kT$.

Gain Coefficient in Thermal Quasi-Equilibrium

The net gain coefficient $\gamma_0(\nu)$ corresponding to the rates of stimulated emission and absorption in (6.3-9) and (6.3-10) is determined by taking a cylinder of unit area and incremental length dz, and assuming that a mean spectral photon-flux density is directed along its axis. If $\phi_\nu(z)$ and $\phi_\nu(z) + d\phi_\nu(z)$ are the mean spectral photon-flux densities entering and leaving the cylinder, respectively, $d\phi_\nu(z)$ must be the mean spectral photon-flux density emitted from within the cylinder. The incremental number of photons, per unit time per unit frequency per unit area, is simply the number of photons gained, per unit time per unit frequency per unit volume $[r_{\rm st}(\nu) - r_{\rm ab}(\nu)]$, multiplied by the thickness of the cylinder dz. Hence, $d\phi_\nu(z) = [r_{\rm st}(\nu) - r_{\rm ab}(\nu)]\, dz$. Substituting the rates set forth in (6.3-9) and (6.3-10) leads to

$$\frac{d\phi_\nu(z)}{dz} = \frac{\lambda^2}{8\pi\tau_{\rm r}}\, \varrho(\nu)\, [f_e(\nu) - f_a(\nu)]\, \phi_\nu(z) = \gamma_0(\nu)\, \phi_\nu(z). \tag{6.3-14}$$

The net gain coefficient is therefore

$$\boxed{\gamma_0(\nu) = \frac{\lambda^2}{8\pi\tau_r}\,\varrho(\nu)\,f_g(\nu),} \qquad \begin{matrix}(6.3\text{-}15)\\ \text{Gain Coefficient}\end{matrix}$$

where the **Fermi inversion factor** $f_g(\nu)$ takes the form

$$f_g(\nu) \equiv f_e(\nu) - f_a(\nu) = f_c(E_2) - f_v(E_1), \tag{6.3-16}$$

as may be understood from (6.3-1) and (6.3-2), with E_1 and E_2 related to ν by (6.2-5) and (6.2-6). The quantity $\varrho(\nu)\,f_g(\nu)$ in the semiconductor system plays the role of $Ng(\nu)$ in the atomic system. Using (6.2-8), the gain coefficient may be cast in the form

$$\gamma_0(\nu) = D_1\sqrt{h\nu - E_g}\,f_g(\nu), \qquad h\nu > E_g \tag{6.3-17a}$$

with

$$D_1 = \frac{\sqrt{2}\,m_r^{3/2}\lambda^2}{h^2\tau_r}\,. \tag{6.3-17b}$$

The sign and spectral form of the Fermi inversion factor $f_g(\nu)$ are governed by the quasi-Fermi levels E_{fc} and E_{fv}, which in turn depend on the state of excitation of the carriers in the semiconductor. The condition for net emission provided in (6.3-3) shows that this factor is positive (corresponding to a population inversion and net gain) only when $E_{fc} - E_{fv} > h\nu$. When the semiconductor is pumped to a sufficiently high level by means of an external source of power, this condition may be satisfied and net gain achieved. This reflects the physics underlying the operation of semiconductor optical amplifiers and laser diodes.

Absorption Coefficient in Thermal Equilibrium

A semiconductor in thermal equilibrium has only a single Fermi level $E_f = E_{fc} = E_{fv}$, so that

$$f_c(E) = f_v(E) = f(E) = \frac{1}{\exp[(E - E_f)/kT] + 1}\,. \tag{6.3-18}$$

The factor $f_g(\nu) = f_c(E_2) - f_v(E_1) = f(E_2) - f(E_1) < 0$, and therefore the gain coefficient $\gamma_0(\nu)$ is always negative [since $E_2 > E_1$ and $f(E)$ decreases monotonically with E]. This is true whatever the location of the Fermi level E_f. Thus, a semiconductor in thermal equilibrium, whether it be intrinsic or doped, always attenuates light. The attenuation (absorption) coefficient, $\alpha(\nu) = -\gamma_0(\nu)$, is therefore

$$\boxed{\alpha(\nu) = D_1\sqrt{h\nu - E_g}\,\left[f(E_1) - f(E_2)\right],} \qquad \begin{matrix}(6.3\text{-}19)\\ \text{Absorption Coefficient}\end{matrix}$$

where E_2 and E_1 are given by (6.2-5) and (6.2-6), respectively, and D_1 is given by (6.3-17b).

If E_f lies within the bandgap but away from the band edges by an energy of at least several times kT, then $f(E_1) \approx 1$ and $f(E_2) \approx 0$ so that $[f(E_1) - f(E_2)] \approx 1$. In that case, the direct interband contribution to the absorption coefficient is

$$\alpha(\nu) \approx \frac{\sqrt{2}\,c^2 m_r^{3/2}}{\tau_r}\,\frac{1}{(h\nu)^2}\,\sqrt{h\nu - E_g}\,. \tag{6.3-20}$$

EXAMPLE 6.3-1. ***Absorption Coefficient for GaAs in Thermal Equilibrium.*** Equation (6.3-20) is plotted in Fig. 6.3-2 for GaAs, using the following parameters: $n = 3.6$, $m_c = 0.07\,m_0$, $m_v = 0.50\,m_0$, $m_0 = 9.1 \times 10^{-31}$ kg, a doping level such that $\tau_r = 0.4$ ns (this differs from that set forth in Table 5.5-1 because of the difference in doping level), $E_g = 1.42$ eV, and a temperature such that $[f(E_1) - f(E_2)] \approx 1$. As the temperature increases, $f(E_1) - f(E_2)$ decreases below unity and the absorption coefficient provided in (6.3-19) is reduced.

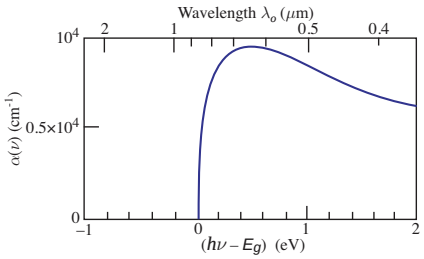


Figure 6.3-2 Calculated absorption coefficient $\alpha(\nu)$ (cm^{-1}) resulting from direct interband transitions, as a function of the photon energy $h\nu$ (eV) and the wavelength λ_0 (μm), for GaAs. This curve should be compared with the empirical result displayed in Fig. 6.1-3, which encompasses all absorption mechanisms.

In accordance with (6.3-20), absorption near the band edge in a direct-bandgap semiconductor should follow the functional form $\sqrt{h\nu - E_g}$. However, the sharp onset of absorption at $h\nu = E_g$ is an idealization. As is evident in Fig. 6.1-3, direct-bandgap semiconductors generally exhibit an exponential absorption tail, known as the **Urbach tail**, with a characteristic width $\approx kT$ that extends slightly into the forbidden band. This is associated with thermal and static disorder in the crystal arising from several factors, including phonon-assisted absorption, randomness in the doping distribution, and variations in material composition. Absorption near the band edge in indirect-bandgap semiconductors (e.g., Ge, Si, and GaP in Fig. 6.1-3) generally follows the functional form $(h\nu - E_g)^2$ rather than the square-root relation applicable for direct-bandgap semiconductors.

□ **Frequency and Wavelength of Maximum Interband Absorption in Thermal Equilibrium.** In accordance with (6.3-20), the absorption coefficient $\alpha(\nu)$ is proportional to $(h\nu - E_g)^{1/2}\,(h\nu)^{-2}$. This function attains its maximum value, ν_p, when its derivative with respect to ν is zero. This occurs when $-2(h\nu_p - E_g)^{1/2} + \frac{1}{2}\,h\nu_p(h\nu_p - E_g)^{-1/2} = 0$ or $\frac{1}{4}h\nu_p = h\nu_p - E_g$ so that $h\nu_p = \frac{4}{3}E_g$. To determine the (free-space) wavelength λ_p at which the absorption is maximized, $\alpha(\nu)$ must be written as $\alpha(\lambda_0)$, and the derivative with respect to λ_0 should be taken. Since $\nu = c_0/\lambda_0$, we have $\alpha(\lambda_0) \propto (hc_0/\lambda_0 - hc_0/\lambda_g)^{1/2}\,(\lambda_0/hc_0)^2 \propto (1/\lambda_0 - 1/\lambda_g)^{1/2}\,(\lambda_0)^2$. Setting the derivative of $\alpha(\lambda_0)$ equal to zero yields $2\,(1/\lambda_p - 1/\lambda_g)^{1/2}\,\lambda_p - \frac{1}{2}\,(1/\lambda_p - 1/\lambda_g)^{-1/2}\,(\lambda_p^2/\lambda_p^2) = 0$ whereupon $4\,(1/\lambda_p - 1/\lambda_g)\,\lambda_p = 1$, which in turn leads to $\lambda_p = \frac{3}{4}\lambda_g$ or $\lambda_p\,(\mu\text{m}) = \frac{3}{4} \cdot 1.24/E_g$ (eV). Although λ_p cannot in general be evaluated as c_0/ν_p, doing so in this case leads to $\frac{3}{4}c_0h/E_g$, which is the correct result. Using GaAs as an example, we have $E_g = 1.42$ eV so that $\lambda_p = \frac{3}{4} \cdot 1.24/1.42 = 0.65$ μm, which lies in the red. Note that these calculations are applicable only for absorption mediated by direct interband transitions. ■

6.4 INJECTION ELECTROLUMINESCENCE

Electroluminescence is a phenomenon in which light is emitted by a material that is subjected to an electric field. An important example of electroluminescence is **injection electroluminescence**, which occurs when an electric current is injected into a semiconductor material. The earliest observation of injection electroluminescence,

which dates to 1907, relied on a Schottky-barrier diode fabricated from an indirect-bandgap semiconductor material (p. 169). Modern LEDs make use of forward-biased semiconductor p–n junctions fabricated from direct-bandgap semiconductors, in which electrons from the conduction band recombine with holes from the valence band to generate photons.

Electroluminescence from Schottky-Barrier Diodes

The electroluminescence experiments carried out in the early 1900s made use of point-contact Schottky-barrier diodes and indirect-bandgap semiconductors. We briefly review this early work.

Cat-Whisker, Point-Contact, and Crystal Diodes. The iconic *car-whisker diode* was a rectifying element that served as a demodulator for amplitude-modulated (AM) signals in the early days of radio technology. The "cat whisker" itself was a short length of curved metallic wire (that sometimes incorporated a coiled section serving as a spring) that was adjusted to exert just the right pressure at its contact with a semiconductor crystal so that a rectifying junction would be formed. These devices were also known as *point-contact diodes* or *crystal diodes*. Jagadis Bose first constructed a device of this form in 1894, using a PbS (galena) semiconductor crystal. Shortly thereafter, Greenleaf Pickard improved on the structure and shepherded the device into widespread use in crystal-radio receivers. Diodes such as these were ubiquitous until the early 1920s, when they were replaced by vacuum tubes, which had the distinct merit that they provided amplification.

The Experiments of Round and Losev. Working at the Marconi Company in 1907, the British radio engineer Henry Round (p. 169) was the first to observe injection electroluminescence.[†] His one-off experiment made use of a forward-biased metal-point-contact/SiC rectifying junction. Evidently unaware of Round's work, in 1927 Oleg Losev (p. 169) conducted similar experiments in which he also detected light emission from forward-biased metal-point-contact/semiconductors consisting of SiC and ZnO crytallites.[‡] Losev investigated this phenomenon over a period of many years and established that the light emission was not of thermal origin. Rather, he posited that it was a form of electroluminescence. Losev discovered that some devices emitted light when either forward- or reverse-biased, while other devices emitted light only when reverse-biased.

In modern terminology, the devices used by both Round and Losev were forward-biased Schottky diodes comprising a metallic point contact and a SiC indirect-bandgap semiconductor.

Schottky-Barrier Diodes. The origin of rectification in metal–semiconductor contacts was elucidated in 1938 when Walter Schottky published a theoretical model that incorporated barrier layers at the surfaces where the two materials made contact. Devices of this type, which continue to be used today as large-bandwidth photodetectors, are known as **Schottky-barrier diodes**. A distinct merit of these devices is that they can be used in material systems in which it is not possible to prepare both p-type and n-type forms.

The band diagram of a Schottky-barrier diode is displayed in Fig. 6.4-1 under equilibrium, moderate forward-bias, and strong forward-bias conditions. Under equilibrium

[†] H. J. Round, A Note on Carborundum, *Electrical World*, vol. 83, p. 309, 1907.

[‡] O. V. Losev, Luminous Carborundum Detector and Detection with Crystals (in Russian), *Telegrafiya i telefoniya bez provodov* (*Wireless Telegraphy and Telephony*), vol. 44, pp. 485–494, 1927.

conditions, electrons diffuse from the semiconductor to the metal on contact, bringing the Fermi levels of the two materials into alignment [Fig. 6.4-1(a)]. This results in a region just inside the semiconductor interface that is depleted of free electrons, so that the accompanying fixed positive charges in the semiconductor cause its valence and conduction bands to bend upward at the interface. The discontinuity in the allowed energy states of the two materials gives rise to the Schottky barrier, which blocks the flow of electrons from the metal back into the semiconductor under equilibrium conditions and is responsible for the rectifying nature of the device.

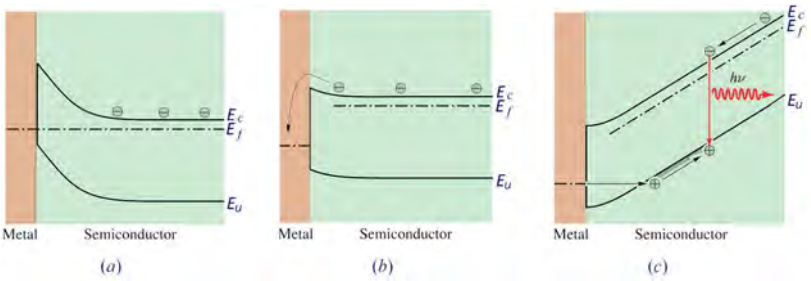


Figure 6.4-1 Band diagram of a Schottky-barrier diode with an n-type semiconductor under: (a) Equilibrium conditions, where the Fermi levels E_f in the two regions align. (b) Moderate forward bias substantially reduces the barrier height. (c) Under strong forward bias conditions, minority carrier injection can occur, allowing the emission of photons with energy comparable to the bandgap energy $E_g = E_c - E_v$.

Forward-Bias Conditions. In their usual mode of operation, Schottky-barrier diodes make use of majority-carrier injection. This is in contrast to p–n junction devices, which function via minority-carrier injection. Under strong forward-bias conditions, however, it is possible to inject minority carriers into the semiconductor region via tunneling through the surface potential barrier, as portrayed in Fig. 6.4-1(c). These minority carriers can then recombine with the n-type majority carriers, emitting recombination radiation in the process. This is the injection electroluminescence that was observed by Round and Losev. The voltage required to achieve minority carrier injection in a Schottky-barrier diode is typically larger than that at which a p–n junction LED operates. Indeed, Round reported operating voltages that ranged between 10 and 110 V.

Reverse-Bias Conditions. Under reverse-bias conditions, minority carriers can be created in Schottky-barrier diodes via avalanche multiplication. In that scenario, impact ionization results in the creation of holes and electrons in the valence and conduction bands, respectively, and light is emitted when the carriers recombine. This explains Losev's observations of electroluminescence under reverse-bias conditions in some devices.

Electroluminescence From p–n Junction Diodes

As described above, the early electroluminescence experiments carried out by Round and Losev made use of Schottky-barrier diodes fabricated using indirect-bandgap semiconductors, principally SiC. In the late 1960s, it became possible to fabricate SiC p–n junction diodes, but these devices turned out to be little better than their progenitor SiC Schottky-barrier diodes in generating electroluminescence light, a consequence of the indirect-bandgap nature of SiC. Modern light-emitting diodes rely on p–n junction diodes fabricated from direct-bandgap semiconductors. These two features, p–n *junctions* and *direct bandgaps*, are responsible for an orders-of-magnitude enhancement of the electroluminescence photon flux and the success of modern-day LED technology.

Semiconductor in Thermal Equilibrium. In principle, electron–hole radiative recombination results in luminescence from a semiconductor material. At room temperature, however, the concentration of thermally excited electrons and holes is so small that the photon flux generated by radiative recombination is essentially undetectable, as illustrated in Example 6.4-1.

EXAMPLE 6.4-1. ***Photon Emission from GaAs in Thermal Equilibrium.*** At room temperature, the intrinsic concentration of electrons and holes in GaAs is $\mathtt{n}_\mathrm{i} \approx 1.8\times10^6$ cm^{-3} (Table 5.4-1). Since the radiative electron–hole recombination coefficient $\mathtt{r}_\mathrm{r} \approx 10^{-10}$ cm^3/s under certain conditions (as specified in Table 5.5-1), the radiative recombination rate $\mathtt{r}_\mathrm{r}\mathtt{np} = \mathtt{r}_\mathrm{r}\mathtt{n}_\mathrm{i}^2 \approx 324$ photons/cm^3-s, as discussed in Sec. 5.5. A 2-μm-thick layer of GaAs therefore produces a photon-flux density $\phi \approx 0.065$ photons/cm^2-s, which is negligible as may be understood by consulting Table 3.4-1 (light emitted from a layer of GaAs thicker than about 2 μm suffers reabsorption). Taking the photon energy $h\nu$ as the bandgap energy for GaAs, $E_g = 1.42$ eV or $1.42e = 2.27 \times 10^{-19}$ J, the emitted intensity turns out to be $I = h\nu\phi \approx 1.5 \times 10^{-20}$ W/cm^2. Of course, the GaAs also emits thermal radiation in accordance with Planck's radiation law (4.7-9).

If thermal equilibrium conditions are maintained, this intensity cannot be appreciably increased (or decreased) by doping the material. In accordance with the law of mass action (5.4-15), the product $\mathtt{np}$ is fixed at $\mathtt{n}_\mathrm{i}^2$ if the material is not too heavily doped so that the radiative recombination rate $\mathtt{r}_\mathrm{r}\mathtt{np} = \mathtt{r}_\mathrm{r}\mathtt{n}_\mathrm{i}^2$ depends on the doping level only through $\mathtt{r}_\mathrm{r}$. An abundance of electrons *and* holes is required for a large recombination rate; in an n-type semiconductor $\mathtt{n}$ is large but $\mathtt{p}$ is small, whereas the converse is true in a p-type semiconductor.

Presence of Carrier Injection. The photon emission rate can be appreciably increased, however, by using external means to increase excess electron–hole pairs in the material. This may be accomplished, for example, by illuminating the material with light, but it is typically achieved by forward biasing a diode, which serves to inject carrier pairs into the junction region. This process is illustrated in Fig. 5.6-3 for a p–n junction diode. The photon emission rate may be calculated from the electron–hole pair injection rate R (pairs/cm^3-s). The photon flux Φ (photons per second), generated within a volume V of the semiconductor material, is directly proportional to the carrier-pair injection rate (Fig. 6.4-2).

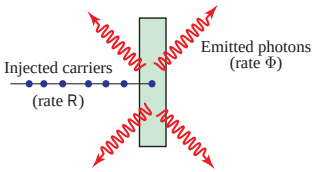


Figure 6.4-2 Spontaneous photon emission resulting from electron–hole radiative recombination, as might occur in a forward-biased Schottky-barrier or p–n junction diode.

Excess Carrier Concentrations. Denoting the equilibrium concentrations of electrons and holes in the absence of excitation (also called **pumping**) as $\mathtt{n}_0$ and $\mathtt{p}_0$, respectively, we use $\mathtt{n} = \mathtt{n}_0 + \Delta\mathtt{n}$ and $\mathtt{p} = \mathtt{p}_0 + \Delta\mathtt{p}$ to represent the steady-state carrier concentrations in the presence of pumping (Sec. 5.5). The excess electron concentration $\Delta\mathtt{n}$ is precisely equal to the excess hole concentration $\Delta\mathtt{p}$ because electrons and holes are produced in pairs. It is assumed that the excess electron–hole pairs recombine at a rate $1/\tau$, where τ is the overall (radiative and nonradiative) electron–hole recombination time. Under steady-state conditions, the generation (pumping) rate must precisely

balance the recombination (decay) rate, so that $R = \Delta n/\tau$. Thus, the steady-state excess-carrier concentration is proportional to the pumping rate, i.e.,

$$\Delta n = R\tau. \tag{6.4-1}$$

For carrier injection rates that are sufficiently low, as explained in Sec. 5.5, we have $\tau \approx 1/r(n_0 + p_0)$ where r is the (radiative and nonradiative) recombination coefficient, so that $R \approx r\Delta n\,(n_0 + p_0)$.

Photon Flux. However, only radiative recombinations generate photons, and the internal quantum efficiency $\eta_{IQE} = r_r/r = \tau/\tau_r$ defined in (5.5-10) and (5.5-12), accounts for the fact that only a fraction of the recombinations are radiative in nature. The injection of RV carrier pairs per second therefore leads to the generation of a photon flux $\Phi = \eta_{IQE}RV$ photons/s, i.e.,

$$\boxed{\Phi = \eta_{IQE}RV = \eta_{IQE}\frac{V\Delta n}{\tau} = \frac{V\Delta n}{\tau_r}.} \tag{6.4-2}$$

The internal photon flux Φ is proportional to the carrier-pair injection rate R and therefore to the steady-state concentration of excess electron–hole pairs Δn.

The IQE plays a crucial role in determining the performance of this electron-to-photon transducer. As schematized in Fig. 6.2-4, photon emission is unlikely in an indirect-bandgap semiconductor. Direct-bandgap semiconductors are used to fabricate LEDs (and LDs) because the IQE is substantially larger than it is for indirect-bandgap semiconductors (e.g., at room temperature $\eta_{IQE} \approx 0.5$ for GaAs, whereas $\eta_{IQE} \approx 10^{-5}$ for Si, as shown in Table 5.5-1). The IQE depends on the doping, temperature, and defect concentration of the material.

EXAMPLE 6.4-2. ***Injection Electroluminescence from GaAs.*** Consider a slab of GaAs for which $\tau = 50$ ns and $\eta_{IQE} = 0.5$ (Table 5.5-1), so that a steady-state excess concentration of injected electron–hole pairs $\Delta n = 10^{17}$ cm^{-3} yields a photon-flux concentration $\eta_{IQE}\Delta n/\tau \approx 10^{24}$ photons/cm^3-s. For photons at the bandgap energy $E_g = 1.42$ eV, the corresponding optical power density is then $\approx 2.3 \times 10^5$ W/cm^3. If the thickness of the slab is 2 μm, the generated optical intensity is ≈ 46 W/cm^2, which is a factor of 10^{21} greater than the value at thermal equilibrium (Example 6.4-1). If the area of the slab is 200 μm $\times$ 10 μm, the emitted optical power under these conditions is calculated to be ≈ 0.9 mW, which is substantial.

Indirect-Bandgap *p–n* Junction Diodes. The electroluminescence light emitted by SiC *p–n* junction diodes was very weak because the IQEs of these devices are minuscule because of their indirect-bandgap nature (Fig. 6.2-4). Nevertheless, SiC diodes continued to be manufactured until the early 1990s because they were the only semiconductor source of violet light available. (The bandgap of the 6H-SiC polytype is $E_g \approx 3.05$ eV, corresponding to $\lambda_0 \approx 407$ nm, as shown in Fig. 5.3-3(*b*).) Only when direct-bandgap, high internal quantum efficiency III–nitride LEDs began to be manufactured in the mid-1990s, did SiC LEDs lose any of the allure they might have had. Also abandoned were blue-emitting LEDs fabricated from direct-bandgap II–VI ZnSe/ZnS, as a consequence of their limited lifetimes.

Direct-Bandgap *p–n* Junction Diodes.

Electroluminescence Spectral Density. We now proceed to elaborate on the spectral density of injection electroluminescence from direct-bandgap semiconductors, which is established with the help of the interband-transition theory presented in Secs. 6.2 and 6.3. As discussed in those sections, spontaneous injection electroluminescence is a result of the recombination of electron-hole pairs, as illustrated in Fig. 6.4-3.

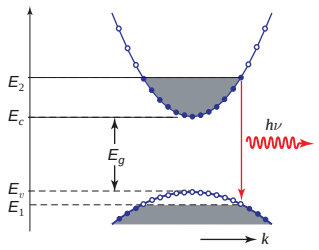


Figure 6.4-3 As explained in Secs. 6.2 and 6.3, the spontaneous emission of a photon results from the recombination of an electron of energy E_2 with a hole of energy $E_1 = E_2 - h\nu$. The transition is represented by a vertical line because the momentum carried away by the photon, $h\nu/c$, is negligible on the scale of the figure.

The rate of spontaneous emission $r_{\rm sp}(\nu)$ (photons/s-Hz-cm^3), as provided in (6.3-8), is given by

$$r_{\rm sp}(\nu) = \frac{1}{\tau_{\rm r}} \varrho(\nu) f_e(\nu), \tag{6.4-3}$$

Spontaneous Emission Rate

where $\tau_{\rm r}$ is the radiative electron–hole recombination lifetime. The optical joint density of states, obtained from (6.2-2)–(6.2-8), is written as

$$\varrho(\nu) = \frac{(2m_r)^{3/2}}{\pi\hbar^2} \sqrt{h\nu - E_g}\,, \qquad h\nu \geqslant E_g, \tag{6.4-4}$$

Optical Joint Density of States

while the emission condition as set forth in (6.3-1) is

$$f_e(\nu) = f_c(E_2)\,[1 - f_v(E_1)]. \tag{6.4-5}$$

Emission Condition

The Fermi functions that appear in the emission condition (6.4-5), $f_c(E_2)$ and $f_v(E_1)$, apply to the conduction and valence bands, respectively, under conditions of thermal quasi-equilibrium.

The semiconductor parameters E_g, $\tau_{\rm r}$, m_v, and m_c, along with the temperature T, determine the spectral distribution $r_{\rm sp}(\nu)$, given the quasi-Fermi levels E_{fc} and E_{fv}. These in turn are determined from the concentrations of electrons and holes provided in (5.4-7) and (5.4-8),

$$\int_{E_c}^{\infty} \varrho_c(E) f_c(E)\, dE = \mathtt{n} = \mathtt{n}_0 + \Delta\mathtt{n}, \tag{6.4-6}$$

$$\int_{-\infty}^{E_v} \varrho_v(E)[1 - f_v(E)]\, dE = \mathtt{p} = \mathtt{p}_0 + \Delta\mathtt{n}, \tag{6.4-7}$$

where n_0 and p_0 are the concentrations of electrons and holes in thermal equilibrium (in the absence of injection), and $\Delta\mathrm{n} = R\tau$ is the steady-state injected-carrier concentration, as specified in (6.4-1). The densities of states near the conduction- and valence-band edges are provided in (5.4-2) and (5.4-3), respectively.

Electroluminescence Photon Flux. The spontaneous photon flux is obtained by integrating the spectral density $r_{\mathrm{sp}}(\nu)$ over all frequencies and volume. With the help of the integral $\int_0^\infty u^{1/2}e^{-au}du = (\sqrt{\pi}/2)a^{-3/2}$, we then obtain

$$\Phi = V\int_0^\infty r_{\mathrm{sp}}(\nu)\,d\nu = \frac{V(m_r)^{3/2}}{\sqrt{2}\,\pi^{3/2}\,\hbar^3\,\tau_{\mathrm{r}}}\,(kT)^{3/2}\,\exp\left(\frac{E_{fc}-E_{fv}-E_g}{kT}\right). \tag{6.4-8}$$

Increasing the pumping level R causes $\Delta\mathrm{n}$ to increase, which, in accordance with (6.4-9), moves E_{fc} toward (or further into) the conduction band, and E_{fv} toward (or further into) the valence band. This results in an increase in the probability $f_c(E_2)$ of finding the conduction-band state of energy E_2 filled with an electron, and the probability $1 - f_v(E_1)$ of finding the valence-band state of energy E_1 empty (filled with a hole). The net result is that the emission-condition probability $f_e(\nu) = f_c(E_2)[1 - f_v(E_1)]$ increases with R, thereby enhancing the spontaneous emission rate given in (6.4-10) and the spontaneous photon flux Φ given in (6.4-8).

□ **Derivation of the Quasi-Fermi Levels of a Pumped Semiconductor.**

(a) Under ideal conditions at $T = 0$ K, when there is no thermal electron–hole pair generation [Fig. 6.4-4(a)], the quasi-Fermi levels are related to the concentrations of injected electron–hole pairs $\Delta\mathrm{n}$ via

$$E_{fc} = E_c + (3\pi^2)^{2/3}\frac{\hbar^2}{2m_c}(\Delta\mathrm{n})^{2/3} \tag{6.4-9a}$$

$$E_{fv} = E_v - (3\pi^2)^{2/3}\frac{\hbar^2}{2m_v}(\Delta\mathrm{n})^{2/3}. \tag{6.4-9b}$$

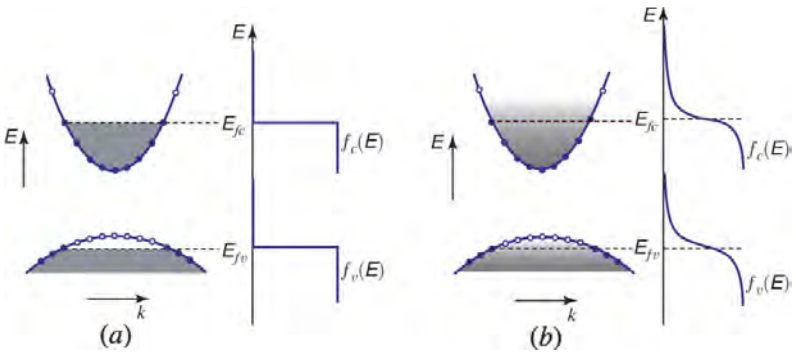


Figure 6.4-4 Energy bands, Fermi functions, and quasi-Fermi levels for a semiconductor in thermal quasi-equilibrium: (a) under ideal conditions when $T = 0$ K; and (b) at a temperature $T > 0$ K.

The calculations that lead to (6.4-9) follow the same path as those used to arrive at (5.4-16); indeed (6.4-9) is identical to (5.4-16) except that $\mathrm{n}^{2/3}$ is replaced by $(\Delta\mathrm{n})^{2/3}$. At $T = 0$ K, the Fermi function $f_c(E) = 1$ obeys $E < E_{fc}$ and 0 otherwise. This result may be used in

conjunction with (5.4-2) and (5.4-7) to evaluate the integral set forth in (5.4-8). Making use of the substitution $x = (E - E_c)$ then provides

$$\Delta \mathrm{n} = \int_{E_c}^{E_{fc}} A(E - E_c)^{1/2} dE = \tfrac{2}{3} A(E_{fc} - E_c)^{3/2}, \tag{6.4-9c}$$

where $A = (2m_c)^{3/2}/2\pi^2\hbar^3$ is a constant. We thus arrive at $E_{fc} - E_c = (3/2A)^{2/3}\,\Delta \mathrm{n}^{2/3}$, from which (6.4-9a) follows, and (6.4-9b) follows suit. Finally, we subtract (6.4-9b) from (6.4-9a), and make use of the reduced mass $1/m_r = 1/m_c + 1/m_v$ defined in (6.2-4), to arrive at

$$E_{fc} - E_{fv} = E_g + (3\pi^2)^{2/3} \frac{\hbar^2}{2m_r} (\Delta \mathrm{n})^{2/3}, \tag{6.4-9d}$$

where $\Delta \mathrm{n} \gg \mathrm{n}_0, \mathrm{p}_0$. Under these conditions, all injected electrons $\Delta \mathrm{n}$ occupy the lowest allowed energy levels in the conduction band, and all injected holes $\Delta \mathrm{p}$ occupy the highest allowed levels in the valence band.

(b) It is useful to sketch the functions $f_e(\nu)$ and $r_{\mathrm{sp}}(\nu)$ for different values of $\Delta \mathrm{n}$. Moreover, the effect of temperature on the Fermi functions illustrated in Fig. 6.4-4(b) can be used to establish the effect of increasing temperature on $r_{\mathrm{sp}}(\nu)$, as illustrated in Fig. 6.4-5.
From (6.2-5), (6.2-6), and (6.3-1), we have $f_e(\nu) = f_c(E_2)[1 - f_v(E_1)]$, with $E_2 = E_c + (m_r/m_c)(h\nu - E_g)$ and $E_1 = E_2 - h\nu$. At $T = 0$ K, the Fermi function $f_c(E_2)$ is unity as long as $E_2 < E_{fc}$ and is 0 otherwise. Similarly, the Fermi function $f_v(E_1)$ is unity for $E_1 < E_{fv}$ and is 0 otherwise. For $h\nu > E_g$, as $h\nu$ increases, we see that E_2 increases and E_1 decreases. But as long as these two values lie below E_{fc} and above E_{fv}, respectively, $f_c(E_2) = 1$ and $1 - f_v(E_1) = 1$, so that $f_e(\nu) = 1$. When $h\nu$ exceeds the value $E_{fc} - E_{fv}$, we see that E_2 exceeds E_{fc} and E_1 lies below E_{fv}, so that $f_c(E_2) = 0$ and $1 - f_v(E_1) = 0$, indicating that $f_e(\nu) = 0$. The function $f_e(\nu)$ is therefore a rectangular function with value 1 for $E_g < h\nu < E_{fc} - E_{fv}$, and value 0 otherwise, as displayed in Fig. 6.4-5(a). According to (6.4-3), the rate of spontaneous emission r_{sp} is proportional to $\varrho(\nu) f_e(\nu)$, where $\varrho(\nu) \propto (h\nu - E_g)^{1/2}$. Therefore, the dependence of r_{sp} on ν is as illustrated in Fig. 6.4-5(a) for $T = 0$ K. The effect of increasing the temperature ($T > 0$ K) is to smooth the Fermi function so that the functions $f_e(\nu)$ and $r_{\mathrm{sp}}(\nu)$ take the forms displayed in Fig. 6.4-5(b).

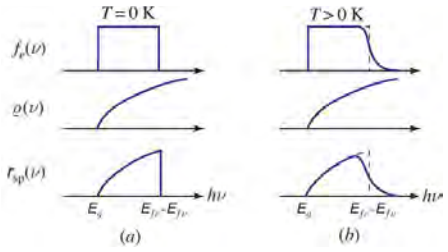


Figure 6.4-5 Effect of temperature on: 1) the probability that the emission condition $f_e(\nu)$ is satisfied; 2) the optical joint density of states $\varrho(\nu)$; and 3) the rate of photon emission $r_{\mathrm{sp}}(\nu)$. (a) $T = 0$ K. (b) $T > 0$ K.

■

Injection electroluminescence is generated by direct-bandgap semiconductors in thermal quasi-equilibrium, whereas thermal light is generated by all systems in thermal equilibrium.

EXAMPLE 6.4-3. ***Spectral Density Under Weak Injection.*** This example provides analytical expressions for the spontaneous spectral density under weak injection. Some of the more salient features of this spectral density, along with representative numerical estimates, are examined in

Example 6.4-4. In accordance with (6.4-3)–(6.4-5), the spectral density, optical joint density of states, and emission condition are given by $r_{sp}(\nu) = \varrho(\nu) f_e(\nu)/\tau_r$, $\varrho(\nu) = [(2m_r)^{3/2}/\pi\hbar^2](h\nu - E_g)^{1/2}$, and $f_e(\nu) = f_c(E_2)[1 - f_v(E_1)]$, respectively. For sufficiently weak injection, such that the Fermi levels lie within the bandgap and away from the band edges by several kT, so that $E_c - E_{fc} \gg kT$ and $E_{fv} - E_v \gg kT$, the Fermi functions may be approximated by their exponential tails, in which case $f_c(E_2) \approx \exp[-(E_2 - E_{fc})/kT]$ and $1 - f_v(E_1) \approx \exp[-(E_{fv} - E_1)/kT]$. The probability that the emission condition is satisfied can then be written as $f_e(\nu) \approx \exp[(E_{fc} - E_{fv})/kT] \cdot \exp[-(E_2 - E_1)/kT] = \exp[(E_{fc} - E_{fv})/kT] \cdot \exp(-h\nu/kT)$. Substituting this into the expression for $r_{sp}(\nu)$ provided in (6.4-3) then yields

$$r_{sp}(\nu) = D\sqrt{h\nu - E_g}\,\exp\left(-\frac{h\nu - E_g}{kT}\right), \quad h\nu \geqslant E_g, \tag{6.4-10a}$$

where

$$D = \frac{(2m_r)^{3/2}}{\pi\hbar^2\tau_r}\,\exp\left(\frac{E_{fc} - E_{fv} - E_g}{kT}\right) \tag{6.4-10b}$$

is an exponentially increasing function of the separation between the quasi-Fermi levels $E_{fc} - E_{fv}$.

Equation (6.4-10), which is displayed in Fig. 6.4-6, has precisely the same shape as the thermal-equilibrium spectral density shown in Fig. 6.3-1. However, its magnitude is larger by the factor $D/D_0 = \exp[(E_{fc} - E_{fv})/kT]$, which can be very large in the presence of injection, leading to a greatly enhanced magnitude for the injection electroluminescence. In thermal equilibrium, when $E_{fc} = E_{fv}$, we recover (6.3-12) and (6.3-13).

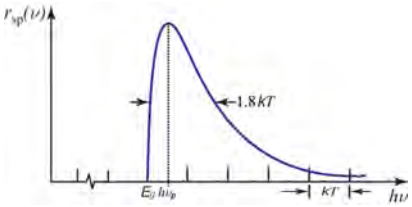


Figure 6.4-6 Spectral density of the direct interband injection-electroluminescence rate $r_{sp}(\nu)$ (photons/s-Hz-cm^3), versus the photon energy $h\nu$, under conditions of weak injection, from (6.4-10). The peak photon energy, $h\nu_p = E_g + kT/2$, is indicated, as is the energy width $h\Delta\nu \approx 1.8\,kT$ (see Example 6.4-4).

EXAMPLE 6.4-4. ***Features of the Spectral Density Under Weak Injection.***

(a) *Peak photon energy $h\nu_p$*: With the help of the substitution $u = (h\nu - E_g)/kT$, (6.4-10) may be written in the form $r_{sp}(\nu) = D(kT)^{1/2}u^{1/2}\exp(-u)$. The function $u^{1/2}\exp(-u)$ has its peak value when its derivative with respect to u vanishes, i.e., when $-u^{1/2}\exp(-u) + \frac{1}{2}u^{-1/2}\exp(-u) = 0$, from which we obtain $u = \frac{1}{2}$, i.e., $(h\nu - E_g)/kT = \frac{1}{2}$. Hence, the spectral density of the emitted light described by (6.4-10) attains its peak value at a photon energy $h\nu_p$ given by

$$h\nu_p = E_g + \tfrac{1}{2}kT\,. \tag{6.4-11}$$

At $T = 300$ K, we obtain $kT/2 = 0.013$ eV. Since $E_g = h\nu_g$ ranges from about 1.5 to 3 eV over the visible region, we have $kT/2 \ll E_g$, indicating that the electroluminescence photon energy exceeds the bandgap energy by only a slight amount.

(b) *Peak wavelength λ_p*: The peak wavelength is calculated by writing (6.4-10) as a function of the free-space wavelength λ_0 and setting $dr_{sp}(\lambda_0)/d\lambda_0 = 0$. In this case, it turns out that the peak wavelength can also be directly determined by inserting (5.1-1) together with the formula $\nu_p = c_0/\lambda_p$ into (6.4-11), which leads to $1/\lambda_p = 1/\lambda_g + kT/2hc_0$ and thence to $\lambda_p = \lambda_g/(1 + \lambda_g kT/2hc_0)$. A Taylor-series expansion then yields an approximate expression for the peak wavelength, which lies just a bit below the bandgap wavelength λ_g:

$$\lambda_p \approx \lambda_g - \lambda_g^2 kT/2hc_0\,. \tag{6.4-12}$$

The visible band is convered by a bandgap wavelength λ_g that extends from about 800 nm to 400 nm. At $T = 300$ K, we have $kT = 0.026$ eV so that $\lambda_g^2 kT/2hc_0 \approx 6.67$ nm at 800 nm and ≈ 1.67 nm at 400 nm. Clearly, $\lambda_g^2 kT/2hc_0 \ll \lambda_g$ over the full visual band, which also justifies the Taylor-series expansion above.

(c) *Photon-energy width $h\Delta\nu$:* The peak of the function $u^{1/2}e^{-u}$ occurs at $u = \frac{1}{2}$, where the function has the value $(\frac{1}{2})^{1/2}e^{-1/2}$. The function reaches half its peak value where $u^{1/2}e^{-u} = \frac{1}{2} \times (\frac{1}{2})^{1/2}e^{-1/2}$, i.e., where $u^{1/2}e^{-u} = (\frac{1}{2})^{3/2}e^{-1/2}$. Squaring both sides of this equation leads to $ue^{-2u} = (\frac{1}{2})^3 e^{-1} = 0.046$. Computation yields the roots of this equation, which are $u_1 \approx 0.051$ and $u_2 \approx 1.84$. The difference between these values, $u_2 - u_1 = 1.79 \approx 1.8$, corresponds to $[(h\nu_2 - E_g)/kT - (h\nu_1 - E_g)/kT] \approx 1.8$ so that $h(\nu_2 - \nu_1) \approx 1.8\,kT$. The FWHM photon-energy width is therefore

$$h\Delta\nu \approx 1.8\,kT. \tag{6.4-13}$$

The photon energy width $h\Delta\nu$ is independent of the photon energy $h\nu$.

(d) *Wavelength spectral width $\Delta\lambda$:* The magnitude of the wavelength spectral width $\Delta\lambda$ is determined from the frequency spectral width $\Delta\nu \approx 1.8\,kT/h$ set forth in (6.4-13) above. Since $\nu = c_0/\lambda_0$, we have $\Delta\nu = -(c_0/\lambda_0^2)\Delta\lambda$ which, with the help of $\lambda_g \equiv c_0/\nu_g$, yields $|\Delta\lambda| \approx (\lambda_g^2/c_0)\Delta\nu = 1.8\,\lambda_g^2 kT/hc_0$. Expressing $\Delta\lambda$ and λ_g in μm, and kT in eV, the foregoing equation is written as $\Delta\lambda\,(\mu\text{m}) \times 10^{-6} \approx 1.8\,[\lambda_g^2\,(\mu\text{m}^2) \times 10^{-12}/hc_0] \cdot [kT\,(\text{eV}) \cdot e]$ or $\Delta\lambda\,(\mu\text{m}) \approx [1.8/(10^6 \times hc/e)] \cdot [\lambda_g^2\,(\mu\text{m}^2)] \cdot [kT\,(\text{eV})]$. Finally, since $(10^6 \times hc_0/e) = 1.24$ and $1.8/1.24 \approx 1.45$, we obtain

$$\boxed{\Delta\lambda \approx 1.45\,\lambda_g^2\,kT.} \tag{6.4-14}$$
($\Delta\lambda$ and λ_g in μm; kT in eV)

In contrast to the frequency spectral width $\Delta\nu$, which is independent of ν_g, the wavelength spectral width $\Delta\lambda$ increases quadratically with λ_g.

(e) *Representative values of $h\nu_p - E_g$, $h\Delta\nu$, $\lambda_g - \lambda_p$, and $\Delta\lambda$ over the visible spectrum:* Values are tabulated below for selected features of the spectral density for p–n junction injection-electroluminescence, under weak injection and at $T = 300$ K, as provided in (6.4-10). The following parameters are examined: 1) the deviation of the peak electroluminescence photon energy from the bandgap energy ($h\nu_p - E_g$); 2) the photon-energy width ($h\Delta\nu$); 3) the deviation of the peak emission wavelength from the bandgap wavelength ($\lambda_g - \lambda_p$); and 4) the wavelength spectral width ($\Delta\lambda$). These parameters are calculated at $E_g = 1.55$ eV ($\lambda_g = 800$ nm) and at $E_g = 3.10$ eV ($\lambda_g = 400$ nm), which lie at the red and blue ends of the visible spectrum, respectively.

Table 6.4-1 Values for $h\nu_p - E_g$, $h\Delta\nu$, $\lambda_g - \lambda_p$, and $\Delta\lambda$ at $\lambda_g = 800$ nm and 400 nm.

BANDGAP ENERGY E_g / BANDGAP WAVELENGTH λ_g	$h\nu_p - E_g = \frac{1}{2}kT$	$h\Delta\nu = 1.8\,kT$	$\lambda_g - \lambda_p = \frac{1}{2}\lambda_g^2 kT/hc_0$	$\Delta\lambda = 1.8\,\lambda_g^2 kT/hc_0$
1.55 eV / 800 nm	0.013 eV	0.047 eV	6.67 nm	24 nm
3.10 eV / 400 nm	0.013 eV	0.047 eV	1.67 nm	6 nm

The width-to-deviation ratio is the same for both the photon energy and the wavelength, i.e., $h\Delta\nu/(h\nu_p - E_g) = \Delta\lambda/(\lambda_g - \lambda_p) = 3.6$.

(f) *Strong Injection:* The formulas presented above are based on (6.4-10), which was derived under the assumption of weak carrier injection, i.e., the Fermi levels were assumed to lie within the bandgap, which allowed the Fermi functions to be approximated by their exponential tails outside the bandgap region. A number of other potential sources of broadening were also ignored in formulating these results. These include thermal and static disorder in the crystal associated with phonon-assisted effects, and randomness in the doping and chemical composition of the materials used in fabrication (alloy broadening). It is therefore judicious to consider the values presented in Table 6.4-1 to be lower bounds. Nevertheless, it can be

inferred from Fig. 6.4-5, which is a sketch of $r_{sp}(\nu)$ for strong pumping, that (6.4-11) and (6.4-13) become

$$h\nu_p - E_g > \tfrac{1}{2}kT \tag{6.4-15}$$

$$h\Delta\nu > 1.8\,kT, \tag{6.4-16}$$

and that appropriate values of $h\nu_p - E_g$ and $\Delta\nu$ can be determined for an arbitrary level of pumping. The conversion of these frequency-based parameters into their wavelength-based counterparts depends solely on the reciprocal relation between wavelength and frequency, and is therefore identical to the route used to derive the results presented in (b) and (d) above from those provided in (a) and (c), respectively. We conclude that the proportionality of $\lambda_g - \lambda_p$ and $\Delta\lambda$ to λ_g^2 displayed in (6.4-12) and (6.4-14), respectively, remains intact for all pumping levels. Furthermore, since (6.4-12) reveals that $\lambda_p \approx \lambda_g$, where λ_p is the peak emission wavelength, the proportionality in (6.4-14) can be equivalently written as

$$\Delta\lambda \propto \lambda_p^2\,. \tag{6.4-17}$$

The peak wavelength of the electroluminescence emitted by a direct-bandgap semiconductor is determined principally by its bandgap wavelength, whereas the peak wavelength of the light emitted by a thermal source is established solely by its thermodynamic temperature via Wien's law.

6.5 QUANTUM WELLS AND MULTIQUANTUM WELLS

Multiquantum-well and superlattice structures were considered in Sec. 5.7. The photon interactions in these structures bear a considerable resemblance to those for bulk semiconductors (Sec. 6.1). Several mechanisms play important roles in absorption and emission in quantum-confined structures:

- Interband (band-to-band) transitions
- Excitonic transitions
- Intersubband transitions
- Miniband transitions

These are illustrated in Fig. 6.5-1 and discussed below.

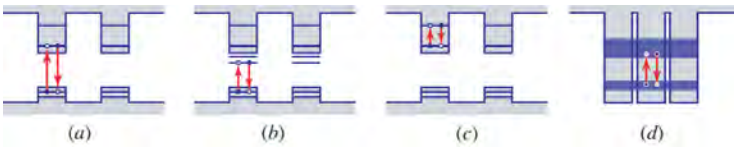


Figure 6.5-1 Photon absorption and emission in multiquantum-well structures. (a) Interband transitions. (b) Excitonic transitions. (c) Intersubband transitions. (d) Miniband transitions in a superlattice structure.

Interband Transitions. Interband emission and absorption takes place between states in the valence and conduction bands [Fig. 6.5-1(a)], much as in bulk semiconductors. Because of quantum confinement, however, the optical joint density of states (6.2-8) must be replaced by the staircase joint density of states

$$\varrho(\nu) = \begin{cases} \dfrac{hm_r}{m_c}\dfrac{m_c}{\pi\hbar^2 l} = \dfrac{2m_r}{\hbar l}, & h\nu > E_g + E_q + E'_q \\ 0, & \text{otherwise,} \end{cases} \tag{6.5-1}$$

which derives from (5.7-5). Interband transitions are responsible for the operation of MQW light-emitting diodes (Fig. 7.3-4), superluminescent diodes, and laser diodes.

Excitonic Transitions. The one-dimensional carrier confinement associated with MQW structures results in an increase in the exciton binding energy. This leads to strong **excitonic transitions**, even at $T = 300$ K, as schematized in Fig. 6.5-1(b). Excitonic transitions play an important role in many quantum-confined photonic devices.

Intersubband Transitions. Transitions that take place between energy levels within a single band of a MQW structure [Fig. 6.5-1(c)] are known as **intersubband transitions**. Devices that operate on the basis of these intraband transitions include the quantum-well quantum cascade laser and the quantum-well infrared photodetector. In the latter device, which offers large bandwidth, the absorption of a photon causes a transition from a bound energy level to the continuum.

Miniband Transitions. In superlattices, the discrete MQW energy levels broaden into minibands that are separated by minigaps. Such **miniband transitions** [Fig. 6.5-1(d)] are important in the operation of superlattice quantum cascade lasers.

6.6 QUANTUM-DOT SINGLE-PHOTON EMITTERS

As discussed in Sec. 5.8, aggregations of quantum dots can serve as photonic devices that range from light-emitting diodes to backlights. Still, individual quantum dots, when embedded in photonic structures (e.g., microcavities, semiconductor heterostructures, and 2D materials) can be used as optically or electrically excited **single-photon emitters (SPEs)**. Because an individual quantum dot can emit only one photon at a time, the photons are separated from each other in time and therefore exhibit natural antibunching and sub-Poisson behavior. Quantum-dot single-photon emitters have been fabricated from II–VI, III–V, and group-IV semiconductors, as well as from organic and perovskite materials.

Efficient, on-demand sources of such pure, highly indistinguishable, single-photon streams are useful for enabling scalable quantum information processing, communications, computing, and cryptography. Although SPEs can be created using other approaches (e.g., diamond defect centers, single-walled carbon nanotubes, and defects in 2D materials), the simplicity and easy availability of quantum dots is appealing. QDs can also be used to generate entangled photons, another form of nonclassical light.

Two examples of quantum-dot single-photon emitters are provided below:

EXAMPLE 6.6-1. ***Quantum-Dot/Micropillar Single-Photon Emitter.***

Indistinguishable photons of high purity (in this case signifying that one and only one photon is emitted at a time) can be generated by making use of resonance fluorescence. One implementation consists of a single InAs/GaAs self-assembled quantum dot embedded in a 2.5-μm-diameter, cryogenically cooled micropillar microcavity, such as that pictured at right. Excitation is provided by 25-nW, 3-ps-duration optical pulses of wavelength $\lambda_0 = 897$ nm, which is resonant with the microcavity, delivered at a repetition rate of 81 MHz. This device offers a substantial Purcell spontaneous-emission enhancement factor ($F_P > 5$) by virtue of its small cavity volume and high quality factor ($Q > 6000$). A carefully coupled single-mode optical fiber can extract $> 3.5 \times 10^6$ single photons/s with an extraction efficiency $\eta_{XTE} \approx 0.65$, yielding an overall system efficiency $\approx 4.5\%$.

EXAMPLE 6.6-2. ***Silicon-Photonics Quantum-Dot Emitter.*** Carrier confinement in a quantum dot leads to a reduction in positional uncertainty Δx. In accordance with the Heisenberg position–momentum uncertainty relation $\Delta x\, \Delta p \geqslant \frac{\hbar}{2}$ set forth in (A.2-9) of Appendix A, a decrease in Δx is accompanied by a concomitant increase in the momentum uncertainty Δp. This obviates the need for phonon momentum to participate in radiative recombination in an indirect-bandgap quantum dot, greatly increasing efficiency. This behavior is analogous to the co-doping of GaP with impurities such as N, which take up residence at sharply localized positions in the crystal, enabling GaP:N LEDs to emit light (Sec. 7.3). In brief, the small size of an indirect-bandgap Si quantum dot significantly enhances radiative recombination via interband transitions. Light emission from porous silicon is also possible by virtue of an enhancement of the radiative rate facilitated by induced surface-localized excitons resulting from surface passivation.

6.7 REFRACTIVE INDEX

The ability to control the refractive index of a semiconductor is important in the design of many photonic devices, particularly those that make use of optical waveguides, laser diodes, and integrated photonics. Semiconductor materials are dispersive, so that the refractive index is dependent on the wavelength. Indeed, the refractive index is related to the absorption coefficient $\alpha(\nu)$ inasmuch as the real and imaginary parts of the susceptibility must satisfy the Kramers–Kronig relations. The group index and refractive index for GaAs, calculated from the Sellmeier equation, are displayed in Fig. 6.7-1. The refractive index depends on temperature and doping level.

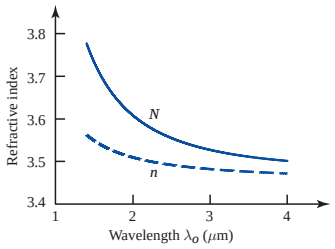


Figure 6.7-1 Refractive index n and group index N for GaAs as a function of the free-space wavelength λ_0. The results are determined from the Sellmeier equation.

The refractive indices of selected elemental and binary bulk semiconductors, under specific conditions and near the bandgap wavelength, are provided in Table 6.7-1. The refractive indices of ternary and quaternary semiconductors can be approximated via linear interpolation between the refractive indices of their components.

Table 6.7-1 Refractive indices n of selected semiconductor materials.[a]

Material	Refractive Index
ELEMENTAL SEMICONDUCTORS	
Ge	4.0
Si	3.5
III–V BINARY SEMICONDUCTORS	
AlN	2.2
AlP	3.0
AlAs	3.2
AlSb	3.8
GaN	2.5
GaP	3.3
GaAs	3.6
GaSb	4.0
InN	3.0
InP	3.5
InAs	3.8
InSb	4.2

[a]Results reported are for photon energies near the bandgap energy of the material ($h\nu \approx E_g$) and at $T = 300$ K.

BIBLIOGRAPHY

Semiconductor Photonics

M. Nisoli, *Semiconductor Laser Photonics*, Cambridge University Press, 2nd ed. 2023.

B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, 3rd ed. 2019, Chap. 17.

B. K. Ridley, *Quantum Processes in Semiconductors*, Oxford University Press, 5th ed. 2013.

W. Barford, *Electronic and Optical Properties of Conjugated Polymers*, Oxford University Press, 2nd ed. 2013.

C. F. Klingshirn, *Semiconductor Optics*, Springer, 4th ed. 2012.

M. Kira and S. W. Koch, *Semiconductor Quantum Optics*, Cambridge University Press, 2012.

S. O. Kasap, *Optoelectronics and Photonics: Principles and Practices*, Pearson, 2nd ed. 2012.

S. Adachi, *The Handbook on Optical Constants of Semiconductors: In Tables and Figures*, World Scientific, 2012.

T. Yoshimura, *Thin-Film Organic Photonics: Molecular Layer Deposition and Applications*, CRC Press/Taylor & Francis, 2011.

M. Fox, *Optical Properties of Solids*, Oxford University Press, paperback 2nd ed. 2010.

S. L. Chuang, *Physics of Photonic Devices*, Wiley, 2nd ed. 2009.

H. Haug and S. W. Koch, *Quantum Theory of the Optical and Electronic Properties of Semiconductors*, World Scientific, 5th ed. 2009.

H. Morkoç, *Handbook of Nitride Semiconductors and Devices*, Volume I: *Materials Properties, Physics and Growth*; Volume II: *Electronic and Optical Processes in Nitrides*, Wiley–VCH, 2008.

T. Meier, P. Thomas, and S. W. Koch, *Coherent Semiconductor Optics: From Basic Concepts to Nanostructure Applications*, Springer, 2007.

A. Moliton, *Optoelectronics of Molecules and Polymers*, Springer, 2006, paperback ed. 2011.

S. O. Kasap and P. Capper, eds., *Springer Handbook of Electronic and Photonic Materials*, Springer, 2006.

P. K. Basu, *Theory of Optical Processes in Semiconductors: Bulk and Microstructures*, Oxford University Press, paperback ed. 2003.

P. T. Landsberg, *Recombination in Semiconductors*, Cambridge University Press, 1991, paperback ed. 2003.

J. I. Pankove, *Optical Processes in Semiconductors*, Prentice Hall, 1971; Dover, paperback ed. 2010.

Single-Photon Emitters

Y. Karli, D. A. Vajner, F. Kappe, P. C. A. Hagen, L. M. Hansen, R. Schwarz, T. K. Bracht, C. Schimpf, S. F. Covre da Silva, P. Walther, A. Rastelli, V. M. Axt, J. C. Loredo, V. Remesh, T. Heindel, D. E. Reiter, and G. Weihs, Controlling the Photon Number Coherence of Solid-State Quantum Light Sources for Quantum Cryptography, arXiv:2305.20017v1, 2023.

C. Couteau, S. Barz, T. Durt, T. Gerrits, J. Huwer, R. Prevedel, J. Rarity, A. Shields, and G. Weihs, Applications of Single Photons in Quantum Metrology, Biology and the Foundations of Quantum Physics, *Nature Reviews Physics*, DOI:10.1038/s42254-023-00589-w, 2023.

P. Lodahl, A. Ludwig, and R. J. Warburton, A Deterministic Source of Single Photons, *Physics Today*, vol. 75, no. 3, pp. 44–50, 2022.

P. Michler, ed., *Quantum Dots for Quantum Information Technologies*, Springer, 2017.

X. Lin, X. Dai, C. Pu, Y. Deng, Y. Niu, L. Tong, W. Fang, Y. Jin, and X. Peng, Electrically-Driven Single-Photon Sources Based on Colloidal Quantum Dots with Near-Optimal Antibunching at Room Temperature, *Nature Communications*, vol. 8, 1132, 2017.

X. He, N. F. Hartmann, X. Ma, Y. Kim, R. Ihly, J. L. Blackburn, W. Gao, J. Kono, Y. Yomogida, A. Hirano, T. Tanaka, H. Kataura, H. Htoon, and S. K. Doorn, Tunable Room-Temperature Single-Photon Emission at Telecom Wavelengths from sp^3 Defects in Carbon Nanotubes, *Nature Photonics*, vol. 11, pp. 577–582, 2017.

C. Palacios-Berraquero, D. M. Kara, A. R.-P. Montblanch, M. Barbone, P. Latawiec, D. Yoon, A. K. Ott, M. Loncar, A. C. Ferrari, and M. Atatüre, Large-Scale Quantum-Emitter Arrays in Atomically Thin Semiconductors, *Nature Communications*, vol. 8, 15093, 2017.

I. Aharonovich and M. Toth, Quantum Emitters in Two Dimensions, *Science*, vol. 358, pp. 170–171, 2017.

I. Aharonovich, D. Englund, and M. Toth, Solid-State Single-Photon Emitters, *Nature Photonics*, vol. 10, pp. 631–641, 2016.

X. Ding, Y. He, Z.-C. Duan, N. Gregersen, M.-C. Chen, S. Unsleber, S. Maier, C. Schneider, M. Kamp, S. Höfling, C.-Y. Lu, and J.-W. Pan, On-Demand Single Photons with High Extraction Efficiency and Near-Unity Indistinguishability from a Resonantly Driven Quantum Dot in a Micropillar, *Physical Review Letters*, vol. 116, 020401, 2016.

N. Somaschi, V. Giesz, L. De Santis, J. C. Loredo, M. P. Almeida, G. Hornecker, S. L. Portalupi, T. Grange, C. Antón, J. Demory, C. Gómez, I. Sagnes, N. D. Lanzillotti-Kimura, A. Lemaítre, A. Auffeves, A. G. White, L. Lanco, and P. Senellart, Near-Optimal Single-Photon Sources in the Solid State, *Nature Photonics*, vol. 10, pp. 340–345, 2016.

P. Lodahl, S. Mahmoodian, and S. Stobbe, Interfacing Single Photons and Single Quantum Dots with Photonic Nanostructures, *Reviews of Modern Physics*, vol. 87, pp. 347–400, 2015.

M. G. Raymer and K. Srinivasan, Manipulating the Color and Shape of Single Photons, *Physics Today*, vol. 65, no. 11, pp. 32–37, 2012.

Injection Electroluminescence: Historical Accounts and Seminal Publications

E. F. Schubert, *Light-Emitting Diodes*, Google Books, 4th ed. 2023.

N. Zheludev, The Life and Times of the LED — a 100-Year History, *Nature Photonics*, vol. 1, pp. 189–192, 2007.

E. E. Loebner, Subhistories of the Light Emitting Diode, *IEEE Transactions on Electron Devices*, vol. ED-23, pp. 675–699, 1976.

H. Gooch, *Injection Electroluminescent Devices*, Wiley, 1973.

H. K. Henisch, Electroluminescence, *Reports on Progress in Physics*, vol. 27, pp. 369–405, 1964.

H. K. Henisch, *Electroluminescence*, Pergamon/Hassell Street Press, 1962.

K. Lehovec, C. A. Accardo, and E. Jamgochian, Injected Light Emission of Silicon Carbide Crystals, *Physical Review*, vol. 83, pp. 603–607, 1951.

H. C. Torrey and C. A. Whitmer, *Crystal Rectifiers*, McGraw–Hill, 1948.

W. Schottky, Halbleitertheorie der Sperrschicht (Semiconductor Theory of the Blocking Layer), *Naturwissenschaften*, vol. 26, p. 843, 1938.

O. V. Losev, Luminous Carborundum Detector and Detection Effect and Oscillations with Crystals, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, ser. 7, vol. 6, no. 39, pp. 1024–1044, 1928.

O. V. Losev, Luminous Carborundum Detector and Detection with Crystals (in Russian), *Telegrafiya i telefoniya bez provodov* (*Wireless Telegraphy and Telephony*), vol. 44, pp. 485–494, 1927.

H. J. Round, A Note on Carborundum, *Electrical World*, vol. 83, p. 309, 1907.

CHAPTER

7

LIGHT-EMITTING DIODES

7.1 PHOTON FLUX AND QUANTUM EFFICIENCY 200

7.2 SPATIAL, SPECTRAL, AND TEMPORAL PROPERTIES 207

7.3 LED MATERIALS AND DEVICE STRUCTURES 210

7.4 LEDS FOR ILLUMINATION 215

7.5 QUANTUM-DOT LIGHT-EMITTING DIODES (QLEDS) 218

7.6 ORGANIC LIGHT-EMITTING DIODES (OLEDS) 220

7.7 PEROVSKITE LIGHT-EMITTING DIODES (PELEDS) 223

7.8 LASER DIODES AND LIGHT-EMITTING DIODES 226

Robert J. Keyes (1927–2012) and **Theodore M. Quist (1931–2013)**, left and right, respectively, displayed the high-efficiency p–n junction light-emitting diode they co-invented while working at MIT Lincoln Laboratory in 1962. Operated at room temperature, this direct-bandgap GaAs semiconductor device generated spontaneous recombination radiation centered at 920 nm in the near infrared.

LED Lighting: Devices and Colorimetry. Malvin Carl Teich.
Google Books. Published 2024.

Light can be emitted from a semiconductor material as a result of electron–hole recombination. Materials capable of emitting such light do not glow at room temperature because the concentrations of thermally excited electrons and holes are too small to produce discernible light. However, an external source of energy can be used to produce electron–hole pairs in sufficient numbers so that they generate copious amounts of spontaneous recombination radiation, causing the material to luminesce. A convenient way of achieving this is to forward bias a p–n junction, which fosters the injection of electrons and holes in the vicinity of the junction. The ensuing recombination radiation is injection electroluminescence, as described in Chapter 6.

A light-emitting diode (LED) is a forward-biased p–n junction fabricated from a direct-bandgap semiconductor material that emits injection electroluminescence [Fig. 7.0-1(a)]. If the forward voltage is increased beyond a certain point, the number of electrons and holes in the junction region can become large enough to achieve a population inversion, whereupon stimulated emission (i.e., emission induced by the presence of photons) becomes more prevalent than absorption. Under those conditions, the junction region may be used as a semiconductor optical amplifier [Fig. 7.0-1(b)] or, with appropriate feedback, as a laser diode (LD) [Fig. 7.0-1(c)].

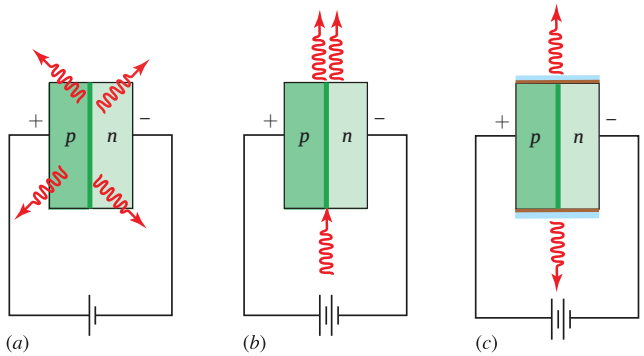


Figure 7.0-1 A forward-biased semiconductor p–n junction diode operated as: (a) a light-emitting diode (LED); (b) a semiconductor optical amplifier; and (c) a laser diode (LD).

The first truly useful LED was a GaAs p–n junction device fabricated in 1962 by Robert Keyes and Theodore Quist at MIT Lincoln Laboratory (p. 198). This device had high output power and a near-ideal (85%) quantum efficiency, as reported in a manuscript submitted to the *Proceedings of the IRE* on 25 May 1962 and published on 1 August 1962.[†] The development of the GaAs LED was announced publicly at the *Solid-State Device Research Conference* in Durham, New Hampshire on 9 July 1962.[‡] Similar results were reported within months by research groups at General Electric, IBM, and RCA, and the first commercial GaAs LED was offered by the Texas Instruments Corporation in the same time frame.

[†] R. J. Keyes and T. M. Quist, Recombination Radiation Emitted by Gallium Arsenide (Correspondence), *Proceedings of the IRE*, vol. 50, pp. 1822–1823, 1 August 1962 (submitted 25 May 1962).

[‡] R. J. Keyes and T. M. Quist, Radiation Emitted by Gallium Arsenide Diodes, presented at the *Solid-State Device Research Conference*, Durham, New Hampshire, 9–11 July 1962; abstract published in *IRE Transactions on Electron Devices*, vol. 9, No. 6, p. 503, July 1962. This GaAs device emitted continuous-wave (CW) spontaneous recombination radiation with a peak wavelength near $\lambda_p = 920$ nm ($h\nu = 1.35$ eV) when operated at 300 K, and close to 855 nm ($h\nu = 1.45$ eV) when operated at 77 K. The authors reported that the emitted light was perceived to be red, which accords with the results of subsequently conducted experiments (Example 8.5-2).

This chapter is devoted to investigating the operation of light-emitting diodes. These highly efficient electronic-to-photonic transducers are indispensable in many applications by virtue of their small size, high intensity, high efficiency, high reliability, ruggedness, and durability. Infrared LEDs are used in remote controls for consumer products such as optical mice, headphones, and keyboards, as well as in short-haul, modest-bit-rate communication systems. Visible LEDs are widely used in **indication applications** (in which the observer directly views the source); examples include mobile phones, indicator lights, computers, games, information displays, signage, traffic signals, backlighting, among others. Ultraviolet LEDs are useful in applications such as water purification, surgical sterilization, equipment decontamination, resin curing, and printing; they are also used for the detection of chemical and biological agents, many of which fluoresce at particular wavelengths when exposed to ultraviolet light.

Chapter 8 will describe the operation of the visual system and the perception of color in preparation for the exposition in Chapter 10 on the use of LEDs in **illumination applications** (in which the observer views the light scattered from objects illuminated by the source).

7.1 PHOTON FLUX AND QUANTUM EFFICIENCY

As is clear from the foregoing discussion, the simultaneous availability of electrons and holes substantially enhances the flux of spontaneously emitted photons from a semiconductor. Electrons are abundant in n-type material, and holes are abundant in p-type material, but the generation of copious amounts of light requires that both electrons and holes be plentiful in the same region of space. This condition may be readily achieved in the junction region of a forward-biased p–n diode (Sec. 5.6). As shown in Fig. 7.1-1, forward biasing causes holes from the p side and electrons from the n side to be forced into the common junction region by the process of minority carrier injection, where they recombine and emit photons.

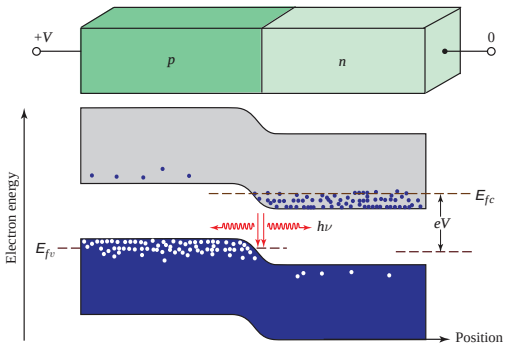


Figure 7.1-1 Energy-band diagram of a heavily doped p–n junction that is strongly forward biased by an applied voltage V (compare with the less strongly forward-biased energy-band diagram in Fig. 5.6-3). The dashed lines represent the quasi-Fermi levels, which are separated as a result of the bias. The simultaneous abundance of electrons and holes within the junction region results in strong electron–hole radiative recombination (injection electroluminescence).

The **light-emitting diode (LED)** is a *forward-biased p–n junction* with a large radiative recombination rate arising from injected minority carriers. The semiconductor material is *direct-bandgap* to ensure high quantum efficiency. In this section we determine the output power, as well as the spectral and spatial distributions of the light emitted from an LED, and derive expressions for the efficiency, responsivity, and response time. We occasionally refer to this genre of LEDs as **electroluminescent LEDs (ELLEDs)** to distinguish them from **phosphor-conversion LEDs (PCLEDs)**, which contain an

auxiliary photoluminescent material designed to modify the wavelength of the emitted electroluminescence (Chapter 10 is devoted to PCLEDs).

Internal Photon Flux and Internal Quantum Efficiency

Current and Current Density. A sketch portraying a simple forward-biased p–n homojunction diode is presented in Fig. 7.1-2. A DC current i injected into the junction region gives rise to a current density within the device given by

$$J = i/A\,, \tag{7.1-1}$$

where J is the current density and A is the junction area. The injected current increases the steady-state carrier concentration $\Delta\mathrm{n}$, which in turn fosters radiative recombination and spontaneous emission.

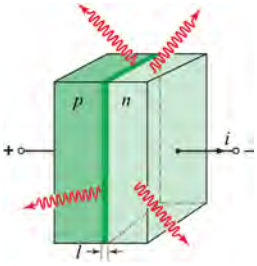


Figure 7.1-2 Schematic representation of a forward-biased LED. The injected current leads to an increase in the steady-state carrier concentration and to radiative recombination in the active region, and thence to spontaneous emission. The junction region has thickness l, area A, and volume V.

Carrier Injection and Concentration. The total number of carriers per second passing through the junction region is i/e, where e is the electronic charge, so the carrier injection rate R (carriers/s-cm^3) is

$$R = \frac{i/e}{V}\,. \tag{7.1-2}$$

Since $R = \Delta\mathrm{n}/\tau$, as per (6.4-1), the steady-state carrier concentration is

$$\Delta\mathrm{n} = \frac{(i/e)\tau}{V}\,. \tag{7.1-3}$$

Combining (7.1-2) with (6.4-2), which specifies that the **internal photon flux** is given by $\Phi = \eta_{\mathrm{IQE}} RV$, then leads to

$$\Phi = \eta_{\mathrm{IQE}}\,(i/e)\,. \tag{7.1-4}$$

Internal Quantum Efficiency (IQE). The **internal quantum efficiency** (IQE) defined in (5.5-10) and (5.5-12) is therefore given by

$$\boxed{\eta_{\mathrm{IQE}} = \frac{\Phi}{(i/e)}\,.} \tag{7.1-5}$$

Internal Quantum Efficiency

This simple and intuitively appealing formula quantifies the production of photons by electrons in the LED junction region. It states that the IQE is simply the ratio of the

generated photon flux Φ (photons/s) to the injected electron flux i/e (electrons/s), i.e., it is the fraction of the injected electron flux that is converted to photon flux. On a microscopic scale, $\eta_{\mathtt{IQE}}$ is the probability that an individual injected electron generates an emitted photon in the junction.

Enhancing the Internal Quantum Efficiency. The internal quantum efficiency can be enhanced by using a double-heterostructure configuration (Sec. 5.6), and even more so by employing a multiquantum-well active region (Sec. 5.7). Structures such as these accommodate higher carrier concentrations, which reduces the radiative lifetime (5.5-13) and thus increases the internal quantum efficiency (5.5-12). The internal quantum efficiency can also be enhanced by lattice matching the heterostructure confinement layers to the active region. Narrow quantum wells confine carriers even more tightly than double heterostructures, further enhancing the IQE. Still, the number of useful quantum wells is often limited because of the difficulty of populating all of them. Performance is also optimized by using high-quality materials to minimize defects and by avoiding the presence of surfaces to which both electrons and holes have access, which minimizes nonradiative recombination.

Yet another approach for increasing the IQE relies on making use of a **plasmonic LED**, in which metallic nanoparticles are embedded in a layer adjacent to a MQW active region. This serves to engender coupling between localized surface plasmons (LSPs) of the metallic nanoparticles and the light emitted from the proximate MQWs. This can provide a substantial enhancement of the spontaneous-emission rate $r_{\mathrm{sp}}(\nu)$ via the Purcell effect, which in turn leads to an increase in the IQE and increased LED output power.

Extraction Efficiency

The **extraction efficiency (XTE)**, also called the **transmission efficiency**, is a measure of the fraction of the internal photon flux that can be successfully extracted from an LED. In practice, it is calculated via a series of steps that recite the transmission through, and Fresnel reflection from, the various elements of the device structure.

Although the photon flux generated in the junction region is radiated uniformly in all directions, the flux that emerges from the device depends on the direction of emission. This is didactically illustrated by considering the photon flux transmitted through a planar material into air along three possible ray directions, denoted A, B, and C in the geometry of Fig. 7.1-3:

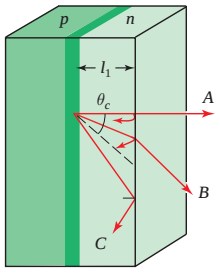


Figure 7.1-3 Not all light generated in an LED with a planar surface is able to emerge into air. Ray A is partly reflected. Ray B suffers more reflection. Ray C lies outside the critical angle and therefore undergoes total internal reflection, so that it is trapped in the structure.

Ray Traveling in a Direction Normal to the Surface. The photon flux traveling in the direction of ray A is attenuated by the factor

$$\eta_1 = \exp(-\alpha l_1), \tag{7.1-6}$$

where α is the absorption coefficient of the n-type material and l_1 is the distance from the junction to the surface of the device. Furthermore, for normal incidence, reflection

at the semiconductor–air boundary permits only a fraction of the light,

$$\eta_2 = 1 - \frac{(n-1)^2}{(n+1)^2} = \frac{4n}{(n+1)^2}\,, \tag{7.1-7}$$

to be transmitted, where n is the refractive index of the semiconductor material. The reflectance $\mathcal{R}$ for the photon flux, as established by the Fresnel equations, is given by (2.6-23). For GaAs, $n = 3.6$, so that $\eta_2 = 0.68$. The overall transmittance for the photon flux traveling in the direction of ray A is therefore $\eta_A = \eta_1\eta_2$.

Ray Traveling in a Direction Within the Critical Angle. The photon flux traveling in the direction of ray B has farther to travel and therefore suffers a larger absorption; it also has greater reflection losses. Thus, $\eta_B < \eta_A$.

Ray Traveling in a Direction Outside the Critical Angle. The photon flux traveling in the direction of a ray that lies outside a cone defined by the critical angle $\theta_c = \sin^{-1}(1/n)$, such as that depicted by ray C in Fig. 7.1-3, suffers total internal reflection in an ideal material and is not transmitted [see (1.3-2)]. As derived below, the fraction of the emitted photon flux that lies within the solid angle subtended by this cone, and is therefore extractable, is

$$\eta_3 = \tfrac{1}{2}(1 - \cos\theta_c) = \tfrac{1}{2}\left(1 - \sqrt{1 - 1/n^2}\right) \approx 1/4n^2. \tag{7.1-8}$$

For a material with refractive index $n = 3.6$, as an example, only 1.9% of the total generated photon flux can be transmitted. For a parallelepiped of refractive index $n > \sqrt{2}$, the ratio of the isotropically radiated photon flux that can emerge, to the total generated photon flux, is $3(1 - \sqrt{1 - 1/n^2}\,)$, as shown in Example 1.3-1. However, some fraction of the photons emitted outside the critical angle can be absorbed and reemitted within this angle, so that in practice, η_3 may assume a value larger than that specified by (7.1-8). Loss and Fresnel reflection must also be incorporated for these rays.

□ **Extractable Fraction of Photon Flux Specified in (7.1-8).** Snell's law for a ray traveling at the critical angle in a material of refractive index n and escaping into air ($n = 1$) at the surface is given by $n \sin\theta_c = 1 \cdot \sin(90°)$, so $\sin\theta_c = 1/n$ and $\cos\theta_c = \sqrt{1 - \sin^2\theta_c} = \sqrt{1 - 1/n^2}$. The area of the spherical cap atop the cone defining the critical angle is $A = \int_0^{\theta_c} 2\pi r \sin\theta\, r\, d\theta = 2\pi r^2(1 - \cos\theta_c)$. Since the area of the entire sphere is $4\pi r^2$, the fraction of the emitted photon flux that lies within the solid angle subtended by the cone is $A/4\pi r^2$. Hence, $\eta_3 = \frac{1}{2}(1 - \cos\theta_c) = \frac{1}{2}\left(1 - \sqrt{1 - 1/n^2}\right)$. Since $\sqrt{1 - 1/n^2} \approx 1 - 1/2n^2$ for $1/n^2 \ll 1$, we have $\eta_3 \approx \frac{1}{2}(1/2n^2) = 1/4n^2$, as in (7.1-8). ■

EXAMPLE 7.1-1. ***Extraction of Light from a Planar-Surface LED.***

(a) As established above, the critical angle within which light can escape from a material of refractive index n into air at a planar surface is $\theta_c = \sin^{-1}(1/n)$. In accordance with (7.1-8), if absorption and Fresnel reflection are ignored, the fraction of the photon flux that is not trapped by total internal reflection, and can therefore be extracted, is $\eta_3 \approx 1/4n^2$. The numerical values for θ_c and η_3 for GaAs ($n = 3.6$), GaN ($n = 2.5$), and a transparent polymer ($n = 1.5$) are therefore:

$$\begin{aligned} \theta_c(\text{GaAs}) &= \sin^{-1}(1/3.6) = 16.1° \quad \text{and} \quad \eta_3(\text{GaAs}) = 0.019 \\ \theta_c(\text{GaN}) &= \sin^{-1}(1/2.5) = 23.6° \quad \text{and} \quad \eta_3(\text{GaN}) = 0.040 \\ \theta_c(\text{polymer}) &= \sin^{-1}(1/1.5) = 41.8° \quad \text{and} \quad \eta_3(\text{polymer}) = 0.111. \end{aligned}$$

(b) In the absence of Fresnel reflection, the fraction of the photon flux that can be extracted can be enhanced by coating the LED surface with a transparent material whose refractive index lies between that of the LED material and air. Consider the example of a GaAs LED ($n_1 = 3.6$) coated with a transparent polymer ($n_2 = 1.5$), and ignore absorption and Fresnel reflection at the semiconductor–polymer boundary. The critical angle of a ray traveling from GaAs into the polymer is established from Snell's law, $n_1 \sin\theta_{c1} = n_2$ so that $\theta_{c1} = \sin^{-1}(n_2/n_1) = \sin^{-1}(1.5/3.6) = 24.6°$. Using (7.1-8), we therefore arrive at $\eta_3 = \frac{1}{2}[1 - \cos(24.6°)] = 0.045$. As specified above in (a) , light escaping from GaAs into air yields $\eta_3(\text{GaAs}) = 0.019$ so the enhancement in the fraction of extracted light offered by the polymer is $0.045/0.019 \approx 2.4$.

(c) It can be shown, however, that if Fresnel reflection at the semiconductor–polymer and polymer–air interfaces are incorporated (but absorption is ignored), use of a material of intermediate refractive index does not enhance the fraction of photon flux that can be transferred from the LED into air. The additional Fresnel reflection at the added interface negates the benefit of the reduction in refractive-index mismatch attained by incorporating that interface. The use of an intermediate-index material in the form of a quarter-wave film can sometimes be useful in this connection, however.

Enhancing the Extraction Efficiency.

Antireflection coatings and other techniques can be used to reduce Fresnel reflection and thereby increase the XTE, as discussed below.

Geometry. The extraction efficiency (XTE) can be enhanced in a multitude of ways. One approach involves selecting a geometry for the **LED die** (**LED chip**) that allows a greater fraction of the light to escape. A spherical dome surrounding a point source at its center, for example, permits all rays to escape, although they remain subject to Fresnel reflection. As illustrated in Fig. 7.1-4, several other geometries offer enhanced extraction efficiencies in comparison with the parallelepiped: hemispherical domes, cylindrical structures (which have an escape ring along the perimeter in addition to the escape cone toward the top surface), inverted cones, and truncated inverted pyramids. However, geometries that entail complex processing steps are often avoided in practice because of increased manufacturing costs. Simple planar-surface-emitting LEDs are suitable when the intended viewing angle deviates little from the normal or when the light is coupled into an optical fiber, as it is in telecommunications applications.

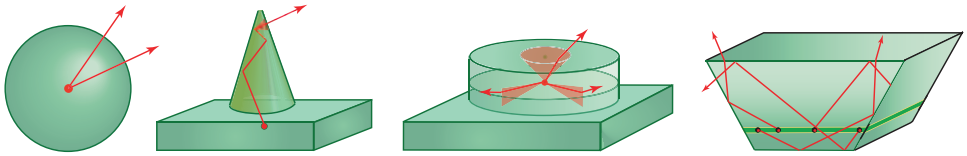


Figure 7.1-4 LED-die geometries that offer enhanced extraction efficiencies relative to the parallelepiped.

Surface Roughening. Another approach is to roughen the planar surface, which enhances the extraction efficiency by permitting rays beyond the critical angle to escape via scattering, as illustrated in Fig. 7.1-5. Indeed, an irregular surface appears automatically under certain growth conditions. Alternatively, the emission surface can be textured, such as with an array of microscopic cones or pyramids, or with nanoparticles. Another twist is to make use of the morphology of the light-emitting organs of some biological organisms, such as fireflies, which serve to enhance light extraction by reducing refractive-index mismatch and total internal reflection. *Bioinspired surface patterning* has been successfully used to increase the extraction efficiency of LEDs and OLEDs.

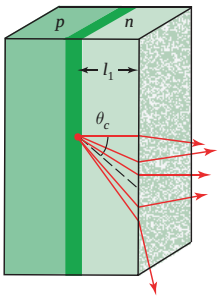


Figure 7.1-5 An LED with a roughened planar surface permits rays beyond the critical angle to escape, thereby increasing the extraction efficiency (XTE).

Contact Geometry. Top-emitting LEDs often make use of current-spreading layers (also referred to as window layers), which are transparent conductive semiconductor layers that spread the region of light emission beyond that surrounding the electrical contact. Current-blocking layers, which prevent current from entering the active region below the top contact, can also be used to control the light emission. The contact geometry can be designed to maximize light transmission.

Other Techniques. A whole host of other techniques are also used to enhance the extraction efficiency. These include the use of distributed Bragg reflectors between the active layer and an absorbing substrate to reflect the light back toward the desired direction of emission, and reflective and transparent contacts. Another favored approach is the use of a transparent substrate in conjunction with **flip-chip packaging**, which allows the light to be extracted through the substrate rather than through the top surface of the device. The XTE can also be enhanced by guiding light to the surface of the device via a 2D photonic crystal, such as a regular array of 100–250-nm diameter holes formed in the current-spreading layer.

External Photon Flux and External Quantum Efficiency

External Photon Flux.

The **external photon flux** $\Phi_{\mathtt{O}}$, also called the **output photon flux**, is related to the internal photon flux $\Phi = \eta_{\mathtt{IQE}}\,(i/e)$ provided in (7.1-4) by

$$\Phi_{\mathtt{O}} = \eta_{\mathtt{XTE}}\Phi = \eta_{\mathtt{XTE}}\eta_{\mathtt{IQE}}\,(i/e)\,. \tag{7.1-9}$$

While the IQE relates the internal photon flux to the injected electron flux i/e via (7.1-5), the extraction efficiency XTE specifies the fraction of the internal photon flux that is successfully extracted from the structure, as depicted in Figs. 7.1-3–7.1-5.

External Quantum Efficiency (EQE).

A combined quantum efficiency that accommodates both the internal and extraction quantum efficiencies is the **external quantum efficiency (EQE)**,

$$\boxed{\eta_{\mathtt{EQE}} = \eta_{\mathtt{XTE}}\eta_{\mathtt{IQE}}.} \tag{7.1-10}$$

External Quantum Efficiency

Using this quantity, the external photon flux in (7.1-9) can be written as

$$\boxed{\Phi_{\mathtt{O}} = \eta_{\mathtt{EQE}}\,(i/e)\,,} \tag{7.1-11}$$

External Photon Flux

which declares that the EQE is simply the ratio of the external photon flux to the injected electron flux, i.e., the ratio of the flux of emitted photons to that of injected electrons.

Since each photon is endowed with energy $E = h\nu$, the LED optical output power P_0 is readily written in terms of the external photon flux provided in (7.1-11) as

$$\boxed{P_0 = h\nu\Phi_0 = \eta_{\text{EQE}}\, h\nu\, (i/e)\,.} \qquad (7.1\text{-}12)$$

LED Output Power

In the current state of the technology, the internal quantum efficiency (IQE) for an LED usually falls between 50 and 100%, and the average extraction efficiency (XTE) is in the vicinity of 50%. The external quantum efficiency (EQE) therefore generally lies in the range between 25% and 75%.

Power-Conversion Efficiency (PCE). A related measure of LED performance is the **power-conversion efficiency (PCE)**, also known as the **energy-conversion efficiency (ECE)** and the **overall efficiency**. In some quarters, this quantity is also called the wall-plug efficiency but we abstain from use of this terminology to avoid confusion with the wall-plug luminous efficiency, a different (but similarly named) measure that will be introduced in Sec. 8.9.

The PCE is defined as the ratio of the **emitted optical power** P_0 to the **electrical drive power** P_{EL}, where

$$\boxed{P_{\text{EL}} = iV,} \qquad (7.1\text{-}13)$$

Electrical Drive Power

so that

$$\boxed{\eta_{\text{PCE}} = \frac{P_0}{P_{\text{EL}}} = \frac{P_0}{iV} = \eta_{\text{EQE}}\frac{h\nu}{eV}\,,} \qquad (7.1\text{-}14)$$

Power-Conversion Efficiency

where V is the voltage drop across the device. For $h\nu \approx eV$, we obtain $\eta_{\text{PCE}} \approx \eta_{\text{EQE}}$. Like the EQE, therefore, the PCE for a well-designed device lies in the range

$$1/4 \lesssim \eta_{\text{PCE}} \lesssim 3/4\,. \qquad (7.1\text{-}15)$$

The empirical value of the power-conversion efficiency depends on the wavelength at which the LED operates: $\eta_{\text{PCE}} \approx 3/4$ for blue, $\approx 1/2$ for red, and $\approx 1/4$ for green (as a result of the "green gap" discussed in Sec. 7.3); η_{PCE} is even smaller for amber LEDs. The PCE is dimensionless since it has units of W/W.

Resonant-Cavity LEDs. The quantum efficiencies η_{EQE} and η_{PCE} may be enhanced by making use of a **resonant-cavity light-emitting diode (RCLED)**. A pair of mirrors (e.g., distributed Bragg reflectors) is used to confine injection electroluminescence to a wavelength-sized, resonant microcavity in one dimension. RCLEDs exhibit a number of attractive features: 1) the spontaneous-emission rate is enhanced by the Purcell effect, which results in an increase in the IQE; 2) the spectral width of the emitted light is reduced below kT when the cavity resonance is narrower than the spectral-intensity profile;) the temperature stability is then also enhanced because the cavity is less sensitive to temperature changes than is the semiconductor energy gap; and 4) the

emission is more narrowly confined in angle, which results in an increase in the XTE. As illustrated in Fig. 7.1-6, a substantial fraction of the light is emitted into a resonant mode whose angular extent falls principally within the extraction cone.

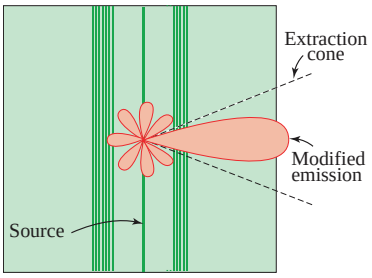


Figure 7.1-6 A plane-parallel-reflector resonant-cavity light-emitting diode (RCLED). Two closely spaced reflectors (the one at left with a reflectance near 100% and the one at right with a reflectance of, say, 50%) form a wavelength-size cavity in one dimension that confines the light and funnels a large portion of it into a spatial region that lies within the extraction cone.

A photonic-crystal structure can also be incorporated in an RCLED to guide much of the residual light toward the surface of the device, thereby further increasing the XTE. The increased values of the IQE and the XTE for RCLEDs lead directly to enhanced values of the external and power-conversion quantum efficiencies, $\eta_{\mathrm{EQE}} = \eta_{\mathrm{XTE}}\eta_{\mathrm{IQE}}$ and $\eta_{\mathrm{PCE}} = \eta_{\mathrm{EQE}}(h\nu/eV)$, respectively. However, RCLEDs are inherently low power devices by virtue of the small sizes of their active regions.

Responsivity

The responsivity R of an LED is defined as the ratio of the emitted optical power P_0 to the injected current i, i.e., $\mathsf{R} = P_0/i$. Using (7.1-12), we obtain

$$\mathsf{R} = \frac{P_0}{i} = \frac{h\nu\,\Phi_0}{i} = \eta_{\mathrm{EQE}}\frac{h\nu}{e}. \tag{7.1-16}$$

The responsivity in W/A, when λ_0 is expressed in μm, is then

$$\boxed{\mathsf{R} = \eta_{\mathrm{EQE}}\frac{1.24}{\lambda_0}.} \qquad \begin{array}{r}(7.1\text{-}17)\\ \text{LED Responsivity}\\ (\text{W/A};\ \lambda_0\ \text{in}\ \mu\text{m})\end{array}$$

For example, if $\lambda_0 = 1.24\ \mu$m, then $\mathsf{R} = \eta_{\mathrm{EQE}}$ W/A; if η_{EQE} were unity, the maximum optical power that could be produced by an injection current of 1 mA would be 1 mW. Thus, for $\eta_{\mathrm{EQE}} = 1/2$ at $\lambda_0 = 1.24\ \mu$m, we have $\mathsf{R} = 1/2$ mW/mA.

In accordance with (7.1-12), the LED output power P_0 is proportional to the injected current i. In practice, however, this relationship is valid only over a restricted range. For the particular device whose **light–current (L–i) curve** is shown in Fig. 7.1-7, the emitted optical power is proportional to the injection (drive) current only when the latter is less than about 20 mA. In this range, the responsivity has a constant value of about 0.3 mW/mA, as determined from the slope of the curve. For larger drive currents, saturation causes the proportionality to fail; the responsivity then declines with increasing drive current. Since $\lambda_0 = 0.420\,\mu$m for this LED, (7.1-17) reveals that it has an EQE $= 0.10$.

7.2 SPATIAL, SPECTRAL, AND TEMPORAL PROPERTIES

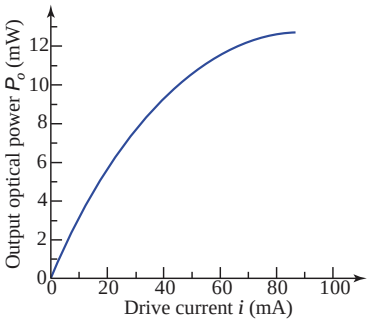


Figure 7.1-7 Optical power at the output of an LED versus injection (drive) current. This MQW InGaN/GaN LED emits in the violet region of the spectrum at $\lambda_0 = 420$ nm; the device structure is exhibited in Fig. 7.3-4.

Spatial Distribution

The far-field radiation pattern for light emitted into air from a planar surface-emitting LED is similar to that of a Lambertian radiator. The intensity varies as $\cos\theta$, where θ is the angle from the emission-plane normal to the direction of view; the intensity decreases to half its value at $\theta = 60°$. This angular pattern represents the projected area of a uniform surface intensity when viewed at different angles. The uniform intensity distribution in turn follows from the randomized photon directions inside the LED that emerge from multiple photon scatterings.

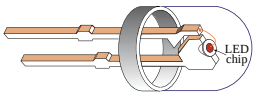


Figure 7.2-1 Polymer-encapsulated LED in a 5-mm-diameter dual in-line package (DIP). Encapsulation protects the LED chip (die), increases light extraction by reducing refractive-index mismatch, and serves as a lens to shape the beam.

LEDs are often encapsulated in transparent polymer lens domes such as epoxy or silicone for a number of reasons (Fig. 7.2-1). Lenses of different shapes alter the emission pattern in different ways, as illustrated schematically for hemispherical and parabolic lenses in Fig. 7.2-2. Polymer lenses can also enhance the XTE. A lens with a refractive index close to that of the semiconductor optimizes the extraction of light from the semiconductor into the polymer. The shape of the lens can then be tailored so as to maximize the extraction of light at the polymer–air interface. Polymer materials usually have refractive indices that are intermediate between those of semiconductors and air and, in practice, yield an enhancement in light extraction by a factor of 2 to 3. Molded acrylic or polycarbonate collimators that make use of total internal reflection in conjunction with refraction are often used to provide parallel light rays for LED lighting applications, as illustrated in Fig. 1.4-3.

A parameter that is often used to represent the angular width of a light beam emitted from an LED is the **viewing angle** (or **50%-power angle**) $2\theta_{1/2}$, which is defined as twice the half-angle at which the intensity decreases to half its maximum value. The radiation pattern from edge-emitting LEDs and LDs is usually quite narrow and the intensity can often be empirically described by the function $\cos^s\theta$, with $s > 1$. If $s = 10$, for example, the intensity decreases to half its value at $\theta \approx 21°$.

Spectral Density

The spectral density $r_{sp}(\nu)$ of the spontaneous injection electroluminescence emitted by an LED is sketched as a function of frequency in Fig. 6.4-5. The direct-bandgap theory that underlies this result, which is summarized in Sec. 6.4, assumes that the current injected into a semiconductor p–n junction induces quasi-equilibrium conditions. In accordance with the results presented in Example 6.4-4, the frequency spectral width

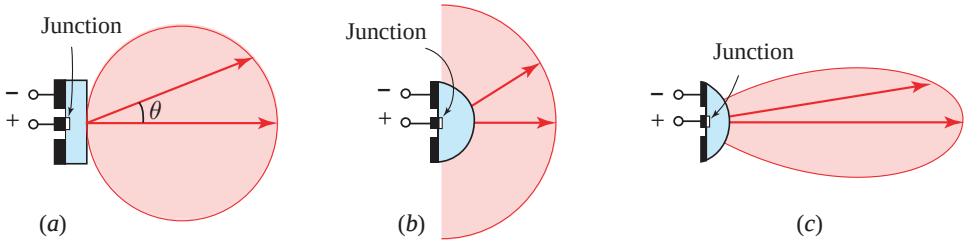


Figure 7.2-2 Radiation patterns of surface-emitting LEDs: (a) In the absence of a lens, the spatial radiation pattern is roughly Lambertian so that the intensity decreases to half its value at $2\theta_{1/2} = 120°$. (b) Spatial radiation pattern with a hemispherical lens. (c) Spatial radiation pattern with a parabolic lens.

$\Delta\nu$ of the LED spectrum is independent of peak frequency, while the wavelength spectral width $\Delta\lambda$ increases quadratically with peak wavelength λ_p, i.e.,

$$\Delta\lambda \propto \lambda_p^2 . \tag{7.2-1}$$

LED Spectral Width

The validity of this result is borne out in Example 7.2-1.

EXAMPLE 7.2-1. ***Behavior of LED Spectral Width with Wavelength.*** Figure 7.2-3 displays the observed spectral densities for a collection of LEDs with peak wavelengths that cover the ultraviolet (indicated as magenta), visible (curve colors match the peak wavelengths), and near infrared (indicated as gray). AlN, with the smallest bandgap wavelength of all binary III–nitride compounds, generates light at 210 nm in the mid-ultraviolet region of the spectrum. As will be discussed in Sec. 7.3, AlGaN and AlInGaN are typically used to fabricate LEDs in the near- and mid-ultraviolet; InGaN usually serves the violet, blue, and green; and AlInGaP is the material of choice in the yellow, orange, and red. InGaAsP is usually used in the near infrared. The spectral widths displayed in Fig. 7.2-3 roughly increase as λ_p^2 over the full spectral range, in accordance with (7.2-1).

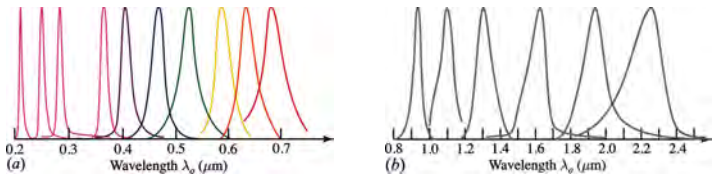


Figure 7.2-3 Spectral density versus wavelength for a collection of LEDs operating: (a) in the ultraviolet and visible regions; and (b) in the near-infrared region. The peak intensities are all normalized to unity. As indicted in (6.4-17), the wavelength spectral widths $\Delta\lambda$ increase roughly as λ_p^2 in all spectral regions. Note the different abscissa scales in the two panels.

Temporal Response

The response time of LEDs used for illumination is usually limited by the RC time constant of the device, τ_{RC}, because the junction area, and therefore the capacitance, is large. The response time of communication-system LEDs, in contrast, is generally

limited principally by the lifetime τ of the injected minority carriers that are responsible for radiative recombination. For a sufficiently small injection rate R, the injection/recombination process can be described by a first-order linear differential equation (Sec. 5.5), and therefore by the response to sinusoidal signals. An experimental determination of the highest frequency at which an LED can be effectively modulated is easily obtained by measuring the output light power in response to sinusoidal electric currents of different angular frequencies χ. If the injected current assumes the form $i = \bar{\imath} + i_1 \cos(\chi t)$, where i_1 is sufficiently small so that the emitted optical power P_0 varies linearly with the injected current, the emitted optical power behaves as $P_0 = \overline{P} + P_1 \cos(\chi t + \varphi)$.

The associated transfer function, which is defined as $H(\chi) = (P_1/i_1)\exp(j\varphi)$, assumes the form

$$H(\chi) = \frac{\mathrm{R}}{1 + j\chi\tau}, \tag{7.2-2}$$

which is characteristic of a resistor–capacitor circuit. The risetime of the LED is τ (s) and its 3-dB bandwidth is $B = 1/2\pi\tau$ (Hz). A larger bandwidth B is therefore attained by decreasing the risetime τ, which comprises contributions from both the radiative lifetime τ_r and the nonradiative lifetime τ_nr through the relation $1/\tau = 1/\tau_\mathrm{r} + 1/\tau_\mathrm{nr}$. However, reducing τ_nr results in an undesirable reduction of the internal quantum efficiency $\eta_\mathrm{IQE} = \tau/\tau_\mathrm{r}$. It may therefore be desirable to maximize the internal quantum efficiency–bandwidth product $\eta_\mathrm{IQE}B = 1/2\pi\tau_\mathrm{r}$ rather than the bandwidth alone. This requires a reduction of only the radiative lifetime τ_r, without a reduction of τ_nr, which can be achieved by carefully choosing the semiconductor material and doping level. Typical risetimes of LEDs are in the range 1 to 50 ns, corresponding to bandwidths of hundreds of MHz.

7.3 LED MATERIALS AND DEVICE STRUCTURES

LED Materials

Photonics was revolutionized in the 1950s by the development of single-crystal III–V semiconductors, materials that do not occur in nature. The properties of important binary, ternary, and quaternary III–V semiconductors, in the context of their constituent elements and their placement in the periodic table, were introduced in Sec. 5.3. Many of these materials have direct bandgaps and high internal quantum efficiencies, as well as long lifespans.

Today's LED (and LD) industries are built almost exclusively around direct-bandgap ternary and quaternary III–V material systems, such as those identified in Fig. 7.3-1. The upper and lower abscissas of this figure represent bandgap wavelength λ_g (μm) and bandgap energy E_g (eV), respectively [the relationship between these quantities is provided in (5.1-1) and (5.1-2)]. Just two compositionally tunable III–V materials, InGaN and AlInGaP, suffice for generating bright LED light across virtually the entire visual spectrum, as exemplified in Fig. 7.2-3. Devices that make use of these materials can be grown on readily available substrates, are robust in the face of degradation induced by defects, and can be manufactured reliably and inexpensively. Nevertheless, the performance of red, orange, and amber LEDs fabricated from these materials, and particularly green LEDs, is inferior to that available with blue LEDs.

It is apparent from Fig. 7.3-1 (and Table 5.3-1) that the bandgap wavelengths for selected triplets of binary III–V compounds obey:

$$\lambda_g(\mathrm{GaN}) < \lambda_g(\mathrm{GaP}) < \lambda_g(\mathrm{GaAs}) \tag{7.3-1}$$

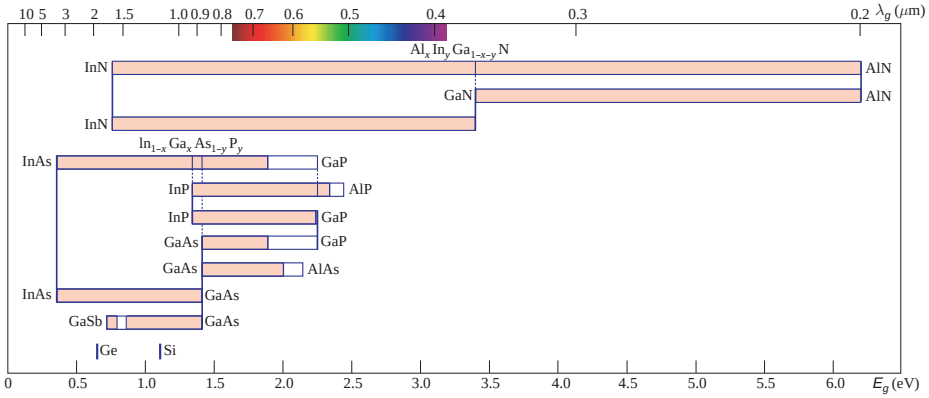


Figure 7.3-1 Bandgap wavelength λ_g (μm) (upper abscissa), and corresponding bandgap energy E_g (eV) (lower abscissa), for commonly used elemental and III–V binary, ternary, and quaternary semiconductor materials. Consecutive rows, beginning from the top, represent AlInGaN, AlGaN, InGaN, InGaAsP, AlInGaP, InGaP, GaAsP, AlGaAs, InGaAs, and GaAsSb. Shaded and unshaded regions indicate direct- and indirect-bandgap compositions, respectively.

$$\lambda_g(\text{AlP}) < \lambda_g(\text{GaP}) < \lambda_g(\text{InP})\,. \tag{7.3-2}$$

The entries in (7.3-1) all have Ga in common while the entries in (7.3-2) all have P in common. These inequalities are a consequence of the fact that N, P, and As in (7.3-1), and Al, Ga, and In in (7.3-2), reside in rows of the periodic table (Fig. 5.3-1) that have successively larger principal quantum numbers. This signifies progressively larger atomic radii, less tight binding, and larger bandgap wavelengths.

The bandgap wavelength λ_g of a semiconductor is instrumental in determining the wavelength λ_0 of the light emitted by an LED fabricated from that material.

Early LEDs

GaAs. The first III–V material to play an important role in photonics was GaAs. Robert Keyes and Theodore Quist, working at MIT Lincoln Laboratory in 1962, used this direct-bandgap, binary semiconductor to fabricate the first truly useful LED (see p. 198 and footnotes on p. 199). The near-ideal (85%) quantum efficiency of the device was the result of a novel fabrication technique in which the p-type layer was formed by indiffusion of Zn into the bulk of the n-type single crystal (Zn is a column-II acceptor for GaAs, as discussed in Sec. 5.3). The peak emission wavelength of this device was $\lambda_p = 920$ nm at $T = 300$ K, which Keyes and Quist perceived as red (Example 8.5-2). Within months, a number of other research groups had also fabricated LEDs (as well as LDs) from GaAs.

Not long thereafter, several other direct-bandgap, binary III–V semiconductors, grown by vapor-phase epitaxy (VPE) and liquid-phase epitaxy (LPE), were also fabricated in the form of LEDs and LDs, emitting light in the vicinity of their bandgap wavelengths. These included GaSb, InP, InAs, and InSb (see Table 5.3-1). Yet other binary semiconductors, including II–VI compounds (Fig. 5.3-3), followed suit.

GaAsP. Adding phosphorus to GaAs forms the ternary semiconductor $\text{GaAs}_{1-x}\text{P}_x$, a material whose bandgap wavelength decreases as the mole-fraction of phosphorus increases (Fig. 7.3-1). The first $\text{GaAs}_{1-x}\text{P}_x$ LED was fabricated by Holonyak and Bevacqua in 1962 (see p. 338 and footnote on p. 339); this device functioned as an LED at room temperature and as a laser diode at sufficiently low temperatures. However, the

external quantum efficiency of $GaAs_{1-x}P_x$ degrades markedly as the mole-fraction of phosphorus increases beyond $\approx 40\%$ because the substantial lattice mismatch between the GaAs substrate and the GaAsP epilayer gives rise to a high density of dislocations. Moreover, the bandgap ultimately changes from direct to indirect with increasing phosphorus level, which is an impediment to attaining efficient emission at wavelengths shorter than the red. However, the light emitted by a GaAsP LED appears brighter to the eye than that emitted by a GaAs device of the same optical power because its shorter wavelength is more effective at stimulating the photopic visual system (Fig. 8.5-3). Nevertheless, the light emitted from both a GaAsP and a GaAs LED is perceived to have the same hue (Example 8.5-2).

GaAsP:N. George Craford (see p. 338) and his colleagues demonstrated in the early 1970s that emission in the red, orange, yellow, and green could be elicited from GaAsP, as well as from other indirect-bandgap materials such as GaP, by doping the material with isoelectronic nitrogen impurities (GaAsP:N or GaP:N). The impurities serve as a layer of optically active traps within the bandgap. Since isoelectronic impurities have highly localized wavefunctions, the nitrogen atoms can be viewed as residing at sharply localized positions in the lattice (small Δx). By virtue of the uncertainty principle set forth in (A.2-9) of Appendix A, a small value of Δx is accompanied by a large value of Δp, thereby enabling large momentum changes to be accommodated. The emission process can therefore proceed via a nonradiative transition from the conduction-band minimum to the nitrogen level and a radiative transition from the nitrogen level to the valence-band maximum, the momentum change being absorbed by the isoelectronic nitrogen atom. Indeed, the external quantum efficiency of GaAsP:N devices exceeds that of GaAsP devices over the entire range of emission wavelengths. Despite the fact that all external quantum efficiencies are typically $< 1\%$, LEDs made of GaAsP, GaAsP:N, and GaP:N are inexpensive to fabricate and continue to be used in low-intensity applications such as indicator lamps.

AlGaAs. Just as adding phosphorus to GaAs leads to a reduction in bandgap wavelength, so too does adding aluminum. As is evident in Fig. 7.3-1, the bandgap wavelength of the ternary compound $Al_xGa_{1-x}As$ can be compositionally tuned over the direct-bandgap range $630 \leqslant \lambda_g \leqslant 873$ nm. This encompasses the near infrared and the red, falling just short of the orange (Fig. 2.4-1). Unlike GaAsP, AlGaAs has the merit that lattice matching to GaAs is maintained for all mole fractions of aluminum [Fig. 5.3-2(a)]. However, nonuniform carrier distributions in the active region of $Al_xGa_{1-x}As/GaAs$ multiquantum-well structures tend to adversely affect performance, so AlGaAs LEDs are often instead fabricated in the form of double-heterostructure configurations that make use of different barrier and well compositions, i.e., as $Al_xGa_{1-x}As/Al_yGa_{1-y}As$. AlGaAs structures are generally considered to be less reliable than AlInGaP devices inasmuch as layers with high Al content are subject to corrosion and oxidation.

Red, Orange, and Yellow LEDs

AlInGaP. The quaternary semiconductor $Al_xIn_yGa_{1-x-y}P$ provides bright emission in the red, orange, and yellow regions of the spectrum (Fig. 2.4-1), and is widely used in LED lighting applications (see p. 338). Much as with GaAs, the addition of Al and In to GaP decreases and increases λ_g, respectively. The bandgap wavelength of AlInGaP can be compositionally tuned over a region that includes the longer wavelength reaches of the visible band and a limited portion of the near-infrared band (Fig. 7.3-1). Specifically, emission can be attained over the range 506 nm (AlP) $\leqslant \lambda_g \leqslant$ 919 nm (InP), although indirect-bandgap behavior prevails over the shortest wavelength reaches of this range. As such, this material is not suitable for fabricating devices that emit in

the green. The substrate for AlInGaP devices is usually GaAs, although efficiency can be increased by making use of a transparent, wafer-bonded GaP substrate instead. Other enhancements include the use of multiquantum-well active regions and resonant-cavity (RC) configurations that provide directed emission patterns.

Green, Blue, and Violet LEDs

InGaN. In the same way that AlInGaP provides bright emission in the red, orange, and yellow, the direct-bandgap ternary semiconductor $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{N}$ provides bright emission in the green, blue, and violet (Fig. 2.4-1). Together with AlInGaP, InGaN is widely used in LED lighting applications (see p. 306). The addition of In to GaN increases its bandgap wavelength, much as it does for GaAs and GaP. The bandgap wavelength for InGaN can, in principle, be compositionally tuned over the (broad) wavelength range 366 nm (GaN) $\leqslant \lambda_g \leqslant$ 1.91 μm (InN) (Fig. 7.3-1). In practice, however, it becomes increasingly difficult to grow InGaN as the concentration of In increases, so this material is typically used only over the more modest wavelength range $366 \leqslant \lambda_g \leqslant 580$ nm, comprising the near-ultraviolet, violet, blue, and green. While InGaN devices do serve the green, they do so with reduced external quantum efficiency. III–nitride compounds are often grown by MBE, MOCVD, or HVPE, and usually make use of sapphire, Si, or SiC substrates. Lattice mismatch and large dislocation concentrations are well-tolerated by the III–nitrides (unlike the III–V arsenides and phosphides). Buffer layers are also used for accommodating differences in thermal-expansion coefficients. As with AlInGaP, other enhancements include the use of MQW active regions and RC configurations. As will be discussed extensively in Chapter 10, the emission of blue light by InGaN LEDs enables the direct generation of bright white light via phosphor-conversion devices.

Green Gap. It is apparent from the discussions presented above that AlInGaP is not suitable for fabricating green-emitting devices because of its transition to indirect-bandgap behavior, and that InGaN exhibits reduced external quantum efficiency in the green as a result of indium-related materials-growth issues. These fundamental limitations lead to what is known as the "green gap." A number of novel semiconductor materials are currently under consideration for possible use in fabricating green LEDs; these include the III–V compounds $\mathrm{Al}_x\mathrm{In}_{1-x}\mathrm{P}$ and $\mathrm{GaN}_{1-x}\mathrm{As}_x$, certain II–IV–N alloys, and halide perovskites (Sec. 5.9). It will be some time before it can be established whether any of these materials can rise to the challenge.

Ultraviolet LEDs

GaN. Gallium nitride is a direct-bandgap binary semiconductor whose bandgap wavelength $\lambda_g = 366$ nm falls in the near-ultraviolet region. Much as GaAs was the progenitor of InGaAs, AlGaAs, and InGaAsP, GaN served as the progenitor of InGaN, AlGaN, and AlInGaN. The growth of GaN was not perfected until the early 1990s, an achievement that was hailed as an important breakthrough in LED technology at the time because it signaled that blue InGaN LEDs, and therefore metameric-white LEDs, were not far behind (see p. 306 and footnote on p. 307).

AlInGaN. While InGaN and AlGaN can be compositionally tuned over a broad range of bandgap wavelengths (Figs. 5.3-2(b) and 7.3-1), the direct-bandgap quaternary semiconductor $\mathrm{Al}_x\mathrm{In}_y\mathrm{Ga}_{1-x-y}\mathrm{N}$ has the additional merit that it can be lattice matched to a GaN substrate for certain values of x and y, thereby increasing device quantum efficiency. This lattice matching is analogous to that of AlInGaP to GaAs and of InGaAsP to InP. LEDs fabricated from lattice-matched AlInGaN are generally employed over a wavelength range extending from 250 nm in the MUV (the bandgap wavelength of

AlInN that is lattice matched to GaN) to 366 nm in the NUV (the bandgap wavelength of GaN). Multiquantum-well structures of the form AlInGaN/InGaN/AlInGaN serve as active regions for these devices. The inclusion of indium in AlGaN also yields an enhancement in the internal quantum efficiency. AlInGaN can also serve as a transparent contact layer.

Device Structures

LEDs can be constructed either in surface-emitting or edge-emitting geometries (Fig. 7.3-2). The surface-emitting LED emits light from a face of the device that is parallel to the plane of the active region. The edge-emitting LED, in contrast, emits light from the edge of the active region. We provide brief descriptions of the principal III–V

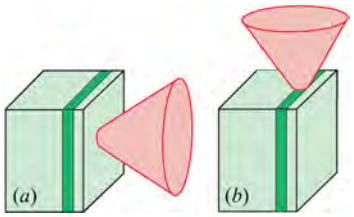


Figure 7.3-2 Sketch of (*a*) a surface-emitting LED; (*b*) an edge-emitting LED.

semiconductor compounds used to fabricate LEDs, along with schematic illustrations of several representative device structures. Along the way, we highlight a number of applications for LEDs and LDs in the IR, visible, and UV.

EXAMPLE 7.3-1. ***Surface-Emitting AlInGaP LED.*** AlInGaP is widely used in the red, orange, yellow-orange (amber), and yellow regions of the spectrum. Multiquantum-well devices fabricated from AlInGaP and from InGaN (for green and blue) are the devices of choice for applications such as traffic signals, signage, and LED lighting. AlInGaP/InGaP LEDs are also sometimes used in plastic fiber-optic communication systems that operate in the red (Fig. 7.3-3).

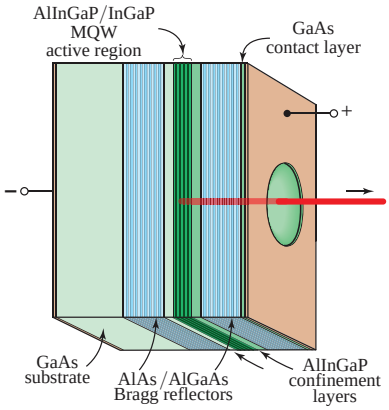


Figure 7.3-3 A Surface-emitting AlInGaP/InGaP 650-nm MQW RCLED used in short-haul, plastic-fiber communication systems. A top-emitting structure is employed because of the opacity of the GaAs substrate in this device. The distributed Bragg reflectors comprise AlAs/AlGaAs layers with an aluminum content that is sufficiently high so that the 650-nm light is transmitted through it. A lens enhances coupling of the light to a fiber.

EXAMPLE 7.3-2. ***A Surface-Emitting InGaN LED.*** As with AlInGaP, the use of MQW structures such as GaN/InGaN (Fig. 7.3-4) enhances device quantum efficiency. The substrate is often GaN on sapphire. However, the number of quantum wells is generally limited because of population limits imposed by the hole diffusion length; low and/or thin barriers are preferred. Performance can also be enhanced by the use of resonant-cavity devices.

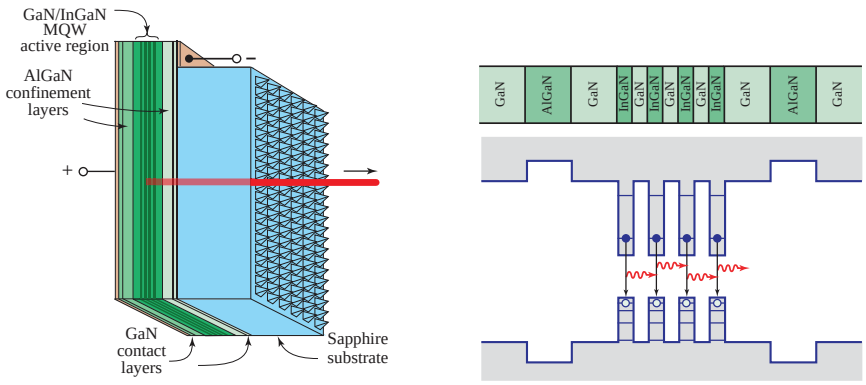


Figure 7.3-4 Flip-chip packaged, surface-emitting GaN/InGaN MQWLED operating at $\lambda_0 = 420$ nm in the violet spectral region. The light is extracted through the GaN-on-sapphire transparent substrate, which is textured with an array of tiny pyramids to increase the extraction efficiency. In the structure illustrated, the active region comprises 5-nm GaN barriers and 2.5-nm $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{N}$ wells.

MicroLEDs

MicroLEDs have been developed for use in applications that are well-served by close-packed individually addressable μm-sized LEDs. These include high-resolution displays in smartphones, medical devices, TVs, outdoor signage, augmented-reality (AR) devices, and wearable devices that require compact and power-efficient displays. Fabricated using conventional LED materials such as InGaN, microLEDs are endowed with similar salutary features. They offer high brightness, high contrast, and superior stability. Filters, phosphors, and quantum dots provide pixels of different colors. MicroLEDs offer improved display quality, efficiency, and lifespan in comparison with conventional liquid-crystal displays (LCDs), and arguably with OLED displays, since the challenges of technological implementation and manufacturing are being met. They have also been promoted for use in microdisplays, video walls, TV walls, and cinema screens. Visible-spectrum light communications and optogenetics are other potential applications.

7.4 LEDS FOR ILLUMINATION

Crystalline III–V multiquantum-well LEDs rely on a well-established and mature manufacturing technology with a long history of commercialization. Highly reliable devices with long lifespans and low maintenance are ubiquitous. It is not an exaggeration to say that MQWLEDs have revolutionized lighting worldwide.

Colors

Under daylight (photopic) conditions, human vision is maximally sensitive at a wavelength of 555 nm, which falls in the yellowish-green region of the spectrum (Fig. 8.5-3). LEDs used for illumination are generally fabricated from AlInGaP and InGaN compound semiconductors. As discussed above, AlInGaP is a direct-bandgap semiconductor over the longer wavelengths of the visible spectrum and is ideal for fabricating red, orange, amber, and yellow LEDs. InGaN, also a direct-bandgap semiconductor, is ideally suited for the fabrication of green, blue, and violet LEDs. Multiquantum-well structures are invariably used since they provide superior performance.

The semiconductor material most commonly used for fabricating red, orange, and yellow LEDs for illumination applications is AlInGaP, while the material

generally used for fabricating green, blue, and violet LEDs is InGaN.

An example of the use of these materials in a traffic-signal indicator application is provided in Fig. 7.4-1. White light for illumination is often synthesized by the judicious combination of light of several colors, as will be elucidated in Fig. 9.1-2 and Sec. 11.3.

Figure 7.4-1 LED traffic signal based on III–V semiconductor materials.

Representative Parameters

Representative values for the key measures generally used to characterize the performance of LED sources for lighting applications (as detailed in Secs. 8.8 and 8.9) are displayed in the upper portion of Table 7.4-1. These data deliberately focus on small-area MQWLEDs, which enables their performance to be directly compared with the performance of small-area quantum-dot LEDs (QLEDs) as presented in the lower portion of Table 7.4-1. The table entries confirm that the blue, green, and red InGaN and AlInGaP MQWLEDs exhibit superior external quantum efficiency, energy conversion efficiency, optical power (radiant flux), luminous flux, and wall-plug luminous efficacy. Fortunately, the excellent performance of the MQWLEDs scales with device area and power, as will become clear in Chapter 10.

Drive Circuitry

An LED is usually driven by a current source, as schematically illustrated in Fig. 7.4-2(a). This is most simply implemented by means of a constant-voltage source in series with a resistor, as depicted in Fig. 7.4-2(b). The emitted light is readily modulated by varying the injected current. Analog and digital modulation are portrayed in Figs. 7.4-2(c) and 7.4-2(d), respectively. The performance of LED drivers can be enhanced by incorporating circuitry that regulates bias current, matches impedance, and provides nonlinear compensation to limit the maximum current. Fluctuations in the intensity of the emitted light may be stabilized by monitoring it with a photodetector and using the output as feedback to control the injected current.

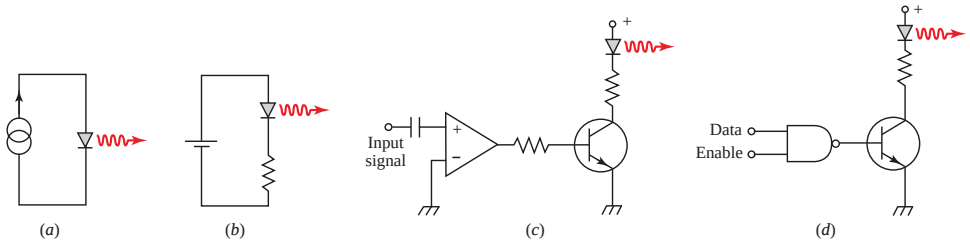


Figure 7.4-2 Schematics of circuits used as LED drivers: (a) an ideal DC current source; (b) a DC current source formed by a constant-voltage source in series with a resistor; (c) transistor control of the current injected into an LED to provide analog modulation of the emitted light; and (d) transistor switching of the current injected into an LED to provide digital modulation of the emitted light.

Table 7.4-1 ***UPPER TABLE (MQWLEDs)*:** Specifications for representative, discrete, small-area blue, green, and red III–V semiconductor multiquantum-well light-emitting diodes. (Data for Cree (Wolfspeed) XP-E2 LEDs. These devices are supplied in pockets on a tape containing 1000 LEDs; the tape is wound onto a reel to facilitate automated, high-throughput assembly.)

***LOWER TABLE (QLEDs)*:** Representative parameter values for discrete, small-area blue, green, and red quantum-dot light-emitting diodes comprising 40-nm-thick, single emissive layers of solution-processed II–VI colloidal CdSe/ZnSe core–shell quantum-dots. (Data adapted from H. Shen, Q. Gao, Y. Zhang, Y. Lin, Q. Lin, Z. Li, L. Chen, Z. Zeng, X. Li, Y. Jia, S. Wang, Z. Du, L. S. Li, and Z. Zhang, Visible Quantum Dot Light-Emitting Diodes with Simultaneous High Brightness and Efficiency, *Nature Photonics*, vol. 13, pp. 192–197, 2019.)

***UNITS*:** The successive columns display: peak emission wavelength λ_p (nm), photopic luminous efficiency function at peak wavelength $V(\lambda_p)$, device active area A (cm^2), voltage V (V), current density J (mA/cm^2), current i (mA), electrical power consumption P_{EL} (mW), external quantum efficiency (EQE), power-conversion efficiency (PCE), radiant flux P_0 (mW), radiance L (W/sr-m^2), luminance L_{V} (lm/sr-m^2), luminous flux P_{V} (lm), luminous efficacy of radiation (LER) (lm/W), wall-plug luminous efficacy (WPE) (lm/W), and wall-plug luminous efficiency (WPC).

DISCRETE III–V SEMICONDUCTOR MULTIQUANTUM-WELL LEDs

LEDS[a]	λ_p	$V(\lambda_p)$[b]	A	V	J	i[c,d]	P_{EL}^e	η_{EQE}^e	η_{PCE}^e	P_0^e	L^f	L_{V}^g	$P_{\mathrm{V}}^{i,j}$	η_{LER}^j	η_{WPE}^j	η_{WPC}^k
Blue	465	0.106	0.07	3.1	4930	350	1085	0.40	0.35	380	17040	1230000	27.5	72.4	25.3	0.037
Green	528	0.834	0.07	3.2	4930	350	1120	0.30	0.22	250	11210	6390000	140	570	125	0.183
Red	625	0.361	0.07	2.2	4930	350	770	0.40	0.36	280	12550	3100000	68.5	247	89.0	0.130

DISCRETE II–VI SEMICONDUCTOR QUANTUM-DOT LEDs

QLEDS[a]	λ_p	$V(\lambda_p)$[b]	A	V	J	i[d]	P_{EL}^e	η_{EQE}^e	η_{PCE}^e	P_0^e	L^f	$L_{\mathrm{V}}^{g,h}$	P_{V}^j	η_{LER}^j	η_{WPE}^j	η_{WPC}^k
Blue	483	0.195	0.04	3.8	105	4.20	16.0	0.081	0.060	0.96	75.9	10100	0.127	133	7.94	0.012
Green	532	0.878	0.04	3.8	53.0	2.12	8.06	0.229	0.137	1.10	87.5	52500	0.660	600	81.9	0.120
Red	602	0.666	0.04	3.1	18.0	0.72	2.23	0.216	0.165	0.37	29.2	13300	0.167	455	74.9	0.110

[a] Table entry values are rounded.

[b] Photopic luminous efficiency function at peak emission wavelength (Fig. 8.5-3).

[c] Increasing the drive current to 1 A results in an approximate doubling of the radiant flux and luminous flux, at the expense of reduced values of η_{EQE} stemming from efficiency droop.

[d] The current i (mA) is the product of the current density J (mA/cm^2) and the device active area A (cm^2).

[e] The electrical power consumption $P_{\mathrm{EL}} = iV$, EQE, PCE, and output optical power P_0 are interrelated via (7.1-12)–(7.1-14).

[f] The radiance is given by $L = P_0/\Omega A$, where Ω is the solid angle and A is the emission area (Table 8.8-1). The LEDs are assumed to radiate light with a Lambertian profile [Fig. 7.2-2(a)] so that $\Omega = \pi$.

[g] For a source that is (nearly) monochromatic, the radiance L is related to the luminance L_{V} provided in (8.8-4) by (8.9-1) and (8.9-2), which yield $L_{\mathrm{V}} = \eta_{\mathrm{LER}} L \approx 683\, V(\lambda_0)\, L$. The luminance is often used in place of the radiance for characterizing the performance of LEDs that operate in the visible.

[h] The highest values of the luminance attained for the blue, green, and red QLEDs were $L_{\mathrm{V}} = 62600$, 614000, and 356000 cd/m^2, respectively — these levels were associated with reduced values of η_{EQE} as a result of efficiency droop.

[i] Measured at a junction temperature of 25°C and a viewing angle $2\theta_{1/2} \approx 130°$.

[j] Radiometric and photometric units are linked via the luminous efficacy of radiation η_{LER} (lm/W), as specified in (8.9-1) and (8.9-5). For monochromatic light, $\eta_{\mathrm{LER}} \approx 683\, V(\lambda_0)$ lm/W, in accordance with (8.9-2). By definition, $\eta_{\mathrm{LER}} \leqslant \eta_{\mathrm{LER}}^{\mathrm{MAX}} = 683$ lm/W (the maximum value is attained for a monochromatic yellowish-green source at $\lambda_0 = 555$ nm); also, $\eta_{\mathrm{WPE}} \leqslant \eta_{\mathrm{WPE}}^{\mathrm{MAX}} = 683$ lm/W, as set forth in (8.9-8).

[k] The wall-plug luminous efficiency η_{WPC} is related to the wall-plug luminous efficacy η_{WPE} via $\eta_{\mathrm{WPC}} = \eta_{\mathrm{WPE}}/683$, as stated in (8.9-9). For (nearly) monochromatic LED light, the relationship can be written as $\eta_{\mathrm{WPC}} \approx \eta_{\mathrm{PCE}} V(\lambda_p)$, where λ_p is the peak wavelength, as specified in (8.9-10).

When it is desired to simply adjust the light intensity emitted by an LED, it is usually convenient to use pulse-width modulation (PWM) to regulate the applied current, as depicted in Fig. 7.4-3. This approach has the advantage that the average drive current supplied to the LED, which determines the intensity of the emitted light, is established by changing the duty cycle of the current, i.e., the proportion of time that the current is applied. Because the current level is fixed whenever it is present, this scheme avoids the intrinsic nonlinearity between the LED output intensity and the input current that is inherent in the device.

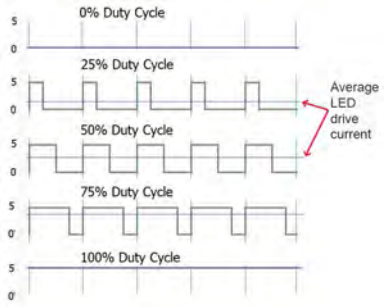


Figure 7.4-3 In pulse-width modulation (PWM), the average LED drive current (which determines the light intensity generated by the LED) is regulated by modifying the duty cycle of the applied current pulses. Since the current always has a fixed value when applied, the use of this approach circumvents the intrinsic nonlinearity between LED output intensity and input current.

Pulse-frequency modulation (PFM), a closely related technique, operates similarly. In this approach, the average current supplied to the LED is determined by varying the frequency of brief current pulses of fixed duration, rather than by varying the widths of the individual pulses as in PWM. In both cases, the LED will appear to be continuously illuminated as long as the rate at which the LED is turned on and off exceeds the human **flicker-fusion threshold**.

7.5 QUANTUM-DOT LIGHT-EMITTING DIODES (QLEDS)

As discussed in Sec. 5.8, colloidal quantum dots can be grown from a broad selection of semiconductor materials, organic compounds, and perovskites, all of which can serve as emissive media for light-emitting diodes. A number of salutary features of QDs render them suitable for fabricating electrically addressable, efficient **quantum-dot light-emitting diodes (QLEDs)**. The discussion in this section is devoted to a consideration of electroluminescent QLEDs fabricated from II–VI (chalcogenide) semiconductors (Fig. 5.3-3). The chalcogenides are the most commonly used materials for QLEDs since the synthesis of gallium-based III–V semiconductor quantum dots is a difficult enterprise using currently available solution-based growth techniques. QLEDs usually employ **core–shell quantum dots** rather than bare QDs since the presence of the shell mitigates surface defects and suppresses nonradiative recombination, thereby providing enhanced luminescence, tunability, and lifespan. We consider discrete (single-color), tandem (multicolor), and **white quantum-dot light-emitting diodes (WQLEDs)** in turn. We also provide a comparison of the performance of small-area discrete QLEDs and MQWLEDs in the current state of their technologies.

Discrete Single-Color Devices

An electrically pumped, monochromatic QLED, in its simplest conception, is sketched in Fig. 7.5-1. Electron and hole injection layers facilitate the injection of charge carriers into the emissive quantum-dot (QD) layer sandwiched between them, where the carriers

recombine and emit spontaneous photons. However, attaining high external quantum efficiency requires a more elaborate multilayer device structure to foster both a large internal quantum efficiency and a proper balance of electron and hole injection currents. Balanced injection obviates the accumulation of long-lived charged excitons that abet nonradiative **Auger recombination** (Fig. 5.5-2), a process whereby the energy released in the course of recombination is transferred to other charge carriers rather than to the creation of useful photons. Auger recombination, along with the **quantum-confined Stark effect (QCSE)**, contribute to **efficiency droop**, an undesirable effect in which device efficiency decreases with increasing current density.

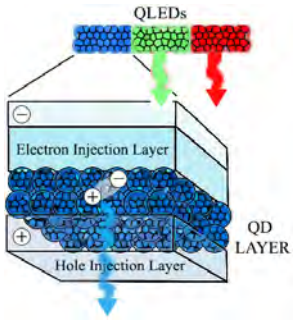


Figure 7.5-1 Structure of a simple, electrically pumped, II–VI, blue QLED. Injected carriers recombine in the QD layer, emitting spontaneous photons. Quantum dots of different sizes emit electroluminescence of different wavelengths, as exemplified in Fig. 5.8-1(*a*) for photoluminescence. Electrically addressable horizontal or serial vertical stacks of blue, green, and red QLED emitters generate light with an arbitrary mixture of these primaries. Tandem blue, green, and red quantum-dot layers designed to emit white light are known as white quantum-dot light-emitting diodes (WQLEDs).

Comparing QLEDs and MQWLEDs. Representative parameter values for small-area blue, green, and red discrete QLEDs fabricated from colloidal, II–VI CdSe/ZnSe, core–shell quantum dots are presented in the lower portion of Table 7.4-1. We briefly compare the performance of these devices with that of blue, green, and red discrete III-V MQWLEDs of comparable area, as reported in the upper portion of Table 7.4-1:

Advantages of QLEDs Over MQWLEDs.

- Superior tunability over a broader range of wavelengths.
- Narrower emission linewidths.
- Larger selection of saturated colors and hence color gamuts.
- Greater range of device design choices.
- More suitable for displays because of superior color purity.

Advantages of MQWLEDs Over QLEDs.

- Superior technology and more reliable manufacturing processes.
- Greater stability at high temperatures.
- Larger current densities and radiative recombination rates.
- Superior external, power-conversion, and wall-plug luminous efficiencies.
- Substantially larger radiant flux and luminous flux.
- More suitable for lighting applications.

Tandem Multicolor Devices

Horizontally or vertically stacked, electrically addressable structures can be used to simultaneously generate red, green, and blue light (Fig. 7.5-1). Of substantial significance is the **tandem configuration**, in which two or more **light-emitting units (LEUs)** are serially stacked and linked by **intermediate connection layers (ICLs)**, also known as **charge-generation layers**. Electrons and holes generated in these layers are efficiently injected into the adjacent LEUs, resulting in multiple photon emissions from a single electron–hole pair. This minimizes current flow, thereby averting device degradation

and Auger recombination (Fig. 5.5-2); and also allows the emitters in each LEU to be separately confined, which mitigates nonradiative energy transfer among them. Fully solution-processed tandem QLEDs with high efficiency have been fabricated, as have flexible tandem and individually addressable devices.

White Quantum-Dot Devices (WQLEDs)

Tandem QLEDs that generate white light, known as WQLEDs, comprise serial stacks of blue, green, and red quantum-dot layers that function independently and are optimized for the emission of white light. These devices usually operate on the basis of additive color mixing (Fig. 9.1-2 and Sec. 11.3). Single-layer WQLEDs in which the three colors of QDs are mixed are generally avoided because white light emission is attained only for a specific value of the drive current. Representative operating parameters for a solution-processed, small-area, tandem WQLED fabricated using alloyed 11-nm-diameter core–shell CdSe/ZnS red and green quantum dots, and 13-nm-diameter core–shell ZnCdS/ZnS blue quantum dots, are displayed in the top row of Table 7.6-1. The performance of these devices is laudable, but efforts directed toward their improvement continue. Analogous data for small-area white organic light-emitting diodes (WOLEDs) and small-area quantum-dot white perovskite light-emitting diodes (QPeWLEDs) are tabulated in the middle and bottom rows of Table 7.6-1, respectively. Performance comparisons for QLEDs withWOLEDs and QPeWLEDs are provided in Sec. 7.6.

7.6 ORGANIC LIGHT-EMITTING DIODES (OLEDS)

Organic light-emitting diodes can be fabricated from small organic molecules or from conjugated polymer chains (Sec. 5.9). Small-molecule **organic light-emitting diodes**, called **SMOLEDs** or simply **OLEDs**, are efficient generators of electroluminescence for the primary colors blue, green, and red. **Polymer light-emitting diodes**, called **PLEDs** or **P-OLEDs**, resemble OLEDs in their construction but usually have an n-type active region into which holes are injected by a p-type organic layer. **TADF-OLEDs**, which make use of **thermally activated delayed fluorescence (TADF)** emitters, offer an effective mechanism for attaining high efficiency. The energy gap between the singlet and triplet excited states (S_1 and T_1, respectively) in these materials is sufficiently small that temperature fluctuations can drive transitions to the singlet state. **White organic light-emitting diodes** are called **WOLEDs**.

Organic light-emitting diodes can be fabricated either by vacuum deposition or by solution processing; the latter includes screen, inkjet, and microcontact printing, as well as spin-coating and blade coating. While vapor deposition is useful for constructing complex multilayer device structures, it is a time-consuming process and is limited to small-area devices. Solution processing, in contrast, is simpler, faster, and less expensive, and can be used to construct devices that are not only large in area, but are also flexible and stretchable. Solution-processing technology is widely used for fabricating efficient SMOLEDs, PLEDs, and TADF-OLEDs.

Small-Molecule Devices (SMOLEDs)

A SMOLED is formed from two thin ($\approx$ 100-nm) organic semiconductor films juxtaposed to form an organic heterostructure. As portrayed in Fig. 7.6-1(a), this structure is sandwiched between two inorganic electrodes, an anode that injects holes and one or more cathodes that inject electrons. This is in contrast to carrier injection in inorganic LEDs, which relies on heavily doped p- and n-type crystalline materials together with strong forward bias.

The injected carriers are transported to the heterojunction (the active region), and

form bound excitons that recombine to generate spontaneous emission. Different heterostructure materials give rise to recombination radiation of different wavelengths, enabling a multicolor OLED to be constructed by patterning several heterostructures on a single substrate. These heterostructures can be fabricated side-by-side, in a *striped configuration*, to form a color-tunable *horizontal stack*, such as that sketched in Fig. 7.6-1(a). They can alternatively be fabricated one atop the other, to form a serial *vertical stack* with a blue emitter on top, a green emitter sandwiched in the middle, and a red emitter on bottom, as sketched in Fig. 7.6-1(b).

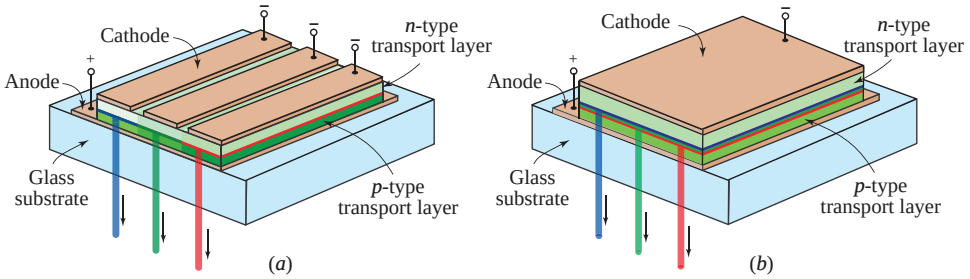


Figure 7.6-1 OLED structures fabricated in the form of: (a) a horizontal stack of blue, green, and red emitters, which is color tunable but requires patterning, and (b) a serial, vertical stack of emitters. Calcium (capped by Al to avoid degradation) and indium tin oxide (ITO, which is transparent) are often used as the cathode and anode materials, respectively. The exciton recombination radiation emitted at the organic heterojunctions exits through the transparent anode and glass substrate. Organic semiconductors such as electron-transporting aluminum tris(8-hydroxyquinoline) (Alq_3) and hole-transporting triphenyl diamine derivative (TPD) were initially used to fabricate OLEDs, but today's devices incorporate a dizzying array of organic materials. Luminescent dopants can be infused into the active regions to enhance the IQE and to create white light.

The energy levels of bound excitons in organic materials are similar to those of electrons in dye molecules, comprising both singlet (S) and triplet (T) states. In a singlet state, the electron spin is antiparallel to that of the remainder of the molecule, so that the total spin angular-momentum quantum number $S = 0$ and the spin multiplicity $2S + 1 = 1$. In a triplet state, in contrast, the electron spin is parallel to that of the remainder of the molecule, which results in $S = 1$ and $2S + 1 = 3$. Hence, the spin multiplicity of the triplet state is three times that of the singlet state.

Radiative transitions that take place between two states of the same multiplicity (i.e., $S \rightarrow S$ or $T \rightarrow T$) are spin-allowed, and the resulting luminescence process is called **fluorescence**. The luminescence resulting from spin-forbidden transitions (i.e., $S \rightarrow T$ or $T \rightarrow S$), in contrast, is called **phosphorescence**. The lifetimes of fluorescent transitions are usually far shorter than those of phosphorescent transitions (e.g., nsec vs. msec) as a result of the forbidden nature of the latter. Since the ground states of most organic compounds are singlets, radiative decay of singlet excitons is highly favored.

Nevertheless, triplet-exciton radiative recombination can be facilitated by infusing the active region of the device with special fluorophores that bind to the organic molecules or conjugated polymer chains comprising the heterostructure. This enables the triplet excitons to efficiently transfer their energy to the fluorophore while their spin angular momentum is concomitantly transferred to the organic molecule or polymer to which the exciton is bound. This in turn serves to increase the internal quantum efficiency of the device by a factor of four, since $2S + 1 = 3$ for the triplet state. This approach has the additional merit that it allows the color of the emitted light to be determined by the choice of fluorophore rather than by the exciting excitons.

Polymer Devices (PLEDs)

PLEDs are often comprised of light-emitting polymer materials (**LEPs**) that are derivatives of poly(p-phenylene vinylene) (PPV) and polyfluorene. The substitution of appropriate side chains on the polymer backbone can modify the color of the emitted light. Although solution-processed P-OLEDs can be readily printed and are less expensive to fabricate than OLEDs, they generally have lower efficiencies and shorter lifespans. The desirable features of small-molecule and large-molecule polymeric organic materials can be combined by using molecules called **phosphorescent dendrimers**. These are large molecular balls that contain a heavy-metal ion core (e.g., Ir(2-phenylpyridine)$_3$) that facilitates triplet-exciton radiative recombination via spin–orbit coupling by virtue of the layers of branching-ring structures bonded around it.

White Organic Devices (WOLEDs)

As discussed in Sec. 7.5, small-area tandem WQLEDs, which have been fabricated using solution-processed, inorganic II–VI quantum dots, offer excellent performance (top row of Table 7.6-1). To reiterate, tandem architectures involve serially stacking multiple light-emitting units (LEUs) and linking them via intermediate connection layers (ICLs); solution processing offers low manufacturing cost and large-scale production capabilities. White organic light-emitting diodes (WOLEDs), which operate on the basis of additive color mixing (Fig. 9.1-2 and Sec. 11.3), have been fabricated using serial stacks such as those displayed in Fig. 7.6-1(b). (WOLEDs are the unit cells of large-area white OLED light panels, such as discussed in Sec. 11.7.)

However, attaining high efficiency in this type of organic device is challenging. In particular, it turns out that solution-processed WOLED ICLs are hampered by suboptimal charge injection, along with issues related to surface wettability, orthogonal solubility, and chemical corrosion. In the current state of WOLED technology, an effective alternative approach is to make use of a single emissive layer consisting of blended light-emitting compounds (guest materials) in a matrix (host material). One way of implementing such a structure is via a **hyperfluorescence device**, in which the emissive layer contains a host material and two blended, interacting emitters. Singlet excitons converted from the first triplet excited state T_1 of a TADF sensitizer are captured and used by a traditional fluorescent emitter via a process known as **Förster resonance energy transfer (FRET)**. This mechanism plays a role analogous to that played by photoluminescence in phosphor-conversion LEDs (Sec. 10.2).

Comparing WOLEDs and WQLEDs. Representative operating parameters for a small-area, tandem **hyperfluorescence WOLED** that operates on the basis of a vacuum-evaporated, TADF blue emitter/sensitizer and yellow fluorescent emitter are displayed in the middle row of Table 7.6-1. Analogous data for a small-area, white quantum-dot light-emitting diode (WQLED) and a small-area white quantum-dot perovskite light-emitting diode (QPeWLED) are provided in the top and bottom rows of Table 7.6-1, respectively. Parameter-value comparisons in the following are based on the entries in Table 7.6-1.

Advantages of WOLEDs Over WQLEDs.

- Longer lifespans.
- Thinner structures.
- Flexible and bendable devices.
- Larger color gamut.
- Superior black expression.
- Wider viewing angle.
- Greater current density and wall-plug luminous efficiency.

Advantages of WQLEDs Over WOLEDs.

- Access to wavelength tunability via quantum-dot size adjustment.
- More established technology and more reliable manufacturing processes.
- Availability of high-efficiency tandem structures.
- More facile solution-processing.
- Fewer functional layers.
- Superior photostability.
- Superior color purity.
- Larger current density, external efficiency, luminance, and luminous flux.
- More suitable for lighting applications.

QWOLEDs. **Quantum-dot white organic light-emitting diodes (QWOLEDS)** have been grown in the laboratory. They possess many of the advantages of WOLEDs, and have the additional merits of being more amenable to solution processing and offering superior color purity and brightness. On the other hand, they suffer from increased temperature sensitivity are require more complex and costly manufacturing procedures. They have not yet been called upon for the fabrication of photonic devices.

7.7 PEROVSKITE LIGHT-EMITTING DIODES (PELEDS)

Perovskites, which were introduced in Sec. 5.9, are versatile photonic materials whose composition can range from organic to hybrid organic-inorganic to fully inorganic. Like chalcogenide quantum dots (Secs. 5.8, 6.6, and 7.5) and organic semiconductors (Secs. 5.9 and 7.6), perovskites can often be inexpensively fabricated by making use of solution-processing methods such as spin coating.

Some perovskites exhibit high carrier mobilities, long carrier lifetimes, and large absorption coefficients, endowing them with superior charge-transport properties and enables them to serve as highly efficient sources of electroluminescence and photoluminescence. As with MQWLEDs, QLEDs, and OLEDs, variations on the theme of **perovskite light-emitting diodes (PeLEDs)** abound. And like their antecedents, PeLEDs are compositionally tunable: the emission wavelength depends on the particular perovskite recipe. They can generate monochromatic light in the visible, as well as in the near-infrared and near-ultraviolet. PeLEDs should not be conflated with polymer light-emitting diodes known as PLEDs.

Of particular importance for the fabrication of PeLEDs are **metal-halide perovskites (MHPs)**, materials that possess a compelling combination of chemical robustness and high-quality optical properties (Sec. 5.9). PeLEDs can consist of a single perovskite or can make use of a mixture of different perovskites. The emissive region can be fabricated in the form of a thin polycrystalline film, which creates a **polycrystalline-film perovskite light emitting diode (PPeLED)**, or in the form of an assembly of perovskite quantum dots (nanocrystals), which creates a **quantum-dot perovskite light-emitting diode (QPeLED)**. **White perovskite light-emitting diodes (PeWLEDs)** are increasingly being used to generate white light, as are are **quantum-dot white perovskite light-emitting diodes (QPeWLEDs)**.

Metal-Halide (MHP) Devices

The physical and chemical properties of fully inorganic, direct-bandgap metal-halide perovskites (MHPs), such as the cesium-lead halides $CsPbI_3$, $CsPbBr_3$, and $CsPbCl_3$, make them particularly suitable for use in LEDs and other photonic components. MHPs that make use of lead, which operate in the visible region, are also called **lead-halide**

Table 7.6-1 ***TOP ROW (WQLED)*:** Representative parameter values for a small-area, tandem, white quantum-dot light-emitting diode with individual 20-nm-thick, solution-processed, ZnCdS/ZnS blue-QD, CdSe/ZnS green-QD, and CdSe/ZnS red-QD emissive layers, in a structure comprising ITO/PEDOT-PSS[a]/TFB-PVK[b]/blue-QDs/ZnO-PMA[c]/TFB/green-QDs/ ZnO-PMA/TFB/red-QDs/ZnO/Ag.(Data adapted from C. Jiang, J. Zou, Y. Liu, C. Song, Z. He, Z. Zhong, J. Wang, H. Yip, J. Peng, and Y. Cao, Fully Solution-Processed Tandem White Quantum-Dot Light-Emitting Diode with an External Quantum Efficiency Exceeding 25%, *ACS Nano*, vol. 12, pp. 6040-6049, 2018.)

***MIDDLE ROW (WOLED)*:** Representative parameter values for a small-area, white organic light-emitting diode with a 100-nm-thick, vacuum-evaporated, single emissive layer comprising a host matrix (DBFDPO)[d], a TADF blue emitter/sensitizer (ptBCzPO$_2$TPTZ)[e], and a yellow fluorescent emitter (TBRb)[f,g]. (Data adapted from D. Ding, Z. Wang, C. Duan, C. Han, J. Zhang, S. Chen, Y. Wei, and H. Xu, White Fluorescent Organic Light Emitting Diodes with 100% Power Conversion, *Research*, vol. 2022, no. 0009, DOI:10.34133/research.0009, 2022.)

***BOTTOM ROW (QPeWLED)*:** Representative parameter values for a small-area, white perovskite light-emitting diode with a 15-nm-thick, single emissive layer comprising solution-processed $CsPbI_3$ perovskite quantum dots of mixed α (cubic) and δ (orthorhombic) phases. (Data adapted from J. Chen, J. Wang, X. Xu, J. Li, J. Song, S. Lan, S. Liu, B. Cai, B. Han, J. T. Precht, D. Ginger, and H. Zeng, Efficient and Bright White Light-Emitting Diodes Based on Single-Layer Heterophase Halide Perovskites, *Nature Photonics*, vol. 15, pp. 238–244, 2021.)

***UNITS*:** The successive columns display: device active area A (cm^2), forward voltage V (V), current density J (mA/cm^2), current i (mA), electrical power consumption P_{EL} (mW), external quantum efficiency (EQE), current luminous efficacy (CLE) (cd/A), luminance L_{V} (lm/sr-m^2), luminous flux P_{V} (lm), wall-plug luminous efficacy (WPE) (lm/W), wall-plug luminous efficiency (WPC), chromaticity coordinates x and y, and correlated color temperature T_c (K).

WHITE QLED, WHITE OLED, and WHITE QUANTUM-DOT PeLED

SOURCE[h]	A	V	J	i^i	P_{EL}	η_{EQE}	η_{CLE}	L_{V}^j	P_{V}^k	η_{WPE}^l	η_{WPC}^m	x^n	y^n	T_c^o
WQLED[p]	0.04	15.2	20.0	0.800	12.2	19.5	45.6	9400	0.118	9.71	0.014	0.42	0.42	3420
WOLED[q]	0.09	4.05	1.16	0.104	0.42	0.269	86.3	1000	0.028	65.3	0.096	0.30	0.42	6225
QPeWLED[r]	0.04	4.55	8.30	0.332	1.51	0.065	12.2	1015	0.013	8.53	0.012	0.38	0.41	4210

[a] PEDOT-PSS: poly(3,4-ethylenedioxythiophene):poly(styrenesulfonate).

[b] TFB-PVK: poly(9,9-dioctylfluorene- *co-N*-(4-butylphenyl)diphenylamine)/poly(9-vinylcarbazole).

[c] ZnO-PMA: ZnO nanoparticle and polyoxometalate phosphomolybdic acid intermediate connection bilayer.

[d] DBFDPO: 4,6-bis(diphenylphosphoryl)dibenzofuran.

[e] TADF: 9-(4-(4,6-Bis(4-(diphenylphosphoryl)phenyl)-1,3,5-triazin-2-yl)phenyl)-3,6-di-*tert*-butyl-carbazole.

[f] TBRb: 2,8-di-*tert*-butyl-5,11-bis(4-*tert*-butylphenyl)-6,12-diphenyltetracene.

[g] The components were combined in the following proportions: DBFDPO:40%-ptBCzPO$_2$TPTZ:0.1%-TBRb.

[h] Table entry values are rounded.

[i] The current i (mA) is the product of the current density J (mA/cm^2) and the device active area A (cm^2).

[j] The luminance L_{V} (cd/m^2) is the product of η_{CLE} (cd/A) and J (A/m^2), as provided in (8.9-11).

[k] The luminous flux P_{V} (lm) is the product of the luminance L_{V} (lm/m^2-sr), the device active area A (m^2), and the radiation solid angle Ω. For a Lambertian radiator [Fig. 7.2-2(a)], we have $\Omega = \pi$ so that $P_{\text{V}} = \pi A L_{\text{V}}$.

[l] The wall-plug luminous efficacy η_{WPE} is defined as $P_{\text{V}}/P_{\text{EL}}$, as specified in (8.9-4). Equation (8.9-8) provides that $\eta_{\text{WPE}} \leqslant \eta_{\text{WPE}}^{\text{MAX}} = 683$ lm/W (the maximum value is attained for a monochromatic source at $\lambda_0 = 555$ nm).

[m] The wall-plug luminous efficiency η_{WPC} is related to the η_{WPE} via $\eta_{\text{WPC}} = \eta_{\text{WPE}}/683$, in accordance with (8.9-9).

[n] The chromaticity coordinates x and y, which are measures of perceived color, are defined in Sec. 9.6.

[o] The correlated color temperature T_c, a measure of the color of a source of light, is defined in Sec. 9.8.

[p] Maximum parameter values observed for devices of this type: $\eta_{\text{EQE}}^{\text{MAX}} = 0.274$, $\eta_{\text{CLE}}^{\text{MAX}} = 60.7$ cd/A, and $L_{\text{V}}^{\text{MAX}} = 210000$ cd/m^2. Current maximum value for solution-processed tandem QLEDs: $\eta_{\text{CLE}}^{\text{MAX}} \approx 183$ cd/A.

[q] Maximum parameter values observed for devices of this type with supplementary outcoupling enhancement: $\eta_{\text{EQE}}^{\text{MAX}} = 0.307$, $\eta_{\text{CLE}}^{\text{MAX}} = 96.1$ cd/A, $L_{\text{V}}^{\text{MAX}} = 40\,105$ cd/m^2, and $\eta_{\text{WPE}}^{\text{MAX}} = 120$ lm/W. Current maximum value for solution-processed tandem OLEDs: $\eta_{\text{CLE}}^{\text{MAX}} \approx 93$ cd/A.

[r] Maximum parameter values observed for devices of this type: $\eta_{\text{EQE}}^{\text{MAX}} = 0.065$, $\eta_{\text{CLE}}^{\text{MAX}} = 12.2$ cd/A, and $L_{\text{V}}^{\text{MAX}} = 12\,200$ cd/m^2.

perovskites (LHPs). The bandgap wavelength of these materials can be compositionally tuned to cover the entire visible region.

Polycrystalline-Film Devices (PPeLEDs)

PeLEDs often have an emissive layer prepared in the form of a thin polycrystalline films, enabling the carriers to be spatially localized within the individual grains, which increases the efficiency of radiative recombination. Alternatively, they can be constructed in the form of 2) perovskite grains embedded in a polymer matrix or as perovskite quantum dots. As discussed in connection with (5.5-10) and (10.2-1), the internal electroluminescence quantum efficiency (IQE) and the photoluminescence quantum yield PLQY are maximized by making the radiative decay rate as large as possible. The individual grains of the thin polycrystalline film spatially localize the carriers in a common location, which enhances the radiative recombination and increases the quantum efficiency in turn. Embedding perovskite grains in a polymer matrix achieves the same goal.

The structures and charge-transport characteristics of PeLEDs share a number of features in common with those of OLEDs: the electrodes sandwiching the emissive region deliver the electrons and holes that generate recombination radiation in this thin region (Fig. 7.6-1).

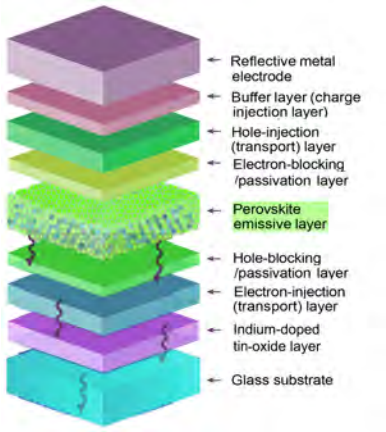


Figure 7.7-1 Schematic of a generic perovskite LED. The device consists of a collection of layers whose functions are indicated. For a PPeLED, the perovskite emissive layer (highlighted) consists of a thin polycrystalline film that contains one or more perovskite compounds, whereas for a QPeLED it comprises an assembly of perovskite quantum dots (nanocrystals). The photons generated in the emissive layer result from radiative recombination of the injected electron–hole pairs. The emissive layer is surrounded by a pair of passivation layers that mitigate nonradiative recombination fostered by surface defects. The passivation layers are in turn surrounded by carrier-injection (transport) layers, which make contact with the conductive electrodes that provide forward bias to the device.

EXAMPLE 7.7-1. ***External Quantum Efficiency of a $FAPbI_3$ Perovskite LED.*** The external quantum efficiency of PeLEDs can be substantially enhanced by adding various multifunctional molecules to the material. For example, the performance of an α-phase formamidinium lead triiodide ($FAPbI_3$) perovskite LED, which has a bandgap $E_g = 1.5$ eV and operates at 800 nm in the NIR, is substantially improved by the addition of 2-(4-(methylsulfonyl)phenyl)ethylamine (MSPE) (see Y. Sun, L. Ge, L. Dai, C. Cho, J. F. Orri, K. Ji, S. J. Zelewski, Y. Liu, A. J. Mirabelli, Y. Zhang, J.-Y. Huang, Y. Wang, K. Gong, M. C. Lai, L. Zhang, D. Yang, J. Lin, E. M. Tennyson, C. Ducati, S. D. Stranks, L.-S. Cui, and N. C. Greenham, Bright and Stable Perovskite Light-Emitting Diodes in the Near-Infrared Range, *Nature*, vol. 615, pp. 830–835, 2023). The additive mitigates nonradiative pathways by removing nonradiative dark regions in the perovskite films while simultaneously suppressing the quenching of perovskite luminescence at the interface with charge-transport layers. The $FAPbI_3$ perovskites were prepared by spin-coating from FAI, PbI_2, and MSPE precursors. The external quantum efficiency depends strongly on the molar fraction of MSPE relative to PbI_2, and attains its maximum value of ($\approx 20\%$) for a molar fraction MSPE/PbI_2 that is ≈ 0.5.

Quantum-Dot Devices (QPeLEDs)

However, the defects at the various surfaces of these perovskites films, such as the grain boundaries, limit the efficiency of light emission. These effects can often be mitigated by making use of core–shell perovskite quantum dots. Like conventional chalcogenide quantum dots, core–shell structures have a high tolerance for defects and can be synthesized in monodisperse form from inexpensive commercial precursors using solution-based methods. As a result of compositional tuning and quantum-size effects, MHP quantum dots can emit over the full visible spectrum as well as in the near infrared and near ultraviolet.

The colors of the photoluminescence elicited from $CsPbX_3$ colloidal quantum dots of cubic shape and cubic crystal structure are illustrated in Fig. 5.8-1(b) for different compositions and dot sizes [X = I (red), Br (green), and Cl (violet)]. Mixed MHP QDs, such as I/Br and Br/Cl, emit in the green and blue, respectively. All of these materials are stable in the α (cubic) phase although $CsPbI_3$ can also easily be prepared in the δ (orthorhombic) phase. The photoluminescence from $CsPbX_3$ quantum dots is characterized by narrow emission linewidth ($\Delta\lambda_{\text{FWHM}} \approx$ 10–45 nm), high photoluminescence quantum yield (PLQY $\rightarrow$ 100%) (Sec. 10.2), high carrier mobility, but short radiative lifetime ($\tau_{\text{r}} \approx$ 1–30 ns). The performance and lifespan of these materials are currently limited by sensitivity to moisture, oxygen, and light, which degrade performance.

As with chalcogenide quantum dots (Sec. 7.5), perovskite quantum dots are useful for fabricating LEDs. Emissive regions employing MHP quantum dots, which possess excellent color purity and high PLQY, result in LEDs with high external quantum efficiency that can be further enhanced by molecular additives. MHP quantum dots can also serve as photoluminescent media in a phosphor-conversion configuration or as matrices that operate as high-mobility charge-transport materials. Thin sheets of these materials can be inexpensively fabricated using solution processing and printed on flexible substrates such as plastic, generally at room temperature. This is in distinction to III–V colloidal quantum dots, which are difficult to grow, and to organic quantum dots, which are seldom used for device fabrication for the reasons explained in Sec. 5.9.

As discussed in Sec. 7.6, white organic light-emitting diodes (WOLEDs) can be fabricated in many different configurations. Metal-halide perovskites (MHPs) can be similarly configured in ways that are attractive for use as **white perovskite light-emitting diodes (PeWLEDs)**. Moreover, **quantum-dot white perovskite light-emitting diodes (QPeWLEDs)** are increasingly available.

7.8 LASER DIODES AND LIGHT-EMITTING DIODES

The compounds discussed in Sec. 7.3 also used not only for fabricating light-emitting diodes, but also for laser diodes, quantum-confined lasers, microcavity lasers, and nanocavity lasers. As illustrated in Fig. 7.0-1, light-emitting diodes (LEDs) and laser diodes (LDs) both have, at their heart, a forward-biased p–n junction fabricated from a direct-bandgap semiconductor. The essential distinction is that light emitted from an LED is spontaneous emission, whereas light from an LD is stimulated emission. The transition from a partially coherent beam generated by a light-emitting diode to a coherent beam generated by a laser diode as the drive current increases, as illustrated in Fig. 7.0-1, is marked by a threshold current beyond which the light intensity increases sharply, as well as a pronounced narrowing of both the spectrum and divergence angle above threshold. We discuss III–V semiconductor laser diodes, metal-halide perovskite laser diodes, and silicon photonics in turn.

Traditional III–V Semiconductor Laser Diodes

The earliest laser diodes comprised single, heavily doped p–n junctions of GaAs and GaAsP, which emitted in the near infrared and red, respectively. Four independent groups, at GE, IBM, and MIT Lincoln Laboratory, reported the operation of these devices in November and December 1962, just months after the initial report of the GaAs LED by Keyes & Quist (p. 198). A GaAs p–n junction diode fabricated at MIT Lincoln Laboratory in 1963, which functioned as a laser diode when cooled and as an LED at room temperature, is depicted in Fig. 7.8-1.

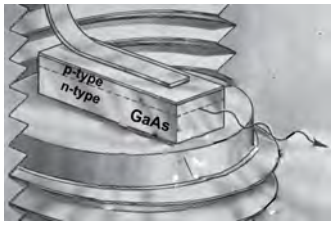


Figure 7.8-1 Sketch of a GaAs laser diode fabricated at MIT Lincoln Laboratory in 1963. The width and length of the device were 200 μm and 1 mm, respectively. When operated in pulsed mode at 77 K, this device emitted coherent light at 845 nm. (Courtesy MIT Lincoln Laboratory; Image adapted from M. C. Teich, "Two Quantum Photoemission and dc Photomixing in Sodium," Ph.D. Dissertation, Cornell University, February 1966, Fig. 3, p. 16).

For both LEDs and LDs, the source of energy is the electric current injected into the junction. It is far easier to elicit partially coherent light from an LED than coherent light from an LD for three principal reasons: 1) Operation as a laser diode requires an injected current that generates a density of electrons and holes in the junction region that is large enough to create a population inversion, thereby rendering stimulated emission more prevalent than absorption and providing gain; 2) Operation as an LD requires a suitable feedback mechanism that can initiate and sustain laser action. In the simplest case, this can be implemented by cleaving the semiconductor material along its crystal planes, which results in a sharp refractive-index discontinuity between the crystal and the surrounding air and thereby gives rise to substantial reflection. The semiconductor crystal then simultaneously acts as a gain medium and as a Fabry–Perot optical resonator, as illustrated in Fig. 7.8-2; 3) Operation as an LD laser requires a population inversion, which is facilitated by low temperatures.

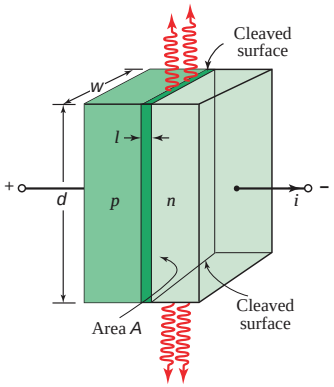


Figure 7.8-2 In its simplest configuration, a laser diode is a forward-biased, heavily doped p–n junction in which two surfaces perpendicular to the plane of the junction act as reflectors in a Fabry–Perot resonator. These surfaces can be created by cleaving the structure along its crystal planes to ensure that they are parallel. The other two surfaces perpendicular to the plane of the junction are then roughened to eliminate feedback in that direction.

From the inception of their development in the early 1960s, LEDs were able to operate at room temperature in a continuous-wave (CW) mode. LDs, on the other hand, were able to operate only in pulsed mode and at cryogenic temperatures. Laser diodes that were operated at temperatures that were too high, or were driven by current pulses that were too long, behaved as LEDs. When the reduction of temperature in these

devices was sufficient, the result was substantial spectral and spatial narrowing as the partially coherent LED light moved toward coherent LD light.

Today's semiconductor lasers operate CW at room temperature. They also assume a bewildering variety of forms. They function at wavelengths that stretch from the mid ultraviolet to the far infrared — and at output powers that range from nW (for nanolasers) to W (for individual laser diodes) to kW (for banks of laser diodes). They come in all shapes and sizes, including quantum-confined devices and compact microcavity and nanocavity lasers.

Laser diodes find extensive use in long-haul optical fiber communication systems, where they can be readily modulated by controlling the injected current. They are also used in high-density optical data-storage systems such as DVD players, and in scanning, reading, and high-resolution color-printing systems. LDs are also employed in lidars and in directional lighting applications, such as automotive headlights. Banks of laser diodes are used to optically pump optical fiber amplifiers and solid-state lasers, thereby converting the relatively broadband, multimode laser-diode light into the narrowband, single-mode light emitted by diode-pumped solid-state lasers. There is a school of thought that promulgates expanding the use of laser diodes in place of light-emitting diodes for specialized illumination applications that make use of their directionality and the absence of efficiency droop above laser threshold.

The advent of quantum-confined semiconductor lasers such as multiquantum-well, multiquantum-dot, and quantum cascade lasers, together with compact lasers such as vertical-cavity, microdisk, photonic-crystal, and nanolasers, has greatly facilitated the integration of lasers with other optical components in compact configurations, which in turn has opened the door to myriad new uses.

Silicon Photonics Light Sources

Silicon has long been the leading materials platform for **integrated electronics**, for a whole host of reasons: it is 1) abundant and inexpensive; 2) readily grown in pure form and in bulk; 3) easy to dope, oxidize, and manipulate; 4) stable at high temperatures; and 5) compatible with **CMOS (omplementary metal-oxide-semiconductor) technology**. Its ubiquity, availability, and properties have also made it an attractive platform for **integrated photonics**. The high refractive-index contrast of silicon and its oxides allows strong optical confinement in a compact volume. This, together with its transparency in the 1.3–1.6-μm telecommunications band, and its CMOS compatibility, promotes its use for devices used in telecommunications.

A notable exception to the adaptability of Si is centered on its use as an active medium for LEDs and LDs. The development of Si-based light sources has been hampered by its indirect bandgap, which restricts its ability to generate light efficiently via interband transitions (Fig. 6.2-4). Over the years, extensive efforts have been devoted to surmounting this roadblock, either by mitigating the indirect nature of silicon's bandgap or by avoiding it altogether. Early efforts directed toward increasing the efficiency of light emission involved the use of alternatives to its crystalline form, such as porous silicon (in which nanopores pervade the diamond structure); silicon nanocrystals, superlattices, and quantum dots (Example 6.6-2); and Er^{3+}-doped silicon-based hosts and superlattices. To date, none of these approaches has been particularly successful, however. A more fruitful approach has been to co-opt light-emitting interactions in silicon other than those associated with interband transitions. In particular, the silicon Raman laser relies on stimulated Raman scattering and is thus indifferent to the nature of the bandgap. Still, Raman devices require optical rather than electrical pumping, which reduces their appeal for many applications. Yet, silicon Raman lasers have been successfully integrated with direct-bandgap emitters such as InP that serve as an optical pump.

Fortunately, substantial progress has been made in recent years in implementing

silicon-based on-chip light sources for use in **photonic integrated circuits (PICs)**. Three approaches are currently in use, each with its own limitations and merits:

1. Flip-chip integration (direct-mounting integration) of III–V laser diodes into a separately fabricated silicon platform, often with optical butt coupling. This approach, which makes use of solder bumps, requires sub-micrometer-scale alignment precision and is not scalable to large wafer volumes or complex laser designs, but it is straightforward.

2. Heterogeneous integration (hybrid approach) of III–V lasers into prepatterned silicon circuits, typically via wafer bonding and with optical evanescent coupling to evade lattice-matching limitations. This approach is incompatible with the clean CMOS-foundry environment. However, it accommodates a whole host of materials and can also relegate photon storage to the undoped silicon platform (with its low loss and high Q) via hybrid modes, thereby facilitating the fabrication of narrow-linewidth, dense-comb, and mode-locked lasers.

3. Direct heteroepitaxial growth of III–V lasers on Si substrates using intermediate buffer layers to minimize dislocations in the light-emitting region. This approach is encumbered with the large lattice-constant and thermal mismatches between Si and III–V materials, which result in dislocations that reduce efficiency by acting as nonradiative recombination centers. However, this can be largely counterbalanced by employing quantum-dot, rather than quantum-well emitters, since: 1) quantum dots are less affected by the threading dislocations initiated by lattice and thermal mismatches, and 2) quantum dots enjoy substantially reduced sensitivity to temperature changes.

On balance, direct heteroepitaxy appears to be the most attractive alternative for large-scale, low-cost, fabrication of silicon-based on-chip light sources.

It is worth noting that group-IV photonics also offers a route to the development of on-chip light sources via combinations of indirect-bandgap semiconductors such as Si, Ge, Sn, and C. Germanium-based structures are leading the way, although considerable challenges remain. Interestingly, the use of such materials is not new: the first LED, dating to 1907, was a forward-biased SiC Schottky diode (p. 169).

BIBLIOGRAPHY

Light-Emitting Diodes (LEDs)

See also the bibliographies in Chapters 5, 6, 10, and 11.

C. J. Praharaj, *Group III–Nitride Semiconductor Optoelectronics*, Wiley, 2024.

E. F. Schubert, *Light-Emitting Diodes*, Google Books, 4th ed. 2023.

V. Kumar, V. Sharma, and H. C. Swart, eds., *Advanced Materials for Solid State Lighting*, Springer, 2023.

J. Li, J. Wang, X. Yi, Z. Liu, T. Wei, J. Yan, and B. Xue, *III–Nitrides Light Emitting Diodes: Technology and Applications*, Springer, 2020.

G. B. Nair and S. J. Dhoble, *The Fundamentals and Applications of Light-Emitting Diodes*, Woodhead/Elsevier, 2020.

J. Li and G. Q. Zhang, eds., *Light-Emitting Diodes: Materials, Processes, Devices and Applications*, Springer, 2019.

T.-Y. Seong, J. Han, H. Amano, and H. Morkoç, eds., *III–Nitride Based Light Emitting Diodes and Applications*, Springer, 2nd ed. 2017.

R. Karlicek, C.-C. Sun, G. Zissis, and R. Ma, eds., *Handbook of Advanced Lighting Technology*, Springer, 2017.

J.-J. Huang, H.-C. Kuo, and S.-C. Shen, eds., *Nitride Semiconductor Light-Emitting Diodes (LEDs): Materials, Technologies and Applications*, Woodhead/Elsevier, 2014.

J. Iveland, L. Martinelli, J. Peretti, J. S. Speck, and C. Weisbuch, Direct Measurement of Auger Electrons Emitted from a Semiconductor Light-Emitting Diode under Electrical Injection: Identification of the Dominant Mechanism for Efficiency Droop, *Physical Review Letters*, vol. 110, 177406, 2013.

H. Morkoç, *Handbook of Nitride Semiconductors and Devices*, Volume III: *GaN-Based Optical and Electronic Devices*, Wiley–VCH, 2008.

R. K. Willardson and E. R. Weber, eds., *Semiconductors and Semimetals*, Volume 48, *High-Brightness Light Emitting Diodes*, G. B. Stringfellow and M. G. Craford, eds., Academic Press, 1997.

F. A. Kish, F. M. Steranka, D. C. DeFevere, D. A. Vanderwater, K. G. Park, C. P. Kuo, T. D. Osentowski, M. J. Peanasky, J. G. Yu, R. M. Fletcher, D. A. Steigerwald, M. G. Craford, and V. M. Robbins, Very High-Efficiency Semiconductor Wafer-Bonded Transparent-Substrate $(Al_xGa_{1-x})_{0.5}In_{0.5}P/GaP$ Light-Emitting Diodes, *Applied Physics Letters*, vol. 64, pp. 2839–2841, 1994.

V. L. Colvin, M. C. Schlamp, and A. P. Alivisatos, Light-Emitting Diodes Made from Cadmium Selenide Nanocrystals and a Semiconducting Polymer, *Nature*, vol. 370, pp. 354–357, 1994.

K. Kobayashi, S. Kawata, A. Gomyo, I. Hino, and T. Suzuki, Room-Temperature CW Operation of AlInGaP Double-Heterostructure Lasers, *Electronics Letters*, vol. 21, pp. 931–932, 1985.

M. G. Craford, R. W. Shaw, A. H. Herzog, and W. O. Groves, Radiative Recombination Mechanisms in GaAsP Diodes With and Without Nitrogen Doping, *Journal of Applied Physics*, vol. 43, pp. 4075–4083, 1972.

MicroLEDs

J.-H. Ahn and J.-H. Kim, eds., *Micro Light Emitting Diode: Fabrication and Devices*, Springer, 2021.

H. Jiang and J. Lin, eds., *Micro LEDs*, in *Semiconductors and Semimetals*, vol. 106, Academic/Elsevier, 2021.

P. J. Parbrook, B. Corbett, J. Han, T.-Y. Seong, and H. Amano, Micro-Light Emitting Diode: From Chips to Applications, *Laser & Photonics Reviews*, vol. 15, 2000133, 2021.

Z. Chen, S. Yan, and C. Danesh, MicroLED Technologies and Applications: Characteristics, Fabrication, Progress, and Challenges, *Journal of Physics D: Applied Physics*, vol. 54, 123001, 2021.

V. Venugopal, The Pricey Promise of MicroLEDs, *Photonics Focus*, vol. 2, no. 6, pp. 22–27, 2021.

Quantum-Dot Light-Emitting Diodes (QLEDs)

H. Meng, *Colloidal Quantum Dot Light Emitting Diodes: Materials and Devices*, Wiley–VCH, 2024.

E. Jang and H. Jang, Review: Quantum Dot Light-Emitting Diodes, *Chemical Reviews*, vol. 123, pp. 4663–4692, 2023.

D. Tian, H. Ma, G. Huang, M. Gao, F. Cai, Y. Fang, C. Li, X. Jiang, A. Wang, S. Wang, and Z. Du, A Review on Quantum Dot Light-Emitting Diodes: From Materials to Applications, *Advanced Optical Materials*, vol. 11, p. 2201965, 2023.

Q. Su, H. Zhang, and S. Chen, Flexible and Tandem Quantum-Dot Light-Emitting Diodes with Individually Addressable Red/Green/Blue Emission, *Flexible Electronics*, vol. 5, 8, 2021.

H. Shen, Q. Gao, Y. Zhang, Y. Lin, Q. Lin, Z. Li, L. Chen, Z. Zeng, X. Li, Y. Jia, S. Wang, Z. Du, L. S. Li, and Z. Zhang, Visible Quantum Dot Light-Emitting Diodes with Simultaneous High Brightness and Efficiency, *Nature Photonics*, vol. 13, pp. 192–197, 2019.

C. Jiang, J. Zou, Y. Liu, C. Song, Z. He, Z. Zhong, J. Wang, H.-L. Yip, J. Peng, and Y. Cao, Fully Solution-Processed Tandem White Quantum-Dot Light-Emitting Diode with an External Quantum Efficiency Exceeding 25%, *ACS Nano*, vol. 12, pp. 6040–6049, 2018.

Y. Shirasaki, G. J. Supran, M. G. Bawendi, and V. Bulović, Emergence of Colloidal Quantum-Dot Light-Emitting Technologies, *Nature Photonics*, vol. 7, pp. 13–23, 2013.

V. L. Colvin, M. C. Schlamp, and A. P. Alivisatos, Light-Emitting Diodes Made from Cadmium Selenide Nanocrystals and a Semiconducting Polymer, *Nature*, vol. 370, pp. 354–357, 1994.

Organic Light-Emitting Diodes (OLEDs)

G. Xie, ed., *Solution-Processed Organic Light-Emitting Diodes*, Woodhead/Elsevier, 2024.

S.-G. Meng, X.-Z. Zhu, D.-Y. Zhou, and L.-S. Liao, Recent Progresses in Solution-Processed Tandem Organic and Quantum Dots Light-Emitting Diodes, *Molecules*, DOI:10.3390/molecules28010134, vol. 28, p. 134, 2023.

S. Huo and Y. Li, *Phosphorescent Materials*, American Chemical Society, 2023.

D. Ding, Z. Wang, C. Duan, C. Han, J. Zhang, S. Chen, Y. Wei, and H. Xu, White Fluorescent Organic Light Emitting Diodes with 100% Power Conversion, *Research*, vol. 2022, no. 0009, DOI:10.34133/research.0009, 2022.

Y. Miao and M. Yin, Recent Progress on Organic Light-Emitting Diodes with Phosphorescent Ultrathin ($<$ 1 nm) Light-Emitting Layers, *iScience*, vol. 25, 103804, DOI:10.1016/j.isci.2022.103804, 2022.

L. Duan, ed., *Thermally Activated Delayed Fluorescence Organic Light-Emitting Diodes (TADF-OLEDs)*, Woodhead/Elsevier, 2022.

M. Koden, *Flexible OLEDs: Fundamental and Novel Practical Technologies*, Springer, 2022.

S.-J. Zou, Y. Shen, F.-M. Xie, J.-D. Chen, Y.-Q. Li, and J.-X. Tang, Recent Advances in Organic Light-Emitting Diodes: Toward Smart Lighting and Displays, *Materials Chemistry Frontiers*, vol. 4, pp. 788–820, 2020.

H. Yersin, ed., *Highly Efficient OLEDs: Materials Based on Thermally Activated Delayed Fluorescence*, Wiley–VCH, 2018.

N. T. Kalyani, H. C. Swart, and S. J. Dhoble, *Principles and Applications of Organic Light Emitting Diodes (OLEDs)*, Woodhead/Elsevier, 2017.

J.-J. Kim, J. Lee, S.-P. Yang, H. G. Kim, H.-S. Kweon, S. Yoo, and K.-H. Jeong, Biologically Inspired Organic Light-Emitting Diodes, *Nano Letters*, vol. 16, pp. 2994–3000, 2016.

H. Kaji, H. Suzuki, T. Fukushima, K. Shizu, K. Suzuki, S. Kubo, T. Komino, H. Oiwa, F. Suzuki, A. Wakamiya, Y. Murata, and C. Adachi, Purely Organic Electroluminescent Material Realizing 100% Conversion from Electricity to Light, *Nature Communications*, vol. 6, 8476, 2015.

D. J. Gaspar and E. Polikarpov, eds., *OLED Fundamentals: Materials, Devices, and Processing of Organic Light-Emitting Diodes*, CRC Press/Taylor & Francis, 2015.

S. Reineke, M. Thomschke, B. Lüssem, and K. Leo, White Organic Light-Emitting Diodes: Status and Perspective, *Reviews of Modern Physics*, vol. 85, pp. 1245–1293, 2013.

J. Kido, M. Kimura, and K. Nagai, Multilayer White Light-Emitting Organic Electroluminescent Device, *Science*, vol. 267, pp. 1332–1334, 1995.

J. H. Burroughes, D. D. C. Bradley, A. R. Brown, R. N. Marks, K. Mackay, R. H. Friend, P. L. Burn, and A. B. Holmes, Light-Emitting Diodes Based on Conjugated Polymers, *Nature*, vol. 347, pp. 539–541, 1990.

C. W. Tang and S. A. VanSlyke, Organic Electroluminescent Diodes, *Applied Physics Letters*, vol. 51, pp. 913–915, 1987.

Perovskite Light-Emitting Diodes (PeLEDs)

H. Meng, *Perovskite Light Emitting Diodes: Materials and Devices*, Wiley–VCH, 2024.

J. P. Martínez-Pastor, P. P. Boix, and G. Xing, *Metal Halide Perovskites for Generation, Manipulation and Detection of Light*, Elsevier/SPIE, 2023.

J. Z. Zhang, Z. Xia, and Q. Pang, *Metal Halide Perovskites: Synthesis, Properties and Applications*, World Scientific, 2023.

Y. Sun, L. Ge, L. Dai, C. Cho, J. F. Orri, K. Ji, S. J. Zelewski, Y. Liu, A. J. Mirabelli, Y. Zhang, J.-Y. Huang, Y. Wang, K. Gong, M. C. Lai, L. Zhang, D. Yang, J. Lin, E. M. Tennyson, C. Ducati, S. D. Stranks, L.-S. Cui, and N. C. Greenham, Bright and Stable Perovskite Light-Emitting Diodes in the Near-Infrared Range, *Nature*, vol. 615, pp. 830–835, 2023.

A. Fakharuddin, M. K. Gangishetty, M. Abdi-Jalebi, S.-H. Chin, A. R. bin Mohd-Yusoff, D. N. Congreve, W. Tress, F. Deschler, M. Vasilopoulou, and H. J. Bolink, Perovskite Light-Emitting Diodes, *Nature Electronics*, vol. 5, pp. 203–216, 2022.

D. Yang, B. Zhao, T. Yang, R. Lai, D. Lan, R. H. Friend, and D. Di, Toward Stable and Efficient Perovskite Light-Emitting Diodes, *Advanced Functional Materials*, vol. 32, p. 2109495, 2022.

J. Chen, H. Xiang, J. Wang, R. Wang, Y. Li, Q. Shan, X. Xu, Y. Dong, C. Wei, and H. Zeng, Perovskite White Light Emitting Diodes: Progress, Challenges, and Opportunities, *ACS Nano*, vol. 15, pp. 17150–17174, 2021.

J. Chen, J. Wang, X. Xu, J. Li, J. Song, S. Lan, S. Liu, B. Cai, B. Han, J. T. Precht, D. Ginger, and H. Zeng, Efficient and Bright White Light-Emitting Diodes Based on Single-Layer Heterophase Halide Perovskites, *Nature Photonics*, vol. 15, pp. 238–244, 2021.

A. Dey et al., State of the Art and Prospects for Halide Perovskite Nanocrystals, *ACS Nano*, vol. 15, pp. 10775–10981, 2021.

J. Shamsi, G. Rainò, M. V. Kovalenko, and S. D. Stranks, To Nano or Not to Nano for Bright Halide Perovskite Emitters, *Nature Nanotechnology*, vol. 16, pp. 1164–1175, 2021.

A. Liu, C. Bi, R. Guo, M. Zhang, X. Qu, and J. Tian, Electroluminescence Principle and Performance Improvement of Metal Halide Perovskite Light-Emitting Diodes, *Advanced Optical Materials*, vol. 9, p. 2002167, 2021.

M. Cao, Y. Xu, P. Li, Q. Zhong, D. Yang, and Q. Zhang, Recent Advances and Perspectives on Light Emitting Diodes Fabricated from Halide Metal Perovskite Nanocrystals, *Journal of Materials Chemistry C*, vol. 7, pp. 14412–14440, 2019.

B. Zhao, S. Bai, V. Kim, R. Lamboll, R. Shivanna, F. Auras, J. M. Richter, L. Yang, L. Dai, M. Alsari, X.-J. She, L. Liang, J. Zhang, S. Lilliu, P. Gao, H. J. Snaith, J. Wang, N. C. Greenham, R. H. Friend, and D. Di, High-Efficiency Perovskite–Polymer Bulk Heterostructure Light-Emitting Diodes, *Nature Photonics*, vol. 12, pp. 783–789, 2018.

P. Szuromi and B. Grocholski, eds., Perovskites (Special Section), *Science*, vol. 358, no. 6364, pp. 732–757, 10 November 2017.

L. Protesescu, S. Yakunin, M. I. Bodnarchuk, F. Krieg, R. Caputo, C. H. Hendon, R. X. Yang, A. Walsh, and M. V. Kovalenko, Nanocrystals of Cesium Lead Halide Perovskites ($CsPbX_3$, X = Cl, Br, and I): Novel Optoelectronic Materials Showing Bright Emission with Wide Color Gamut, *Nano Letters*, vol. 15, pp. 3692–3696, 2015.

Z.-K. Tan, R. S. Moghaddam, M. L. Lai, P. Docampo, R. Higler, F. Deschler, M. Price, A. Sadhanala, L. M. Pazos, D. Credgington, F. Hanusch, T. Bein, H. J. Snaith, and R. H. Friend, Bright Light-Emitting Diodes Based on Organometal Halide Perovskite, *Nature Nanotechnology*, vol. 9, pp. 687–692, 2014.

H. L. Wells, Über die Cäsium- und Kalium-Bleihalogenide, *Zeitschrift für anorganische Chemie*, vol. 3, pp. 195–210, 1893.

Historical Accounts

See also the bibliographies in Chapters 5, 6, 10, and 11.

T. D. Moustakas and R. Paiella, Optoelectronic Device Physics and Technology of Nitride Semiconductors from the UV to the Terahertz, *Reports on Progress in Physics*, vol. 80, 106501, 2017.

O. Shchekin and M. G. Craford, History of Solid-State Light Sources, in R. Karlicek, C. C. Sun, G. Zissis, and R. Ma, eds., *Handbook of Advanced Lighting Technology*, pp. 41–70, Springer, 2017.

J. Hecht, Short History of Laser Development, *Optical Engineering*, vol. 49, 091002, 2010.

R. D. Dupuis and M. R. Krames, History, Development, and Applications of High-Brightness Visible Light-Emitting Diodes, *Journal of Lightwave Technology*, vol. 26, pp. 1154–1171, 2008.

J. Hecht, The Breakthrough Birth of the Diode Laser, *Optics & Photonics News*, vol. 18, no. 7, pp. 38–43, 2007.

Zh. I. Alferov, Double Heterostructure Concept and its Applications in Physics, Electronics and Technology (Nobel Lecture in Physics, 2000), in G. Ekspong, ed., *Nobel Lectures in Physics 1996–2000*, World Scientific, 2003; H. Kroemer, Quasi-Electric Fields and Band Offsets: Teaching Electrons New Tricks (Nobel Lecture in Physics, 2000), in G. Ekspong, ed., *Nobel Lectures in Physics 1996–2000*, World Scientific, 2003.

R. H. Rediker, Semiconductor Diode Luminescence and Lasers — A Perspective, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 6, pp. 1355–1362, 2000.

D. A. Vanderwater, I.-H. Tan, G. E. Höfler, D. C. Defevere, and F. A. Kish, High-Brightness AlGaInP Light Emitting Diodes, *Proceedings of the IEEE*, vol. 85, 1752–1764, 1997.

R. H. Rediker, Research at Lincoln Laboratory Leading up to the Development of the Injection Laser in 1962, *IEEE Journal of Quantum Electronics*, vol. QE-23, pp. 692–695, 1987.

R. H. Rediker, I. Melngailis, and A. Mooradian, Lasers, Their Development, and Applications at M.I.T. Lincoln Laboratory, *IEEE Journal of Quantum Electronics*, vol. QE-20, pp. 602–615, 1984.

Seminal Publications

See also the bibliographies in Chapters 5 and 6.

S. Nakamura, M. Senoh, N. Iwasa, and S.-i. Nagahama, High-Brightness InGaN Blue, Green and Yellow Light-Emitting Diodes with Quantum Well Structures, *Japanese Journal of Applied Physics*, vol. 34, pp. L797–L799, 1995.

S. Nakamura, T. Mukai, and M. Senoh, Candela-Class High-Brightness InGaN/AlGaN Double-Heterostructure Blue-Light-Emitting Diodes, *Applied Physics Letters*, vol. 64, pp. 1687–1689, 1994.

S. Nakamura, N. Iwasa, M. Senoh, and T. Mukai, Hole Compensation Mechanism of P-Type GaN Films, *Japanese Journal of Applied Physics*, vol. 31, pp. 1258–1266, 1992.

K. Itoh, T. Kawamoto, H. Amano, K. Hiramatsu, and I. Akasaki, Metalorganic Vapor Phase Epitaxial Growth and Properties of GaN/$Al_{0.1}Ga_{0.9}N$ Layered Structures, *Japanese Journal of Applied Physics*, vol. 30, no. 9R, p. 1924, 1991.

H. Amano, M. Kito, K. Hiramatsu, and I. Akasaki, P-Type Conduction in Mg-Doped GaN Treated with Low-Energy Electron Beam Irradiation (LEEBI), *Japanese Journal of Applied Physics*, vol. 28, pp. L2112–L2114, 1989.

W. O. Groves, A. H. Herzog, and M. G. Craford, The Effect of Nitrogen Doping on $GaAs_{1-x}P_x$ Electroluminescent Diodes, *Applied Physics Letters*, vol. 19, pp. 184–186, 1971.

T. M. Quist, R. H. Rediker, R. J. Keyes, W. E. Krag, B. Lax, A. L. McWhorter, and H. J. Zeiger, Semiconductor Maser of GaAs, *Applied Physics Letters*, vol. 1, pp. 91–92, December 1962.

N. Holonyak, Jr. and S. F. Bevacqua, Coherent (Visible) Light Emission from $Ga(As_{1-x}P_x)$ Junctions, *Applied Physics Letters*, vol. 1, pp. 82–83, December 1962.

M. I. Nathan, W. P. Dumke, G. Burns, F. H. Dill, Jr., and G. Lasher, Stimulated Emission of Radiation from GaAs p–n Junctions, *Applied Physics Letters*, vol. 1, pp. 62–64, November 1962.

R. N. Hall, G. E. Fenner, J. D. Kingsley, T. J. Soltys, and R. O. Carlson, Coherent Light Emission from GaAs Junctions, *Physical Review Letters*, vol. 9, pp. 366–368, November 1962.

R. J. Keyes and T. M. Quist, Radiation Emitted by Gallium Arsenide Diodes, presented at the *Solid-State Device Research Conference*, Durham, New Hampshire, July 9–11, 1962, *IRE Transactions on Electron Devices*, vol. 9, No. 6, p. 503, July 1962.

R. J. Keyes and T. M. Quist, Recombination Radiation Emitted by Gallium Arsenide (Correspondence), *Proceedings of the IRE*, vol. 50, pp. 1822–1823, August 1962 (Date of Submission: 25 May 1962).

CHAPTER

8

COLOR VISION

8.1 VISUAL PATHWAYS 236

8.2 EYE 237

8.3 RETINA 239

8.4 PHOTORECEPTORS 240

8.5 TRICHROMATIC VISION 245

8.6 OPPONENT CHANNELS 247

8.7 NON-TRICHROMATIC VISION 250

8.8 RADIOMETRIC AND PHOTOMETRIC UNITS 252

8.9 LUMINOUS EFFICACY AND EFFICIENCY 257

The trichromatic theory of color vision was proposed in 1801 by **Thomas Young (1773–1829)**, an English physician, and extended by the German scientist **Hermann von Helmholtz (1821–1894)** in 1850. The theory postulates that there are three types of color receptors in the eye and that they suffice for perceiving all colors in nature.

LED Lighting: Devices and Colorimetry. Malvin Carl Teich.
Google Books. Published 2024.

Parts I and II of this book (Chapters 1–7) are devoted to chronicling the properties of light and describing various means for generating it. Parts III and IV (Chapters 8–11), in contrast, are dedicated to describing the detection and perception of light. The design of LED lighting systems requires a thorough understanding of how the human visual system operates and how colors (including white) are perceived. This chapter introduces the fundamental principles that underlie color vision while Chapter 9 describes the perception of color.

The visual system comprises two parallel but interconnected structures. The scotopic visual system relies on rod photoreceptors that contain the rod opsin photopigment (also called scotopsin) and operate at very low light levels and only in gray-scale. The photopic visual system, in contrast, relies on three types of less-sensitive cone photoreceptors that contain cone opsin photopigments (also called photopsins) and operate at normal light levels and in color. The rods are situated in the peripheral visual field and provide high sensitivity (at the expense of visual acuity), whereas the cones are principally localized in the central visual field and offer high visual acuity (at the expense of sensitivity).

The presentation in this chapter begins with a macroscopic (systemwide) purview and draws to a close from a microscopic (cellular) perspective. We begin with an overview of the central pathways in the visual system that govern the transfer of information from the retina to the cortex (Sec. 8.1). This is followed by a description of the structure and function of the components of the peripheral visual system that involve image formation and image transduction, i.e., the eye (Sec. 8.2) and the retina (Sec. 8.3), respectively. We proceed to discuss the light-sensitive photoreceptors in the retina, namely the rods and cones (Sec. 8.4). In particular, we examine their morphology, phototransduction processes, interplay with other types of cells in the retina, and spatial distribution across the retina and fovea.

We then turn to the trichromatic theory of color vision (Sec. 8.5), which was proposed by Thomas Young (p. 234) in 1802 and refined some fifty years later by Hermann von Helmholtz (p. 234) and James Clerk Maxwell (p. 24). The Young–Helmholtz theory has been eminently successful in explaining color vision.[†] They postulated that the eye contained three types of photoreceptors, now called cones, that responded in different wavelength regions and that different relative strengths of their signals were interpreted by the brain as different visible colors. The three types of photoreceptors were said to be short-, middle-, and long-preferring, now called S-, M-, and L-cones. Young and Helmholtz were evidently unaware of earlier work by Palmer that had espoused the concept of three different color receptors in 1777 (Palmer's manuscript was not found until 1956). We present modern results for the cone spectral sensitivities, the relative effectiveness of light of different wavelengths in stimulating the photopic visual system.

This is followed by an examination of the relation between the trichromatic and opponent-process theories of color vision (Sec. 8.6). We rely on color blindness, tetrachromacy in some humans, and non-trichromatic vision in the animal kingdom to illustrate that trichromatic vision is not universal (Sec. 8.7). We then introduce radiometric and photometric units to provide quantitative metrics that can be used to describe the physical and perceptual features of color vision, respectively (Sec. 8.8). We conclude by examining the collection of luminous efficacies and efficiencies that serve as key

[†] T. Young, The Bakerian Lecture. On the Theory of Light and Colours, *Philosophical Transactions of the Royal Society of London*, vol. 92, pp. 12–48, 1802; H. L. F. von Helmholtz, Physiologische Optik, in G. Karsten, ed., *Handbuch der physiologischen Optik*, Volume 9, *Allgemeine Encyklopädie der Physik*, pp. 1–874, Leopold Voss (Leipzig), 1867 [Translation: *Handbook of Physiological Optics*, N. Wade, ed. and J. P. C. Southall, translator, Thoemmes Continuum (Bristol), 2000].

measures LED lighting.

8.1 VISUAL PATHWAYS

The process of visual detection is complex and relies on a host of interdisciplinary and interconnected processes that involve signaling in different modalities and at different scales. The percept of a simple flash of light, for example, entails a sequence of detections and transductions that involve multiple cascaded mechanical, chemical, and electrical processes.

The principal pathways for the transmission of information from the peripheral visual system to the cortex are highlighted in Fig. 8.1-1. The collection of retinal-ganglion-cell axons that originate at the **retina** and comprise the **optic nerve** (cranial nerve II) carry visual information in the form of action potentials to neurons in the **lateral geniculate nucleus**, which resides in the thalamus (all sensory information, with the exception of olfaction, is routed through the thalamus on its way to the cortex). Retinal-ganglion-cell axons also project to other sub-cortical areas in the midbrain, including the **superior colliculus**, where visual, auditory, and somatosensory information are coordinated and where rapid saccadic eye movements originate; and also to the **pretectal region**, where pupillary reflexes are produced. Axons from the geniculate project to the **primary visual cortex**, a sub-region of the **visual cortex**.

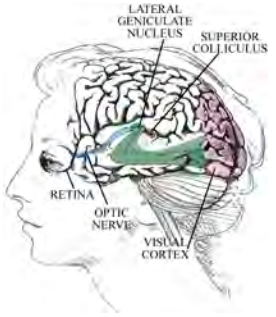


Figure 8.1-1 Visual information flows, via the optic nerve and optic tract, from the retina in the peripheral visual system to the lateral geniculate nucleus and superior colliculus in the central visual system, and thence, via the optic radiation, to the visual cortex in the occipital lobe of the brain.

A more detailed view of the wiring diagram of the visual pathways is provided in the two panels of Fig. 8.1-2. Each optic-nerve bundle comprises approximately 1.2 million **retinal-ganglion-cell axons** (also called optic-nerve fibers). As schematized in this figure, at the **optic chiasm** each eye sends half of these axons to regions on the same (ipsilateral) side of the brain and the other half to regions on the opposite (contralateral) side. The pathways are wired such that each geniculate receives input only from the contralateral half of the **visual field** via the **optic tract**; the two halves of the visual field (**hemifields**) are distinguished by the dark and light shadings at the entrances to the eyes in Fig. 8.1-2. As an example, an object in the right visual field is imaged on the nasal retina of the right eye, whose optic-nerve fibers project solely to the left geniculate. In short, the left geniculate receives information from ganglion-cell axons in both eyes, but all arriving axons are linked to the right visual field. Similarly, the right geniculate is innervated only by ganglion cells linked to the left visual field.

The horizontal field-of-view available in human vision is roughly $200°$, of which about $160°$ is available to each eye and approximately $120°$ is binocular (seen with both eyes), which enables distance and depth to be estimated. The vertical field-of-view is about $135°$ and the solid angle subtended by each eye is estimated to be $\Omega \approx 4.57$ sr.

The geniculate neurons in turn project, via the **optic radiation**, principally to layer 4 of a sub-area of the human visual cortex known by several interchangeable names:

- Primary visual cortex.
- Visual area 1, denoted V1.
- Striate cortex.
- Brodmann's area 17.

Higher-level areas of the visual cortex, denoted V2–V5 and schematically labeled in Fig. 8.1-2, are dedicated to various aspects of visual perception:

1. V1 deals with processing the elementary features of a visual scene such as edges, orientations, colors, and contrast.
2. V2 is involved in processing information related to orientation, luminance, and color.
3. V3 is responsible for the analysis of spatial frequency and stereoscopic depth.
4. V4 plays a key role in color perception and object recognition.
5. V5 (also known as middle-temporal or MT) is dedicated to motion and depth perception.

Each of these areas exhibits a columnar cortical structure and there are extensive feedback connections back to the LGN and among the different areas.

The two brain hemispheres communicate via the **corpus callosum**, which contains some 200 million nerve fibers.

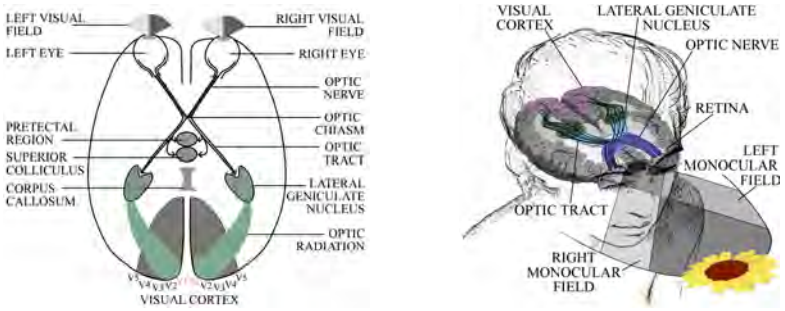


Figure 8.1-2 Top view of the visual system that reveals the information flow from the retina to the superior colliculus and lateral geniculate nucleus via the optic nerve and optic tract. All ganglion-cell axons responsive to a particular half of the visual field, comprising half of all optic nerve fibers from each eye, meet at the optic chiasm and continue via the optic tract to the contralateral geniculate. The optic radiation comprises geniculate-cell axons carrying information to the primary visual cortex (area V1), which resides in the occipital lobe of the brain. The corpus callosum permits the two halves of the brain to share information.

8.2 EYE

The optics of the eye, portrayed in simplified form in Fig. 8.2-1, serves to project a two-dimensional image of visual space onto the retina. The **bulb** of the eye, also called the **globe**, is roughly spherical in shape so that the location of its various features can be conveniently described in terms of the azimuthal angle, polar angle, and radial distance in a spherical coordinate system. Together with its appendages, the globe is recessed in a protective, pyramidal, bony socket in the skull called the **orbit**, within which it may move in three dimensions.

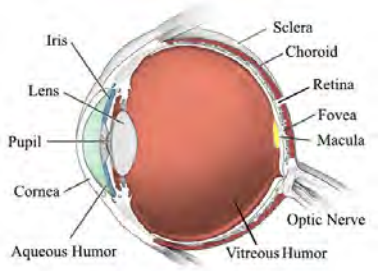


Figure 8.2-1 Principal components of the eye. The spatial configurations of the iris and lens change autonomically as the light level and object distance are modified, respectively. Like an adjustable camera f-stop, the iris governs the amount of light admitted to the eye via the pupil. The thickness of the lens is altered by the ciliary muscle surrounding it, which changes its focusing power and enables it to selectively focus on an object at a particular distance.

The elements of the eye, most of which contain multiple cellular layers, serve the following functions:

- *Cornea.* The cornea protects the eye from damage. Its anterior surface is approximately spherical, with a radius of curvature typically just under 8 mm, which serves as a lens and is responsible for about $^2/_3$ of the eye's refractive power.
- *Aqueous Humor.* The aqueous humor is the transparent fluid that fills the space between the cornea and the lens. It provides nutrients for both, which are avascular to ensure transparency, and maintains the intraocular pressure.
- *Pupil.* The pupil admits light to the eye. Its diameter, which is governed by the smooth muscles of the iris, changes from about 8 mm to 1.5 mm as the ambient light goes from dim to bright. The retinal image is of optimal quality when the effects of aberration and diffraction are balanced, which corresponds to a pupil diameter of about 2.5 mm.
- *Iris.* The iris autonomically expands or contracts with the level of ambient light to control the amount of light entering the pupil. The circular *ciliary body*, which produces the aqueous humor, is an extension of the iris.
- *Lens.* The lens, an avascular and transparent structure, is responsible for the remaining $^1/_3$ of the refractive power of the eye. It autonomically adjusts its focal length f to provide a crisp image at the retina. When viewing objects close at hand, the *ciliary muscle* adjacent to the lens contracts and the lens becomes more bulbous, which decreases its focal length and makes it optically more powerful; this process is known as *accommodation*. The lens is often assumed to obey the thin-lens imaging equation $^1/_{z_1} + {}^1/_{z_2} = {}^1/_f$ (Sec. 1.5), but its refractive index increases toward its center so it is more properly considered as a graded-index (GRIN) optical element (Sec. 2.5).
- *Vitreous Humor.* The vitreous humor is the transparent fluid that fills the space between the lens and the retina. It serves to maintain the spherical shape of the eye while acting as a shock absorber and, like the aqueous humor, helps maintain the lens.
- *Sclera.* The sclera is the opaque white outer layer of the eye that serves as its supporting wall and protects it from injury. Strong and fibrous, it extends from the optic nerve to the outer edge of the cornea, with which it is contiguous. The sclera is covered by *conjunctiva*, clear mucus membranes that lubricate the eye. Extraocular muscles attached to the sclera move the globe as a whole.
- *Choroid.* The choroid, which lies beneath the sclera, contains melanin-pigmented connective and vascular tissue. It provides the blood supply for the photoreceptor layer of the retina and absorbs stray light. The choroid, ciliary body, and iris comprise the *uvea*.
- *Retinal Pigment Epithelium.* The retinal pigment epithelium (RPE) consists of a single layer of melanin-containing cells that lies between the choroid and the photoreceptor layer of the retina, for which it provides protection and maintenance. The RPE absorbs scattered light, thereby preventing back-reflection, and provides

metabolic support.

- *Retina, Macula, and Fovea.* The two-dimensional image projected onto the retina by the optics of the eye is detected by the light-sensitive *rods* and *cones* (collectively called *photoreceptors*) that reside therein. The retina's *neural circuitry*, fed by the photoreceptor outputs, processes and encodes the image information into a parallel stream of *action potentials* carried to higher visual centers by the retinal-ganglion-cell axons. Action potentials are required to traverse these relatively long distances. The macula and fovea are sub-regions of the retina.
- *Optic Disc.* The optic disc (*optic nerve head*) is a raised, nearly circular region on the retina located about 3–4 mm to the nasal side of the fovea, where the retinal-ganglion-cell axons that comprise the *optic nerve* exit the eye. This region is devoid of photoreceptors and corresponds to the *blind spot* of the eye.
- *Fundus.* The posterior lining of the eye that lies opposite the pupil and comprises the retina, macula, optic disc, and their attendant blood vessels is known as the fundus.

The unit of refractive power widely used in visual science is the diopter D***, which is defined as the reciprocal of the focal length expressed in meters, i.e.,*** $\mathrm{D} = 1/f$.

8.3 RETINA

The retina is an intricate layered structure that lines the posterior region of the eye. In adults, it lies approximately 25 mm posterior to the cornea, has an area of about 10 cm^2, and is nominally 270-μm thick. The function of the retina is to detect the optical image focused on it by the optics of the globe, to transduce that image into collections of neural action potentials, and to arrange for these action potentials to be transmitted to higher visual centers in the brain.

The retina is divided into several lateral regions with different anatomical characteristics, as delineated in Fig. 8.3-1. The *macula lutea* is a region at the center of the posterior retina opposite the pupil that has, roughly speaking, a diameter of about 5.5 mm (18.3° of visual field). The macula contains protective, antioxidant, carotenoid pigments that filter out damaging blue light and render it yellowish in color. The macula in turn comprises a number of roughly circular subregions that contain cells with different morphologies and functions (the diameters of these subregions are indicated):

Figure 8.3-1 Diagram displaying the principal lateral regions of the retina and their rough dimensions: *Area and thickness of the* ***retina***: 10 cm^2 and 270 μm, respectively. *Diameter of the* ***macula***: 5.5 mm (18.3°). *Diameter of the* ***fovea***: 1.5 mm (5°). *Diameter of the* ***foveola***: 0.3 mm diameter (1°). Not shown in the diagram are the annular ***perifovea*** and ***parafovea***, nor is the central ***umbo*** displayed. The conversion factor for length to degrees at the retina is $\approx 3.33^\circ/\mathrm{mm}$. The diagram is not to scale.

- *Perifovea.* An annular region in the macula that circumscribes the parafovea. Measured from the umbo, it is an annulus of inner diameter 2.5 mm (8.3°) and outer diameter 5.5 mm (18.3°). The width of the perifoveal band is thus 1.5 mm. The perifovea principally contains rods but also contains some cones.

- *Parafovea*. An annular region of the macula that lies between the perifovea and the fovea. Measured from the umbo, it is an annulus of inner diameter 1.5 mm (5°) and outer diameter 2.5 mm (8.3°). The parafoveal band thus has a width of 0.5 mm.
- *Fovea*. A circular region in the macula with a diameter of 1.5 mm (5°). The fovea takes the form of a depression within which the retinal thickness is reduced from ≈ 270 μm to ≈ 170 μm, since several cellular layers are eliminated. This serves to diminish scattering and thereby to enhance visual acuity. The central region of the fovea that is totally rod-free has a diameter of 0.5 mm (1.7°). Although the fovea comprises $\leqslant 1\%$ of the area of the retina, its output influences some 50% of the cells in the visual cortex.
- *Foveola*. This rod-free, capillary-free, and pedicle-free zone in the central area of the fovea has a diameter of 0.3 mm (1°). The paucity of S-cones (blue cones) renders vision in this small region essentially dichromatic. The foveola contains approximately 18000 specialized red (L-cones) and green (M-cones) cones; their outer segments are about twice the length of those in the parafovea and they are more densely packed than elsewhere. The foveola exhibits higher visual acuity than other areas of the fovea.
- *Umbo*. The very center of the foveola, with a diameter of 0.15 mm (0.5°), offers the highest visual acuity of all regions. The *Müller cells* that reside in the umbo, which are unique in form, are glial cells that provide structural and nutritional support.

8.4 PHOTORECEPTORS

Photoreceptors are specialized retinal neurons that convert absorbed photons into chemical and electrical signals that are processed by the neural circuitry of the retina before being sent to higher centers in the visual system. There are two varieties, **rod photoreceptors** and **cone photoreceptors**, called **rods** and **cones** for short, and their structures are remarkably similar for all vertebrates. The human retina contains about 120 million rods, whose diameters and outer-segment lengths are roughly 2 μm and 30 μm, respectively in the region where their density is highest. It also contains some 6.5 million cones, whose diameters and outer-segment lengths are about 2 μm and 35 μm, respectively in the fovea. Both rods and cones use sophisticated adaptation techniques to ensure that visual function is maintained over the daily twelve order-of-magnitude variation in light level that reaches earth. Photoreceptors also exhibit light-waveguiding capabilities [Fig. 1.6-1(c)]; their long axes point toward the light rays that enter the eye at the pupil.

Rods and Cones Serve Different Functions

- Rods mediate **scotopic vision** (night vision) and offer the ultimate in sensitivity; when fully dark-adapted, and at luminances $L_{\mathrm{v}} \lesssim 10^{-2}$ cd/m^2, rods can reliably signal the absorption of a single photon. The threshold of rod vision lies at about $L_{\mathrm{v}} = 10^{-6}$ cd/m^2. However, rods operate in gray-scale and are rendered inoperative by saturation at photopic light levels. Their response time is also significantly longer than that of cones.
- Cones, so-named because the infoldings of their outer segments taper slightly in some retinal regions, mediate **photopic vision** (day vision) at luminances $L_{\mathrm{v}} > 3$ cd/m^2 and excel at spatial and temporal resolution. Cones, which enable color vision, do not saturate at high levels of steady illumination. However, they are about a factor of 100 less sensitive than rods at scotopic light levels.
- Both rods and cones participate in *mesopic* (twilight) vision, which is operative at luminances in the range $0.01 \lesssim L_{\mathrm{v}} \lesssim 3$ cd/m^2.

Morphology

Although rods and cones have their own unique features, their structure and operation are similar in many respects. As displayed in Fig. 8.4-1, both are divided into outer and inner segments, and are conveniently represented in terms of four subcellular compartments:

- An **outer segment (OS)** that contains lipid bilayers housing opsin molecules covalently bound to chromophores (visual pigment) and is dedicated to the detection of photons and phototransduction.
- An **inner segment (IS)**, containing metabolic machinery (mitochondria), that is devoted to the synthesis of proteins and lipids (Golgi apparatus and endoplasmic reticulum).
- A **cell nucleus** in the inner segment that is the cell's genetic locus and that regulates its activities.
- A **synaptic terminal**, lying at the edge of the inner segment and in the outer plexiform layer, where a decrease in exocytosis of the neurotransmitter glutamate signals an increase in incident light level at the photoreceptor OS.

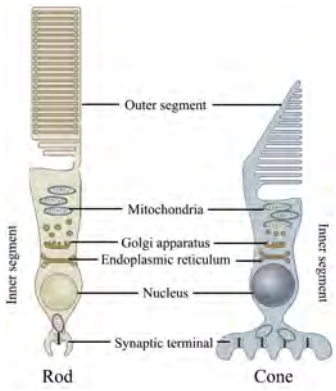


Figure 8.4-1 Cellular features shared by vertebrate rods and cones. The absorption of a photon at the outer segment (OS) triggers a molecular conformational change in an opsin molecule and the activation of its bound chromophore, which then modifies the sodium-ion current at the OS plasma membrane. Endowed with the spatial and temporal information inherent in the optical stimulus, this current modulates the neurotransmitter release at the synaptic terminal. The neurotransmitter signal is fed to the outer and inner plexiform layers of the retina, where analog and digital signal processing take place, respectively. Ultimately, the image information is encoded into neural spike trains carried to higher visual centers by retinal-ganglion-cell axons. The inner segment provides the photoreceptors with energy and maintenance.

Phototransduction

The transduction of light at the retina is the process whereby an optical image incident on the outer segments of a collection of rods and cones is processed and ultimately converted into a parallel stream of neural action potentials that propagate on the retinal-ganglion-cell axons that form the optic nerve and optic tract, and that carry the information to higher visual centers in the brain. The biochemical cascade underlying the transduction process is complex but reasonably well understood.

As indicated in the caption to Fig. 8.4-1, the process is initiated at an individual photoreceptor by the absorption of a photon, which triggers a molecular conformational change in a retinal-based **chromophore** resident in the photoreceptor OS. That in turn activates the **opsin** molecule to which the chromophore is bound, resulting in a modification of the sodium-ion current flow at the OS plasma membrane. Controlled by the photoreceptor cell nucleus in the outer nuclear layer (ONL), the decrease in sodium current causes the transmembrane potential of the entire cell to become **hyperpolarized**. This reduces the internal calcium-ion concentration and ultimately leads to a decrease in **neurotransmitter exocytosis** at the photoreceptor synaptic terminal located at the edge of the inner segment (IS) in the **outer plexiform layer (OPL)**.

The photoreceptor output signal undergoes analog signal processing controlled by *horizontal* and *bipolar* cells in the OPL. Bipolar cells activated in response to an increase in photon absorption at the photoreceptor level are termed ON bipolar cells whereas those activated in response to a decrease in photon absorption are termed OFF bipolar cells. Both are present in the OPL and their concomitant responses, opposite in sign, suggest that the important variable is the *change* in glutamate concentration rather than its absolute value. Because the diurnal variation in ambient light level is so large, the OPL circuitry is designed so that its output relays the *contrast* of various features of the image rather than the absolute level of photoreceptor activity. The *center–surround receptive field* structure of bipolar cells facilitates this process. The **inner plexiform layer (IPL)** is the site where digital signal processing of the temporal and complex features of the stimulus encoded in the photoreceptor output is implemented by *amacrine*, *interplexiform*, and *bipolar cells*, whose somas lie in the inner nuclear layer (INL).

The **ganglion-cell layer (GCL)** contains the somas of the **retinal ganglion cells (RGCs)**. ON (OFF) bipolar cells are presynaptic to ON-CENTER (OFF-CENTER) RGCs; both generate random sequences of action potentials that propagate to the central nervous system on axons found in the nerve-fiber layer (NFL). The information carried on these axons is generally incorporated in the time-varying firing rate, which is superposed on a background of randomly firing spontaneously action potentials that constitutes noise. There are some 30 types of RGCs, each of which forms a mosaic that spatially tiles the retina and constitutes a unique representation of the visual scene. The different representations are computed in parallel by relying on retinal neural circuit elements that are concomitantly configured in multiple ways.

The collected outputs of the RGC mosaics are carried to higher visual centers in parallel by the 1.2 million RGC axons of the optic nerve. The optic nerves from both eyes decussate at the optic chiasm and become the optic tracts, each responsive to a different hemifield. The constituent RGC axons project to the lateral geniculate nuclei in the thalamus, as displayed in Figs. 8.1-1 and 8.1-2.

EXAMPLE 8.4-1. ***OCT Imaging of the Retina.*** The retinal layers involved in the transduction process described above can be viewed by making use of **optical coherence tomography** (**OCT**), a noninvasive, non-contact, *in vivo* imaging technique that relies on light interferometry. A test beam from the instrument directed at the pupil passes through the vitreous and impinges normally on the retina. *Axial sections* that reveal the reflectances and depths of the various retinal layers are obtained at multiple locations. The technique is also useful for imaging the optic nerve head and the anterior

eye. The axial and lateral resolution available with OCT retinal imaging is in the vicinity of 1–15 μm. Superior resolution is attained by making use of *adaptive optics* in conjunction with OCT (AO-OCT). This imaging technique has proven to be highly effective for imaging multilayered media in other disciplines as well.

OCT imaging was initially implemented in the time domain by making use of an interferometric configuration in which echoes were successively recorded from refractive-index boundaries in a sample as an auxiliary mirror increased the observation depth. The development of **spectral-domain OCT (SD-OCT)** constituted a substantial advance since it allowed echoes from the entire depth of the sample to be simultaneously measured, which provided higher detection sensitivity, data-acquisition rates, and resolution. **Swept-source OCT (SS-OCT)**, perhaps the most effective variation on the theme, makes use of a frequency-swept diode laser as the light source and yields superior efficiency. SS-OCT also provides deeper tissue penetration since it operates at longer wavelengths, further improving performance.

An OCT scan collected from the macula of a normal subject is displayed in Fig. 8.4-2. The eight boundaries separating the principal retinal layers are highlighted in color and the layers are identified at the bottom of the figure.

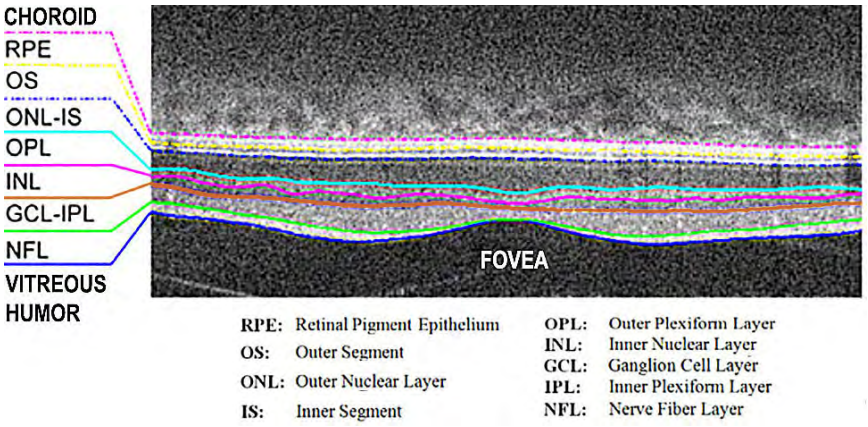


Figure 8.4-2 Cross-sectional image of the normal macula, which contains the fovea, obtained using spectral-domain optical coherence tomography (SD-OCT). The key at the bottom of the figure identifies the retinal layers; the eight boundaries separating them are highlighted in different colors. The roles played by the various layers in the transduction of light at the retina were described earlier. (Data adapted from S. J. Chiu, X. T. Li, P. Nicholas, C. A. Toth, J. A. Izatt, and S. Farsiu, Automatic Segmentation of Seven Retinal Layers in SDOCT Images Congruent with Expert Manual Segmentation, *Optics Express*, vol. 18, pp. 19413–19428, 2010, Fig. 1.)

Photoreceptor Spatial Density

The spatial density of rods and cones across a horizontal retinal strip (number/mm^2), from the umbo at 0° retinal **eccentricity** (azimuthal angle) to approximately 80° in the temporal and nasal retinas, is plotted in Fig. 8.4-3. There is some variation in the curves across individuals. The density of cones (green curve) is highest at the umbo, which is devoid of rods, while the density of rods (blue curve) is highest at $\approx 17.5^\circ$ eccentricity in the temporal retina, where there are few cones. The spatial extents of the fovea and macula are indicated below the plot (horizontal red bars). The five micrographs at the top of the plot illustrate representative configurations of rods (blue dots) and cones (green dots) at several eccentricities. The diameters of the rods and cones become larger toward the periphery, as can be discerned from the sketches. The increasing rod

diameters compensate for the decreasing rod density as the eccentricity increases, so that essentially all of the light falling on the retina is intercepted.

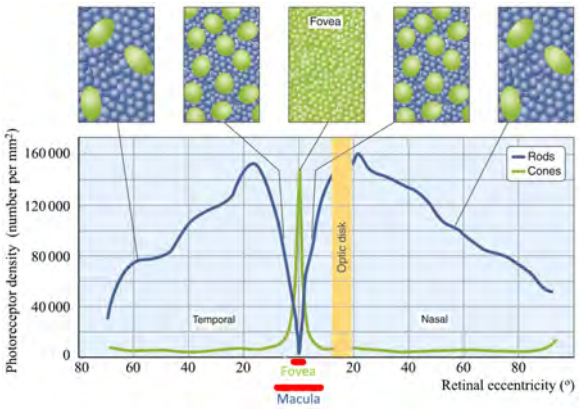


Figure 8.4-3 Spatial density of rods and cones across a horizontal strip of the retina that contains the fovea and optic disc, from the umbo ($0°$ retinal eccentricity) to $\approx 80°$ in the temporal and nasal retinas. The density of cones is highest at the umbo while the density of rods is highest at approximately $17.5°$ eccentricity in the temporal retina, which is more convenient for carrying out experiments since it is not obscured by the optic disc. The spatial extents of the fovea and macula are indicated at the bottom of the plot. The micrographs above the plot are horizontal sections at the photoreceptor layer, all taken with the same magnification, that illustrate representative rod and cone configurations. (Data adapted from citealp*osterberg35; citealp*curcio90; and citealp*rodieck98.)

Foveal Cone Mosaic. The central region of vision with the highest acuity, the fovea, contains mainly cones (Fig. 8.4-3). The distribution of S-, M-, and L-cones for a normal subject at an eccentricity of $1.2°$ nasal, just a tad beyond the foveola, is illustrated in Fig. 8.4-4. Foveal cones are in large part packed tightly in a hexagonal configuration; at this eccentricity the cone spacing and density are $\approx 4\ \mu$m and ≈ 40000 cones$/$mm^2, respectively. The proportion of M- and L-cones varies widely for subjects with normal color vision and each type of cone is spatially distributed in more-or-less random fashion. The S-cones are not nearly as closely spaced as the M- and L-cones because the short-wavelength component of the image at the retina (to which the S-cones are sensitive) is blurred as a result of axial chromatic aberration in the eye's lens; hence, closer spacing would not improve acuity. Approximately half the fibers in the optic nerve carry information from the fovea; the other half carry information from the remainder of the retina.

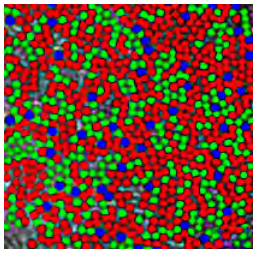


Figure 8.4-4 Foveal cone mosaic. The distribution of S-, M-, and L-cones for a retinal patch at $1.2°$ eccentricity (nasal) for a subject with normal color vision. In this pseudocolor image, the S-, M-, and L-cones are portrayed as blue, green, and red, respectively. Of the roughly 700 cones featured in this image, the S-, M-, and L-cones constitute 6%, 32%, and 62% of the total, respectively. (Data adapted from H. Hofer, J. Carroll, J. Neitz, M. Neitz, and D. R. Williams, Organization of the Human Trichromatic Cone Mosaic, *Journal of Neuroscience*, vol. 25, pp. 9669–9679, Fig. 4, 2005.)

8.5 TRICHROMATIC VISION

In the normal human retina, color vision is mediated by three types of cones, each containing a different opsin and exhibiting a unique spectral response:

- **S-cones** (blue cones), with peak sensitivity at a **S**hort wavelength (≈ 425 nm).
- **M-cones** (green cones), with peak sensitivity at a **M**iddle wavelength (≈ 535 nm).
- **L-cones** (red cones), with peak sensitivity at a **L**ong wavelength (≈ 565 nm).

Cones are widely referred to as blue, green, or red, but these designations are misleading for two reasons: 1) the peak sensitivities of the three types of cones are in the violet, green, and yellow-green, respectively, rather than in the blue, green, and red (Figs. 2.4-1 and 8.5-1); and 2) the color sensation elicited when a particular type of cone is excited is not simply blue, green, or red.

Trichromacy is not a property of the incident light, but rather is a consequence of the presence of three types of cone photoreceptors in the human retina.

Cone Spectral Sensitivities

Humans derive color information from the relative responses of three types of cones, each with a unique visual-sensitivity spectrum. Figure 8.5-1(a) displays the psychophysical spectral sensitivity, on linear coordinates and normalized to unity, for S-, M-, and L-cones. These curves are known as the **cone fundamentals**. The spectral peaks for S-, M-, and L-cones generally fall in the ranges 420–430, 530–535, and 560–565 nm, respectively; the three opsins underlying the cone sensitivities comprise different amino-acid sequences, and there is significant variability of opsin genotype among normal humans. Both energy-based and photon-number-based versions of the sensitivity curves have been established [the relationship is spelled out in (3.4-6)].

The magnitudes of the weighted versions of these curves, displayed in Fig. 8.5-1(b), nominally reflect the relative densities of each cone type in a typical normal observer (see caption of Fig. 8.4-4). Every curve displayed in Fig. 8.5-1(a),(b) is sufficiently broad that it extends over an appreciable portion of the visible spectrum and the collection of curves in each panel spans the entire visible spectrum. Consequently, it is not possible to excite only one type of cone with a physical light source.

EXAMPLE 8.5-1. ***Iodopsin Absorptions Predict Photopic Visual Sensitivity.*** The wavelength dependencies of the behavioral and biological sensitivities associated with the S, M, and L cones are illustrated in normalized form, on semilogarithmic coordinates, in Fig. 8.5-2. The three curves represent human psychophysical measurements carried out with protanopic, deuteranopic, and tritanopic observers (definitions provided in Example 8.7-1), and have been corrected for wavelength-dependent filtering imposed by the lens and macular pigment. A similar plot is presented in Fig. 8.5-1(a) on linear coordinates. The symbols represent the normalized absorption spectra of the underlying cone opsin molecules, determined spectroscopically and also corrected for wavelength-dependent filtering. The cone opsins, or **photopsins**, are also called **iodopsins** when bound to a chromophore. The small volumes of cone opsins available in the eye precluded the direct determination of their absorption spectra until suitable laboratory techniques were developed in the 1990s. Based on the alignment of the curves with the symbols, the behavioral data can be said to follow from the biological data.

Photopic Luminous Efficiency Function

The **photopic luminous efficiency function** $V(\lambda_0)$ displayed in Fig. 8.5-3, also known as the **photopic luminosity function**, represents the overall measured photopic sensitivity curve as a function of the wavelength λ_0 for a standard trichromatic human

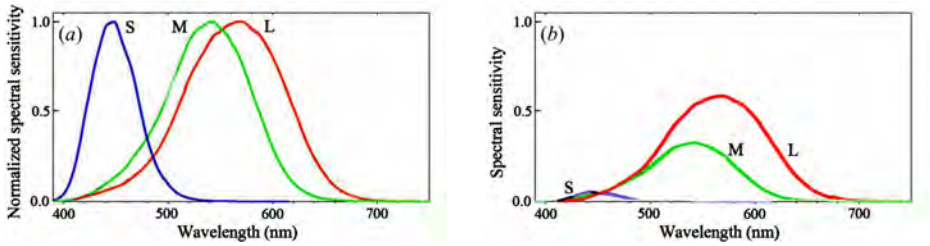


Figure 8.5-1 (a) Normalized psychophysical spectral sensitivity curves for S-, M-, and L-cones on linear coordinates; the peaks have values of unity and occur at the wavelengths 445, 540, and 565 nm, respectively, in these particular data. The M and L curves overlap extensively since the molecular structures of their underlying opsins are similar. These curves, called the cone fundamentals, are also designated $\bar{s}(\lambda_0)$, $\overline{m}(\lambda_0)$, and $\bar{l}(\lambda_0)$, respectively, and serve as color matching functions for the LMS color space discussed in Sec. 9.5. (Data adapted from A. Stockman, L. T. Sharpe, and C. Fach, The Spectral Sensitivity of the Human Short-Wavelength Sensitive Cones Derived from Thresholds and Color Matches, *Vision Research*, vol. 39, pp. 2901–2927, 1999 and A. Stockman and L. T. Sharpe, The Spectral Sensitivities of the Middle- and Long-Wavelength-Sensitive Cones Derived from Measurements in Observers of Known Genotype, *Vision Research*, vol. 40, pp. 1711–1737, 2000.) (b) S-, M-, and L-cone spectral sensitivity curves weighted by their relative densities at an eccentricity of $1.2°$ (see caption of Fig. 8.4-4).

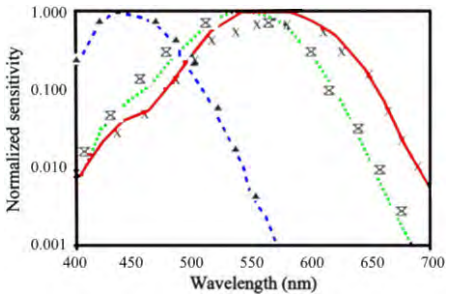


Figure 8.5-2 Normalized human psychophysical sensitivities associated with the three types of cones (S – –, M ···, L —), vs. the free-space wavelength λ_0. Normalized absorption spectra for the three underlying cone iodopsins (S ▲, M ⧖, L x) vs. λ_0. The two sets of data are in accord. (Data adapted from A. Stockman, D. I. A. MacLeod, and N. E. Johnson, Spectral Sensitivities of the Human Cones, *Journal of the Optical Society of America A*, vol. 10, pp. 2491–2521, 1993.)

observer, normalized to unity at its peak ($\lambda_0 = 555$ nm). As such, $V(\lambda_0)$ is a sum of the weighted spectral sensitivity curves for the M- and L-cones portrayed in Fig. 8.5-1(b). The S-cones play a negligible role in this enterprise since there are few in the fovea.

The photopic luminous efficiency function was adopted as a standard by the Commission Internationale de l'Éclairage (CIE, International Commission on Illumination) in 1924 (see introduction to Chapter 9). An updated version for daylight adaptation, denoted $V^*(\lambda_0)$, was set forth in 2005.[†] It should be kept in mind, however, that $V(\lambda_0)$ varies among observers and is affected by the spatial properties of the target, the retinal location, and the mean state of chromatic adaptation. An analogous scotopic luminous efficiency function, $V'(\lambda_0)$, was adopted as a standard by the CIE in 1951.

The photopic luminous efficiency function is a representation of the relative effectiveness of light of different wavelengths in stimulating the photopic visual system.

[†] L. T. Sharpe, A. Stockman, W. Jagla, and H. Jägle, A Luminous Efficiency Function, $V^*(\lambda)$, for Daylight Adaptation, *Journal of Vision*, vol. 5, pp. 948–968, 2005; A Luminous Efficiency Function, $V^*_{D65}(\lambda)$, for Daylight Adaptation: A Correction, *Color Research and Application*, vol. 36, pp. 42–46, 2011.

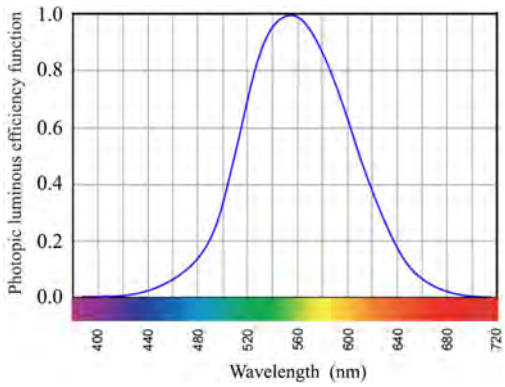


Figure 8.5-3 The photopic luminous efficiency function $V(\lambda_0)$ is the overall measured photopic sensitivity curve, plotted as a function of the free-space wavelength λ_0, for a standard human observer. This dimensionless normalized function, which attains its peak value of unity at $\lambda_0 = 555$ nm, is a sum of the weighted M and L spectral sensitivity curves displayed in Fig. 8.5-1(*b*); the S cones play a negligible role. $V(\lambda_0)$ represents the relative effectiveness of light of different wavelengths in stimulating the photopic visual system.

EXAMPLE 8.5-2. ***The Color of Near-Infrared Light.*** A substantial body of research, dating to Helmholtz's time, shows that the perception of light by humans extends over a wavelength range far greater than that usually recognized, which is 380–780 nm. In experiments carried out with lasers operated at five different near-infrared wavelengths, Sliney reported in 1976 that the photopic luminous efficiency function $V(\lambda_0)$ displayed in Fig. 8.5-3, when plotted on semilogarithmic coordinates, extended over a far broader range of wavelengths, namely 310–1100 nm (the details vary with the radiance of the source and there is variation among individual observers). He established that there was no sharp perceptual dividing line between the spectral regions conventionally referred to as visible and near infrared. Sliney further reported that all wavelengths in the range 625–1100 nm appear red to the normal observer. Indeed, Keyes and Quist (p. 198), in their initial reports on the development of the GaAs LED in 1962, reported that the 930-nm radiation emitted from their device was visible and was perceived as red (see footnote on p. 199).

Univariance

Cones do not directly sense the wavelength of the incoming photons: The absorption of a photon at the outer segment of a cone triggers an all-or-none conformational change in the chromophore, whatever the photon's wavelength. However, for a cone of a particular type, the rate of photon absorption from a beam of light of a given wavelength is proportional both to the absorption spectrum of the cone at that wavelength and to the incident photon flux; the closer the wavelength of the photon is to the spectral peak and/or the larger the photon flux, the greater the absorption rate. Consequently, different combinations of wavelength and photon flux can result in an identical cone response, a principle known as **univariance**. W. A. H. Rushton, the vision scientist who formulated this principle, put it this way: "The output of a receptor depends upon its quantum catch, but not upon what quanta are caught."

Trichromatic vision is mediated by local comparisons of the relative photon-absorption rates of the three types of cones, and not by the direct sensing of the wavelengths of the incident photons. Just as mixing the light of three primary colors suffices for* GENERATING *all visible colors, comparing the light absorbed by three types of cones suffices for* PERCEIVING *all visible colors.

8.6 OPPONENT CHANNELS

Two plausible and competing theories of color vision emerged in the nineteenth century. The first was the Young–Helmholtz *trichromatic theory* of 1801/1850 that follows naturally from the presence of three types of retinal cones with distinct spectral sensitivities [Fig. 8.5-1(*a*)].

The second was a theory proposed by the German physiologist Ewald Hering in 1874 that relied on **opponent channels**. Hering formulated his *opponent-process theory* in response to observations that certain colors, when mixed, yield a composite that has no hint of either of the two constituent colors. He noted, for example, that mixing red and green never produced "reddish-green;" rather, it produced a color such as yellow or gray. Nor did mixing blue and yellow ever produce "bluish-yellow." The constituent colors in these opponent pairs appeared to cancel each other.

This observation is in contrast to mixtures that do retain features of the constituent colors, such as red and blue (magenta), blue and green (cyan), and red and yellow (orange). Hering also observed patterns in the colors of afterimages; after fixating on a red patch, for example, viewing a white screen yields a green afterimage. All of these observations could not be reconciled within the confines of the Young–Helmholtz theory and led Hering to propose that color vision involved three separate, *spectrally opponent* (*mutually antagonistic*) *channels*: black–white (luminance), red–green, and blue–yellow. In the mid-1950s, a series of influential hue-cancelation experiments carried out by Dorothea Jameson and Leo Hurvich provided psychophysical evidence that supported Hering's observations.

Two-Stage Zone Model

The Young–Helmholtz and Hering theories, it turns out, are not mutually exclusive but rather describe sequential stages of visual-system processing. Trichromatic theory describes the initial stage of color vision, which is governed by the spectral characteristics of the light entering the eye and involves the excitation of the L-, M-, and S-cone outer segments (OS, Fig. 8.4-2); this is the first stage of the two-stage zone model. The subsequent additive and subtractive neural interconnections among the outputs of the cones that are configured in the outer plexiform layer (OPL) implement the opponent channels; this is the second stage of the two-stage zone model.

More specifically, the signals from the three types of cones are selectively added and subtracted to form three *cone-opponent channels* in the OPL, as schematically illustrated in Fig. 8.6-1:

1. A black–white (luminance) channel is formed from the addition of the L- and M-cone responses.
2. A red–green cone-opponent channel is formed from the subtraction of the M-cone response from the L-cone response (or vice-versa).
3. A blue–yellow cone-opponent channel is formed from the subtraction of the sum of the L- and M-cone responses from the S-cone response (or vice versa).

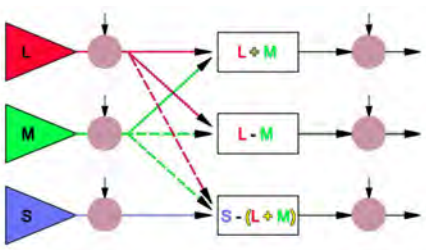


Figure 8.6-1 Schematic of the two-stage zone model. The first stage comprises the OSs of the L-, M-, and S-cones. Additive (solid arrows) and subtractive (dashed arrows) signals are combined to give rise to three channels at the second (OPL) stage: the luminance channel L+M, the red–green cone-opponent channel L–M and the blue–yellow cone-opponent channel S–(L+M). The circles represent generalized gain control that implements adaptation.

The combination of the cone signals, as schematized in Fig. 8.6-1, leads to results that

can be quantitatively illustrated in the spectral domain by making use of the normalized L-, M- and S-cone spectral sensitivity curves displayed in Fig. 8.5-1(a). Subtracting the M curve from the L curve yields the L–M relative spectral sensitivity curve illustrated as red in Fig. 8.6-2(a); this curve exhibits excitation when red is present and suppression when green is present, thus representing opponency in the red–green chromatic channel. Similarly, green–red opponency is represented by the M–L curve shown as green in Fig. 8.6-2(a); this curve reveals excitation when green is present and suppression when red is present. Evidently, suitably balanced mixtures of red and green light can cancel activity on these channels. Analogously, subtracting the sum of the L and M curves from the S curve in Fig. 8.5-1(a) yields the S–(L+M) curve illustrated as blue in Fig. 8.6-2(b); this curve represents opponency in the blue–yellow chromatic channel. Similarly, yellow–blue opponency is represented by the (L+M)–S curve shown as yellow in Fig. 8.6-2(b). Spike-train recordings from color-opponent neurons yield analogous curves, in which the zero level corresponds to the spontaneous firing rate, positive values indicate firing rates above the spontaneous rate, and negative values indicate firing rates suppressed below the spontaneous rate.

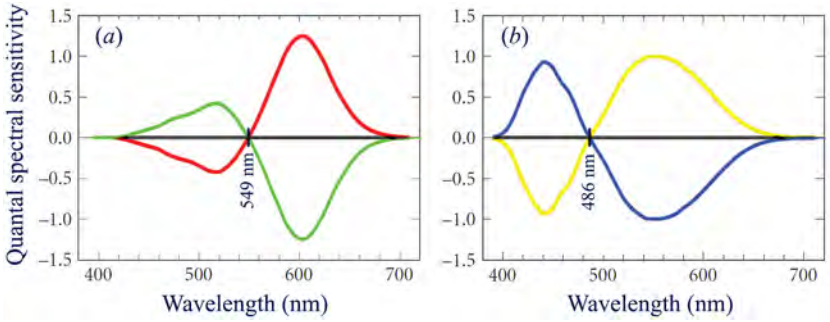


Figure 8.6-2 Relative spectral sensitivities of cone-opponent signals at the output of the second stage of the zone model portrayed in Fig. 8.6-1. (a) Red–green opponency is represented by L–M (red curve) and M–L (green curve). The red curve is positive at longer wavelengths that include red, and negative at shorter wavelengths that include green. (b) Blue–yellow opponency is represented by S–(L+M) (blue curve) and (L+M)–S (yellow curve). The blue curve is positive at shorter wavelengths that include blue, and negative at longer wavelengths that include yellow. The wavelengths at the zero crossings are indicated. Yellow is nominally represented as L+M since it derives from red plus green (Fig. 9.1-2). (Adapted from A. Stockman and D. H. Brainard, Color Vision Mechanisms, in M. Bass, C. DeCusatis, J. Enoch, V. Lakshminarayanan, G. Li, C. Macdonald, V. Mahajan, and E. van Stryland, eds., *Handbook of Optics: Volume III – Vision and Vision Optics*, 3rd ed., McGraw–Hill, Chap. 11, Fig. 4 middle panels, 2010.)

While it is the cones that* RECEIVE *color, it is the visual system that* PERCEIVES *color. Spectrally opponent color channels are silenced in the simultaneous presence of suitably balanced opponent color mixtures.

The information generated in the three opponent channels is ultimately carried from the retina to the lateral geniculate nucleus, and thence to higher visual centers. Red–green information is usually provided to the four dorsal layers of the LGN via the axons of **parvocellular** retinal ganglion cells (also called P, midget, and small-cell RGCs), which are located in the ganglion-cell layer (GCL, portrayed in Fig. 8.4-2). The axons of **magnocellular** retinal ganglion cells (also called M, parasol, and large-cell RGCs) in the GCL, which project to the two ventral layers of the LGN, carry transient and movement information. Certain S-cones in the retina feed **koniocellular** cells, which

reside in the regions surrounding the LGN layers. The processing of color in the cortex is more complex and less-well understood. In the end, however, the range of perceived color gradations is enormous, as illustrated in Fig. 8.7-3.

Separation into Luminance and Chromaticity Components

As indicated above, trichromatic color vision is implemented by the three types of cones that are present in the retina while opponent-channel color vision is implemented by the subsequent neural circuitry that processes the outputs of the cones. The net result is a separation of color into (chromatic) chromaticity components and an (achromatic) luminance component. Transforming the cone signals into opponent signals has the merit that it decorrelates the visual information carried by the cone signals, thereby enhancing signal-transmission efficiency. It will become apparent in the sequel that Grassmann's first law (Sec. 9.2), along with the iconic CIE 1931 XYZ color space (Sec. 9.5), reflect this separation.

8.7 NON-TRICHROMATIC VISION

Although ubiquitous, trichromacy is not a universal feature of human color vision. Other configurations exist, as the following examples attest:

EXAMPLE 8.7-1. ***Color Blindness.*** While humans are ordinarily trichromats, there are numerous variations on the theme of trichromacy. Some observers with **color blindness**, often referred to as **color-vision deficiencies (CVDs)**, have the usual three types of cones but the opsins therein do not function normally; such observers are called anomalous trichromats. Other observers lack one or more types of cones and fall in the category of dichromats or monochromats:

- *Anomalous trichromats* possess three types of functioning cones, as do those with normal vision, but one of their cone types has an anomalous opsin:
  - *Protanomaly* indicates reduced sensitivity to red.
  - *Deuteranomaly* indicates reduced sensitivity to green.
  - *Tritanomaly* indicates reduced sensitivity to blue.
- *Dichromats* have only two types of functioning cones; the third type is absent or impaired:
  - *Protanopia* indicates the absence (or dysfunction) of L-cones.
  - *Deuteranopia* indicates the absence (or dysfunction) of M-cones.
  - *Tritanopia* indicates the absence (or dysfunction) of S-cones.
- *Monochromats* have only one type of functioning photoreceptor. Although they are totally color blind in accordance with the principle of univariance discussed above, they can nevertheless distinguish some color samples in color-matching experiments by using their past experience to associate perceived brightness with semantic color identifiers:
  - *Cone monochromats* have only one type of functioning cone.
  - *Rod monochromats*, also called *achromats*, are devoid of functioning cones but so have functioning rods and are therefore not blind.

In comparison with observers endowed with normal vision, those with color blindness are more likely to conflate different colors in color-matching experiments. The color wheels depicted in Fig. 8.7-1 illustrate that CVD observers can perceive a limited number of spectral hues, or in the case of monochromats, none at all. The precise nature of a color-vision abnormality depends on the number of cone types involved and the extent of their dysfunction.

Females are less prone to color-vision deficiencies than males because the genes that give rise to the L- and M-cone opsins reside on the X-chromosome. Females have two such chromosomes, providing them with redundancy against an anomaly, whereas males have only one. Indeed, only about 0.4% of females are affected by CVDs, while about 8.5% of males are (the most common CVD is deuteranomaly, which affects about 5% of males). The exception that proves the rule is tritanopia: the gene that codes for the production of the S-cone opsin does not reside on the X-chromosome and both females and males do indeed have the same incidence of CVDs.

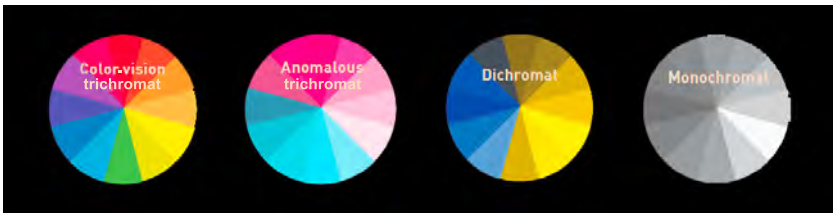


Figure 8.7-1 *Left to Right*: Color wheel as seen by: a normal color-vision trichromat; a deuteranomalous trichromat; a protanopic dichromat; and a monochromat. The deficiencies are defined in the text.

EXAMPLE 8.7-2. ***Tetrachromatic Human Vision.*** Some human observers have color-perception experiences that are superior to those of trichromats. *Retinal tetrachromats* possess opsin genotypes that lead to four, rather than three, types of retinal cones, which offers an enhanced visual space. The perceptual capabilities of tetrachromats have been empirically compared with those of trichromats, and it has been shown that tetrachromats benefit from a wider range of certain color-perception variations, some of which can be accommodated within standard trichromatic theory. Still, because the equipment and techniques used for conducting such experiments are geared to trichromats, they may well be suboptimal for discovering unknown salutary features of tetrachromatic vision.

Concetta Antico, an Australian-American artist who is a tetrachromat, is pictured in Fig. 8.7-2. In principle, she can perceive about 1.5 times the number of color gradations than can a trichromat, but it is difficult to quantify the range of advantages she possesses. Although tetrachromacy is not uncommon in the human population as a whole, females are more likely to possess this enhancement than males. As discussed in connection with Example 8.7-1, this is because the genotypes for the L- and M-cone opsins reside on the X-chromosome; females have two such chromosomes, which enables them to host a different variant on each. More than 15% of human females are tetrachromats but few recognize that they are endowed with this special competence.

Figure 8.7-2 Color expert and artist Concetta Antico is a tetrachromat and can perceive more color gradations than a normal trichromat. More than 15% of human females are tetrachromats, but few recognize that they have this capability.

EXAMPLE 8.7-3. ***Color Vision in the Animal Kingdom.*** While humans are generally trichromats, this is not the default in the animal kingdom. The number of cone types in the retina depends on the species and, as displayed in Fig. 8.7-3, animals can be monochromats, dichromats, trichromats, tetrachromats, or even multichromats (e.g., mantis shrimp, not shown). In analogy with the enhanced incidence of tetrachromats among human females discussed in Example 8.7-2, male squirrel monkeys and marmosets are dichromats while most females of those species are trichromats. The greater the number of cone types, the larger the number of color gradations that can be discerned. Human trichromats are estimated to be able to discern 2 million color gradations, comprising 500 luminance, 200 hue, and 20 saturation gradations. The principle of univariance dictates that the perception of hue relies on the presence of two or more cone types; as an order-of-magnitude estimate, therefore, dichromats and tetrachromats can perhaps discern 100 and 300 hue gradations, respectively.

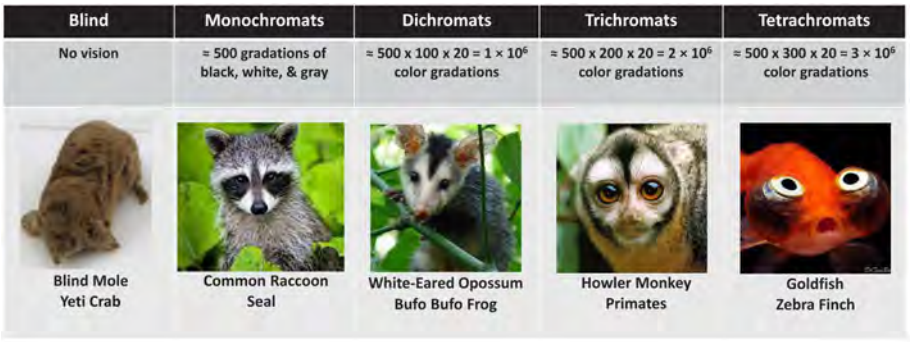


Figure 8.7-3 Animals can be monochromats, dichromats, trichromats, or tetrachromats. The number of discernible color gradations increases with the number of cone types with which a species is endowed. A substantial majority of humans are trichromats.

8.8 RADIOMETRIC AND PHOTOMETRIC UNITS

Radiometric Units

Radiometric units characterize the strength of a source of electromagnetic radiation, such as light or infrared radiation, in terms of its physical properties. In the context of wave optics, as described in Sec. 2.1, the simplest examples of such units are the optical energy E, optical power P, and optical intensity (irradiance) I. Analogous examples from the perspective of photon optics, provided in Sec. 3.4, are the mean photon number $\overline{n}$, mean photon flux Φ, and mean photon-flux density ϕ.

The radiometric measures most often employed in vision science and lighting technology are represented in the left-hand columns of Table 8.8-1. These include: *radiant energy* (J), which is the total energy emitted by a point source in all directions; *radiant flux* (W), which represents the radiant energy per unit time and corresponds to the optical power; *radiant intensity* (W/sr), which is the power radiated per unit solid angle (sr) about the source (radiant intensity should not be confused with optical intensity); *irradiance* (W/m^2), commonly referred to as intensity, which is the power per unit area; and *radiance* (W/m^2-sr), which is the radiant flux emerging from, or incident on, an extended surface, per unit area of that surface, per unit solid angle. The radiance is conserved in ray optics since both the irradiance and solid angle decrease as the inverse square of distance; it is generally used for characterizing optical systems whose acceptance solid angles and apertures are limited. The irradiance, on the other hand, is more often used for characterizing optical systems that deliver light to large areas.

Spectral radiometric measures are also widely used for characterizing broadband sources. Wavelength-based examples of spectral measures include the spectral radiance L_λ, spectral irradiance I_λ, and spectral density S_λ of blackbody and graybody radiation, as discussed in Sec. 9.7. Frequency-based examples of spectral measures include the spectral irradiance I_ν, spectral radiant flux P_ν, and spectral radiant energy E_ν introduced in Sec. 3.4.

Photometric Units

Photometric units incorporate the effectiveness of a source of visible light in exciting the human visual system. Photometric units are designated by use of the subscript 'V' (for visual). The photometric measures most often employed in vision science and lighting technology are specified in the right-hand columns of Table 8.8-1. The photometric counterparts of each of the radiometric measures set forth in the left-hand columns of the table are: the *luminous energy* E_V (lm-s) is analogous to the radiant energy (J); the

Table 8.8-1 Common radiometric and photometric measures used in lighting technology.

RADIOMETRIC[a]			PHOTOMETRIC[b]		
Radiant energy	E	(J or W-s)	Luminous energy	$E_{\mathtt{V}}$	(lm-s)
Radiant flux	P	(W)	Luminous flux[c]	$P_{\mathtt{V}}$	(lm)[d]
Radiant intensity	$\mathcal{I}$	(W/sr)	Luminous intensity	$\mathcal{I}_{\mathtt{V}}$	(cd)[e]
Irradiance	I	(W/m^2)	Illuminance	$I_{\mathtt{V}}$	(lx)[f,g]
Radiance	L	(W/m^2-sr)	Luminance	$L_{\mathtt{V}}$	(cd/m^2)[h,i]

[a] Broadband sources are often also characterized by spectral measures in radiometry.
[b] Photometric units are subscripted with '$\mathtt{V}$' (for visual). [c] Luminous flux is also called luminous power.
[d] The abbreviation for lumen is lm. [e] One candela, abbreviated cd, is one lm/sr.
[f] One lux, abbreviated lx, is one lm/m^2. [g] When the light is emitted from a surface, the illuminance is also called the luminous exitance (or luminous emittance) $M_{\mathtt{V}}$.
[h] One cd/m^2 signifies one lm/m^2-sr and is equivalent to 0.3142 millilambert and to 0.2919 foot-lambert.
[i] As discussed in the sequel, the brightness $B_{\mathtt{V}}$, which is the psychophysical magnitude estimate of the luminance, increases with $L_{\mathtt{V}}$ in fractional power-law fashion [see (8.8-7)].

luminous flux $P_{\mathtt{V}}$ (lm) corresponds to the radiant flux (W); the *luminous intensity* $\mathcal{I}_{\mathtt{V}}$ (cd) has as its counterpart the radiant intensity (W/sr); the *illuminance* $I_{\mathtt{V}}$ (lx) is associated with the irradiance (W/m^2) and assumes that the light is incident on a surface [the *luminous exitance* or *luminous emittance* $M_{\mathtt{V}}$ assumes that the light is emitted from a surface and has the same units (lm/m^2)]; and the *luminance* $L_{\mathtt{V}}$ (cd/m^2) corresponds to the radiance (W/m^2-sr). The solid angle of interest is often that subtended by the pupil of the eye. Just as the radiance is geometrically invariant in ray optics, so too is the luminance.

Historically, the standard source of light used in photometric measurements was a candle manufactured to certain specifications; this later evolved into a blackbody radiator at 2042 K (the freezing point of platinum at atmospheric pressure). Subsequently, photometric units were mathematically related to their radiometric counterparts by invoking a *standard observer*, which obviated the necessity of having to establish photometric values by the direct visual observation of a stimulus by an individual observer.

In *photopic photometry*, the connection between the photometric and radiometric measures provided in Table 8.8-1 is centered on the photopic luminous efficiency function $V(\lambda_0)$ displayed in Fig. 8.5-3. This function represents the relative effectiveness of light of different wavelengths in stimulating the photopic visual system, as discussed in Sec. 8.5; it was established as a standard in the CIE 1924 photometry system, as indicated in the introduction to Chapter 9.

Photometric measures are related to their radiometric counterparts via inner products of the spectral versions of the radiometric measures and the photopic luminous efficiency function $V(\lambda_0)$. The integration is over wavelength, so the inner products maintain their form for all spatial variants of these quantities, including luminous flux, luminous intensity, illuminance, and luminance.

Luminous Flux

The **luminous flux** $P_{\mathtt{V}}$ (lm) is proportional to the inner product of the wavelength-based power spectral density of the incident light $S_\lambda(\lambda_0)$ and the photopic luminous efficiency function $V(\lambda_0)$. The integration is carried out over the range of visible wavelengths, which is conventionally taken to extend from 380 to 780 nm. The luminous flux is

therefore expressed as

$$P_{\mathrm{V}} = 683 \int_{380}^{780} S_{\lambda}(\lambda_0)\, V(\lambda_0)\, d\lambda_0 \,. \tag{8.8-1}$$

Luminous Flux (lm)

The constant of proportionality* 683.002 lm/W ≈ 683 lm/W *linking photometric and radiometric measures reconciles modern and earlier definitions of the candela.

For the special case of a monochromatic source of light of wavelength λ_0, the spectral density $S_{\lambda}(\lambda_0)$ is a delta function of area P_0, so that (8.8-1) can be written as $P_{\mathrm{V}} = 683 \int P_0\, \delta(\lambda - \lambda_0)\, V(\lambda)\, d\lambda$, where λ serves as a dummy variable. With the help of the sifting property of the delta function in the integrand, we therefore obtain

$$P_{\mathrm{V}} = 683\, P_0\, V(\lambda_0) \,. \tag{8.8-2}$$

Luminous Flux (lm)
(Monochromatic Source)

Luminous Intensity

If the visible light emitted by a source takes the form of a well-defined cone of **full vertex angle** (or **radiation angle**) 2θ, where θ is the angle measured from the emission-plane normal, the **luminous intensity** I_{V} may be expressed in terms of the luminous flux P_{V} as

$$I_{\mathrm{V}} = \frac{P_{\mathrm{V}}}{2\pi(1 - \cos\theta)} \,. \tag{8.8-3}$$

The denominator of (8.8-3) represents the solid angle Ω subtended by the cone, which is the area of the spherical cap atop the cone on a unit sphere.

EXAMPLE 8.8-1. ***Luminous Intensity as a Function of Radiation Angle.*** A light source that emits 300 lm into a cone of radiation angle $2\theta = 120°$ (which corresponds to the 50%-power angle of a Lambertian radiator) produces a luminous intensity $I_{\mathrm{V}} = P_{\mathrm{V}}/\pi \approx 95.5$ cd. For a source that emits uniformly in all directions, $2\theta = 360°$ and $I_{\mathrm{V}} = P_{\mathrm{V}}/4\pi$. At the opposite extreme, when the source emits into a sufficiently small radiation angle, such that $\cos\theta \approx 1 - \theta^2/2$, (8.8-3) reduces to $I_{\mathrm{V}} \approx P_{\mathrm{V}}/\pi\theta^2$. In that case, if $P_{\mathrm{V}} = 300$ lm and the radiation angle is $2\theta = 20°$, we have $\theta = 10° \approx 0.1745$ rad so that $I_{\mathrm{V}} \approx P_{\mathrm{V}}/\pi\theta^2 = 3135$ cd.

Luminance and Illuminance

Luminance. As specified in Table 8.8-1, the **luminance** L_{V}, whose units are cd/m^2 (or equivalently lm/m^2-sr), is the photometric counterpart of the **radiance** L, whose units are W/m^2-sr. The relevant area is the projected area A seen by the observer and the relevant solid angle Ω is often that subtended by the pupil of the eye (Example 8.8-2). In analogy with (8.8-1), the luminance is proportional to the inner product of the wavelength-based spectral radiance of the incident light and the photopic luminous

efficiency function over the range of visible wavelengths. The proportionality constant is again fixed at 683 lm/W, which gives rise to

$$L_{\mathrm{v}} = 683 \int_{380}^{780} L_\lambda(\lambda_0)\, V(\lambda_0)\, d\lambda_0 \,. \tag{8.8-4}$$

Luminance (cd/m^2)

For the special case of a monochromatic source of wavelength λ_0, the spectral radiance $L_\lambda(\lambda_0)$ is a delta function of area L, where L is the radiance, so that the sifting property of the delta function in the integrand converts (8.8-4) to

$$L_{\mathrm{v}} = 683\, L\, V(\lambda_0) \,. \tag{8.8-5}$$

Luminance (cd/m^2) (Monochromatic Source)

Luminance of a Lambertian Radiator. The luminance of a Lambertian radiator, such as a source of blackbody radiation, is independent of the angle at which it is viewed since both the intensity of the source and the projected area are proportional to the cosine of the angle from the emission-plane normal [Fig. 7.2-2(a)]. Most flat-surface sources, exit pupils of illuminating optical systems, and diffusely reflecting surfaces behave approximately as Lambertian radiators. Emitters that behave as point sources, in contrast, are isotropic radiators.

Illuminance. In accordance with Table 8.8-1, the **illuminance** I_{v}, with units of lm/m^2 (or equivalently lx), is the photometric counterpart of the **irradiance** I, with units of W/m^2. Again, in analogy with (8.8-1), the illuminance is given by

$$I_{\mathrm{v}} = 683 \int_{380}^{780} I_\lambda(\lambda_0)\, V(\lambda_0)\, d\lambda_0 \,. \tag{8.8-6}$$

Illuminance (lx)

EXAMPLE 8.8-2. ***Solid Angle Subtended by the Pupil of the Human Eye.*** Since a circle has 2π radians and 360°, conversion between the two takes the form $\theta\,(\text{rad}) = (2\pi/360)\,\theta\,(\text{deg})$. The monocular horizontal and vertical fields-of-view subtended by the pupil are known to be $\theta_x \approx 160°$ and $\theta_y \approx 135°$, corresponding to $\theta_x \approx 2\pi 160/360 = 8\pi/9$ rad and $\theta_y \approx 2\pi 135/360 = 3\pi/4$ rad, respectively. The solid angle subtended by the pupil can thus be estimated by constructing a rectangular pyramid with its apex at the pupil, which subtends $\Omega = 4\sin^{-1}[\sin(\theta_x/2)\sin(\theta_y/2)]$ sr. Inserting the values provided above into this formula leads to $\Omega \approx 4\sin^{-1}[\sin(4\pi/9)\sin(3\pi/8)] \approx 4.57$ sr. For a full hemisphere, the fields-of-view are $\theta_x = \theta_y = \pi$ rad, which yields $\Omega \approx 4\sin^{-1}[\sin(\pi/2)\sin(\pi/2)] = 4\sin^{-1}[1] = 2\pi$ sr, as expected.

EXAMPLE 8.8-3. ***Natural-Light Luminance and Illuminance Levels.*** Typical values of the photometric luminance L_{v} and illuminance I_{v} for a number of common sources of natural light are set forth in Table 8.8-2, assuming that the intercepted solid angle $\Omega = 1/2$ sr. The radiometric irradiance (referred to as the intensity in the optics and photonics literature) for some of these same sources of light is reported in Table 3.4-1.

EXAMPLE 8.8-4. ***Luminance Levels for a Tea-Light Candle.*** By way of illustration, an image of a tea-light candle viewed with a luminance camera is portrayed in Fig. 8.8-1. As with the palette of false colors used to represent temperature in thermography (Fig. 4.8-2), the false colors

Table 8.8-2 Luminance and illuminance for various sources of natural light.

SOURCE	Luminance L_v (cd/m^2)	Illuminance I_v (lx)
Starlight	2×10^{-2}	1×10^{-2}
Moonlight	2×10^{-1}	1×10^{-1}
Nighttime	2×10^{0}	1×10^{0}
Twilight	2×10^{1}	1×10^{1}
Dark day	2×10^{2}	1×10^{2}
Overcast daylight	2×10^{3}	1×10^{3}
Bright daylight	2×10^{4}	1×10^{4}
Direct sunlight	2×10^{5}	1×10^{5}

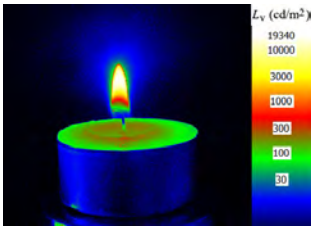


Figure 8.8-1 Image of a tea-light candle, against a black background, viewed with a luminance camera manufactured by Technoteam Bildverarbeitung GmbH. The various luminance levels (cd/m^2), coded as false colors, are indicated in the key on the right-hand side of the figure. (Adapted from an image by Anders Thorseth, 24 May 2019, via Wikimedia Commons.)

used to display luminance levels in Fig. 8.8-1 are chosen arbitrarily. Furthermore, neither bears any relationship to the real colors associated with the color temperatures displayed in Fig. 9.7-1(*b*).

Brightness

Magnitude estimation is a psychophysical paradigm in which an observer is asked to judge the magnitudes of a series of stimuli of different strengths and to assign to them numerical values, say on a scale between 0 and 100. Magnitude-estimation experiments using stimuli of many different modalities were pioneered by S. S. Stevens in the 1950s. The technique has long been known to be highly reliable when the measurements are carefully conducted and properly analyzed.

In visual magnitude-estimation experiments, observers are instructed to view spots of light and to estimate their **brightness** B_v as the luminance L_v is adjusted over a broad range of values. Brightness is an attribute of visual perception according to which an area appears to emit, transmit, or reflect, more or less light (Sec. 9.4). It is a psychophysical variable that depends not only on the responses of the photoreceptors to a stimulus, but also on the neural processing carried out at higher visual centers in the brain.

Under suitably constrained stimulus and observation conditions, the empirical relation that emerges from such experiments is a brightness–luminance formula that obeys a fractional power-law form,

$$B_v = b\left(\frac{L_v}{L_0} - 1\right)^{\beta}, \tag{8.8-7}$$

Brightness–Luminance Relation

where B_v is the perceived brightness, b and L_0 are constants, and L_v is the stimulus luminance. The power-law exponent β typically lies in the range

$$1/4 \leqslant \beta \leqslant 1/2\,, \tag{8.8-8}$$

Brightness Exponent

so that it represents compressive behavior. For circular patches of light subtending a few degrees of visual angle, observations are often consistent with $\beta \approx 1/3$, whereas for point sources they are frequently closer to $\beta \approx 1/2$. Since luminance is invariant to position, so too is brightness.

Brightness and luminance are often conflated. Brightness is a psychophysical variable that relies on the responses of photoreceptors and higher-level neural circuitry, whereas luminance is simply the photometric counterpart of radiance.

The psychophysical brightness–luminance relation presented in (8.8-7) is generally applicable only in the absence of spatial contrast, i.e., for isolated, self-luminous sources presented on a dark background. The observer is assumed to be adapted to dark or to a steady luminance. Brightness enables object recognition over a vast range of diurnal illumination levels (Sec. 8.5). Indeed, many color spaces, such as sRGB, implement an expansive power-law nonlinearity at large values of the luminance, called the **gamma correction**, to compensate for the intrinsic compressive power-law nonlinearity represented in (8.8-7). The viewing situation described above is substantially different from that of normal viewing, where vision is dominated by contrast estimation (i.e., by differences in light level).

EXAMPLE 8.8-5. ***Brightness Ratio for Monochromatic Sources in Photopic Vision.*** In the domain of photopic vision, L_v/L_0 in (8.8-7) is generally $\gg 1$. Using (8.8-5), and assuming that $\beta \approx 1/3$, the brightness–luminance relation provided in (8.8-7) can then be approximated by $B_\mathrm{v}(\lambda_0) \approx b\sqrt[3]{L_\mathrm{v}(\lambda_0)/L_0} = b\sqrt[3]{683\,L/L_0}\,\sqrt[3]{V(\lambda_0)}$. The ratio of perceived brightnesses for two monochromatic sources of wavelengths λ_1 and λ_2, assuming they have the same radiance L and roughly the same proportionality constant b, are then expected to obey

$$\frac{B_\mathrm{v}(\lambda_1)}{B_\mathrm{v}(\lambda_2)} \approx \sqrt[3]{\frac{V(\lambda_1)}{V(\lambda_2)}}, \tag{8.8-9}$$

Brightness Ratio (Monochromatic Sources)

for both simultaneous and sequential viewing.

8.9 LUMINOUS EFFICACY AND EFFICIENCY

Several quantities are commonly used to characterize the performance of light sources for illumination. Among these are measures of *luminous efficacy*, which carry units; and measures of *luminous efficiency*, which are unitless and assume values that range from zero to unity. In this section we compare and contrast the following quantities:

- The luminous efficacy of radiation (LER) (lm/W).
- The wall-plug luminous efficacy (WPE) (lm/W or LPW).
- The wall-plug luminous efficiency (WPC).
- The current luminous efficacy (CLE) (cd/A).

It is particularly important to distinguish between the luminous efficacy of radiation and the more commonly used wall-plug luminous efficacy. As discussed in some detail,

although these quantities are distinct they are often conflated since their appellations are similar and both carry units of lm/W.

Luminous Efficacy of Radiation (LER)

The **luminous efficacy of radiation (LER)** is defined as the ratio of the luminous flux to the radiant flux of a source, or equivalently, as the ratio of the luminance to the radiance,

$$\eta_{\mathtt{LER}} = \frac{P_{\mathtt{V}}}{P_{\mathtt{0}}} = \frac{L_{\mathtt{V}}}{L}\,, \tag{8.9-1}$$

Luminous Efficacy of Radiation (lm/W)

where $P_{\mathtt{V}}$ and $L_{\mathtt{V}}$ are defined in (8.8-1) and (8.8-4), respectively. The LER is a measure of the efficacy of a source of light in exciting the photopic visual system and, as such, converts watts to lumens.

Given the spectral density of an arbitrary source of light $S_\lambda(\lambda_{\mathtt{0}})$, it is straightforward to determine the LER: divide both sides of (8.8-1) by $P_{\mathtt{0}}$ and calculate $\eta_{\mathtt{LER}} = P_{\mathtt{V}}/P_{\mathtt{0}}$ using the normalized spectral density $S_\lambda(\lambda_{\mathtt{0}})/P_{\mathtt{0}}$. For the special case of monochromatic light of wavelength $\lambda_{\mathtt{0}}$, inserting (8.8-2) in (8.9-1) yields

$$\eta_{\mathtt{LER}} = 683\,V(\lambda_{\mathtt{0}})\ \ \text{lm/W}. \tag{8.9-2}$$

Luminous Efficacy of Radiation (Monochromatic Source)

In particular, for a monochromatic source at $\lambda_{\mathtt{0}} = 555$ nm, where $V(\lambda_{\mathtt{0}})$ assumes its maximum value of unity (Fig. 8.5-3), (8.9-2) becomes

$$\eta_{\mathtt{LER}}^{\mathtt{MAX}} = 683\ \ \text{lm/W}. \tag{8.9-3}$$

***For a monochromatic light source of wavelength* 555 nm*, a luminous flux of* 1 lm *corresponds to a radiant flux of* $^1/_{683}$ W*, i.e., to* 1.464 mW.**

Wall-Plug Luminous Efficacy (WPE)

The **wall-plug luminous efficacy (WPE)**, also called the **overall luminous efficacy** and the **luminous efficacy of the source**, is defined as

$$\eta_{\mathtt{WPE}} = \frac{P_{\mathtt{V}}}{P_{\mathtt{EL}}} = \frac{P_{\mathtt{V}}}{iV}\,, \tag{8.9-4}$$

Wall-Plug Luminous Efficacy (lm/W or LPW)

where $P_{\mathtt{V}}$ is the luminous flux defined in (8.8-1) and $P_{\mathtt{EL}} = iV$ is the electrical drive power supplied to the light-emitting device, as provided in (7.1-13); the quantities i and V are the drive current and drive voltage, respectively.

Since the WPE is the ratio of the emitted optical power to the electrical power feeding the device, and the LER is the ratio of the luminous flux to the emitted optical power, the two luminous efficacies are related by combining (7.1-14), (8.9-1), and (8.9-4):

$$\eta_{\mathtt{WPE}} = \eta_{\mathtt{PCE}}\,\eta_{\mathtt{LER}}\,. \tag{8.9-5}$$

Relation of Luminous Efficacies

While the PCE has units of W/W and is thus dimensionless, the LER has units of lm/W, and hence so too does the WPE.

Two distinct definitions of luminous efficacy exist, though both have units of* lm/W: *1) The* WALL-PLUG LUMINOUS EFFICACY (WPE)*, which is the ratio of the luminous flux to the electrical power that feeds a light-emitting device, and 2) The* LUMINOUS EFFICACY OF RADIATION (LER)*, which is the ratio of the luminous flux to the optical power emitted by the device. The* WPE *is the more comprehensive of the two measures since it is a concatenation of the* LER *with the* POWER-CONVERSION EFFICIENCY (PCE)*, which is the ratio of the optical power emitted by the device to the electrical power driving it.

Moreover, since $\eta_{\mathtt{PCE}} \leqslant 1$, it is evident that

$$\eta_{\mathtt{WPE}} \leqslant \eta_{\mathtt{LER}}. \tag{8.9-6}$$

In the particular case of monochromatic light, combining (8.9-2) with (8.9-5) leads to

$$\boxed{\eta_{\mathtt{WPE}} \approx 683\, \eta_{\mathtt{PCE}} V(\lambda_0) \ \text{lm/W}.} \tag{8.9-7}$$

Wall-Plug Luminous Efficacy (Monochromatic Light)

Hence, the maximum value of the wall-plug luminous efficacy, $\eta_{\mathtt{WPE}}^{\mathtt{MAX}}$, is attained when the light-emitting device has unity power-conversion efficiency ($\eta_{\mathtt{PCE}} = 1$) and emits light at 555 nm ($V = 1$), which leads to

$$\eta_{\mathtt{WPE}}^{\mathtt{MAX}} = 683 \ \text{lm/W}. \tag{8.9-8}$$

Wall-Plug Luminous Efficiency (WPC)

The term luminous efficiency is also in common use. The **wall-plug luminous efficiency**, also known as the **wall-plug luminous coefficient (WPC)**, is defined as the wall-plug luminous efficacy normalized to its maximum possible value of 683, as provided in (8.9-8):

$$\boxed{\eta_{\mathtt{WPC}} = \frac{\eta_{\mathtt{WPE}}}{\eta_{\mathtt{WPE}}^{\mathtt{MAX}}} = \frac{\eta_{\mathtt{WPE}}}{683}.} \tag{8.9-9}$$

Wall-Plug Luminous Efficiency

This dimensionless quantity has a value that lies between zero and unity, and is the photometric counterpart of the radiometric power-conversion efficiency $\eta_{\mathtt{PCE}}$. For monochromatic light, combining (8.9-7) and (8.9-9) yields the simple formula

$$\boxed{\eta_{\mathtt{WPC}} \approx \eta_{\mathtt{PCE}} V(\lambda_0).} \tag{8.9-10}$$

Wall-Plug Luminous Efficiency (Monochromatic Light)

As indicated earlier, the various terms used in lighting technology are often conflated so that efficiencies are sometimes (improperly) expressed in lm/W and efficacies are (improperly) expressed as dimensionless fractions.

Numerical values for the wall-plug luminous efficacy and wall-plug luminous efficiency, along with a number of other measures, are provided in Table 11.9-1 for all

manner of light sources. Quantities analogous to those set forth in (8.9-4) and (8.9-9) are also defined for light sources contained within housings that introduce losses of their own. Known as the *luminaire wall-plug luminous efficacy* $\eta_{\mathtt{LUM}}$ and the *luminaire wall-plug luminous efficiency* $\eta_{\mathtt{LUC}}$, these quantities are defined in (11.6-1) and (11.6-3), respectively.

Current Luminous Efficacy (CLE)

A measure that is sometimes used in the technical characterization of light-emitting devices for illumination is the **current luminous efficacy (CLE)**, which is also called the **current efficiency** (see, e.g., Table 7.6-1). The CLE is defined as

$$\eta_{\mathtt{CLE}} = \frac{I_{\mathtt{V}}}{i} = \frac{L_{\mathtt{V}}}{J}, \tag{8.9-11}$$

Current Luminous Efficacy (cd/A)

where $I_{\mathtt{V}}$ and $L_{\mathtt{V}}$ are, respectively, the luminous intensity (cd) and luminance (cd/m^2) of the source, as specified in Table 8.8-1; and i (A) and J (A/m^2) are, respectively, the electrical current and electrical current density interior to the device, which are related by (7.1-1). While the CLE bears some similarity to the WPE defined in (8.9-4), it represents the ratio of different photometric and electrical quantities, and is not nearly as widely used.

EXAMPLE 8.9-1. ***Luminous Efficacy of Radiation for a Laser Pointer.*** Laser pointers emit nearly monochromatic light at various wavelengths that correspond to fully saturated spectral colors. Consider three such devices, constructed as follows:

- *Violet laser pointer:* An InGaN laser diode that emits light at 405 nm.
- *Green laser pointer:* A frequency-doubled YVO_4:Nd^{3+} laser chip pumped by an 808-nm AlGaAs laser diode that emits light at 532 nm.
- *Red laser pointer:* An AlInGaP laser diode that emits light at 650 nm.

The values of the luminous efficiency of radiation $V(\lambda_0)$ at these wavelengths are determined by referring to the photopic luminous efficiency function presented in Fig. 8.5-3. For convenience, this graph is reproduced in Fig. 8.9-1, where it is augmented by photos of the three laser-pointer spots at their appropriate wavelengths.

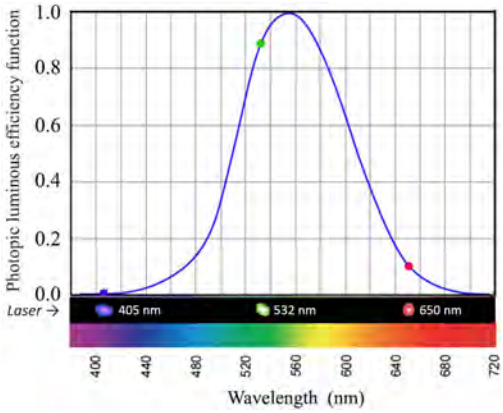


Figure 8.9-1 Photopic luminous efficiency function $V(\lambda_0)$ vs. free-space wavelength λ_0. This plot mimics the one provided in Fig. 8.5-3, differing only in that it includes photographs of the reflected light spots generated by the violet, green, and red laser pointers, which are situated at their corresponding wavelengths on the abscissa. Their values of $V(\lambda_0)$ differ substantially: $V(405\text{ nm}) = 0.0008$, $V(532\text{ nm}) = 0.8804$, and $V(650\text{ nm}) = 0.1070$. In accordance with the discussions provided in Examples 8.8-5 and 8.9-3, the green spot is brighter than the red one, which in turn is brighter than the violet one.

These three values of $V(\lambda_0)$ are also provided in Table 8.9-1. Since the radiant flux (optical power) emitted by all three devices is $P_0 = 2$ mW, it is straightforward to determine the photon flux Φ, the

luminous efficacy of radiation η_{LER}, and the luminous flux P_{V}, all of which are specified in the table. Although the green laser pointer emits a radiant flux of only $P_0 = 2$ mW, its luminous flux P_{V} is > 1 lm since its wavelength falls near the 555-nm peak of the $V(\lambda_0)$ curve, which is in the yellowish-green region.

Table 8.9-1 The parameter values for the nearly monochromatic light emitted by violet, green, and red laser pointers are useful for highlighting the distinction between radiant flux and luminous flux. Successive columns in the table represent: color, wavelength λ_0 (nm), laser material, radiant flux (optical power) P_0 (W), photon flux $\Phi_0 = P_0/h\nu = \lambda_0 P_0/hc_0$ (s^{-1}), photopic luminous efficiency function $V(\lambda_0)$ from Fig. 8.9-1 (dimensionless), luminous efficacy of radiation $\eta_{\text{LER}} = 683\,V(\lambda_0)$ from (8.9-2) (lm/W), and luminous flux $P_{\text{V}} = \eta_{\text{LER}} P_0$ (lm).

COLOR	λ_0	Laser Material	P_0	Φ_0	$V(\lambda_0)$	$\eta_{\text{LER}} = 683\,V(\lambda_0)$	$P_{\text{V}} = \eta_{\text{LER}} P_0$
Violet	405	InGaN	0.002	4.1×10^{15}	0.0008	0.5464	0.001
Green	532	YVO_4:Nd^{3+}	0.002	5.3×10^{15}	0.8804	601.3	1.203
Red	650	AlInGaP	0.002	6.5×10^{15}	0.1070	73.08	0.146

EXAMPLE 8.9-2. ***Efficiencies and Efficacies for a Red Laser Pointer.*** To highlight the distinctions among the various definitions of efficiency and efficacy, and related parameters used in illumination engineering, the following parameter estimates are provided for a 2-mW AlInGaP red laser pointer operated at $\lambda_0 = 650$ nm.

- The *radiant flux* (or *optical power*) $P_0 = 2$ mW is generated by supplying the device with an *electrical current* $i = 20$ mA at an *electrical voltage* $V = 2.0$ V, corresponding to an *electrical drive power* $P_{\text{EL}} = iV = 40$ mW, as specified in (7.1-13).
- The *power-conversion efficiency* is $\eta_{\text{PCE}} = P_0/P_{\text{EL}} = 0.05$, in accordance with the definition provided in (7.1-14).
- The *photopic luminous efficiency function* assumes the value $V(\lambda_0 = 650\text{ nm}) \approx 0.1070$, as displayed in Fig. 8.9-1 of Example 8.9-1.
- The *luminous efficacy of radiation* is $\eta_{\text{LER}} = 683 \cdot V(650\text{ nm}) = 683 \times 0.1070 \approx 73$ lm/W, as provided in Table 8.9-1 of Example 8.9-1.
- The *wall-plug luminous efficacy* is $\eta_{\text{WPE}} = \eta_{\text{PCE}}\,\eta_{\text{LER}} = 0.05 \times 73 = 3.65$ lm/W, in accordance with (8.9-4) and (8.9-5).
- The *wall-plug luminous efficiency* is given by $\eta_{\text{WPC}} = \eta_{\text{WPE}}/683 = 0.0053$, following (8.9-9).
- The *luminous flux* corresponding to the radiant flux $P_0 = 2$ mW is determined via the relation $P_{\text{V}} = \eta_{\text{LER}} P_0 = 0.146$ lm, in accordance with (8.9-1) and Table 8.9-1 of Example 8.9-1.
- The *luminous intensity* for a source of luminous flux P_{V} emitting into a small radiation angle 2θ was established in Example 8.8-1 to be $I_{\text{V}} \approx P_{\text{V}}/\pi\theta^2$; in particular for $P_{\text{V}} = 0.146$ lm and $\theta \approx 10^{-3}$ rad, we arrive at $I_{\text{V}} = (0.146/\pi) \times 10^6 \approx 46000$ cd.
- The *current luminous efficacy* for a source of luminous intensity $I_{\text{V}} \approx 46000$ cd generated by a drive current of 0.02 A is $\eta_{\text{CLE}} \approx 2.3 \times 10^6$ cd/A, in accordance with (8.9-11).

To be clear, although both the luminous efficacy of radiation (LER) and the wall-plug luminous efficacy (WPE) both have units of lm/W, their values can differ substantially: in the example at hand, $\eta_{\text{LER}} = 73$ lm/W whereas $\eta_{\text{WPE}} = 3.65$ lm/W. Indeed, (8.9-6) mandates that $\eta_{\text{WPE}} \leqslant \eta_{\text{LER}}$, which follows from the fact that the WPE is a concatenation of the LER and the power-conversion efficiency (PCE), as specified in (8.9-5), and $\eta_{\text{PCE}} \leqslant 1$.

EXAMPLE 8.9-3. ***Brightness Ratios for Light from Different-Color Laser Pointers.*** The validity of the formula for the brightness ratio (8.8-9) is qualitatively supported by making use of the photographs of the reflected violet, green, and red laser-pointer spots portrayed in Example 8.9-1.

The *experiment* was conducted as follows:

1. The light from the violet, green, and red laser pointers, each with a radiant flux $P_0 = 2$ mW, impinged on a matte-black foam board at near normal incidence.

2. The diffusely reflecting black board had a wavelength-independent reflectance $\mathcal{R} \approx 0.1$ and behaved as a Lambertian reflector, so that the luminance was invariant to the observer's angle of view.
3. The laser spots were viewed with the naked eye. The optics of the eye guided the light to the retina and established the manner in which it was distributed over solid angle and area. This enabled the radiance at the retina L to be determined from the radiant flux P_0.
4. The laser spots were photographed on reflection from the matte-black foam board with an iPhone 13 Pro camera that made use of a 12.2-Megapixel, visible/NIR, Sony IMX703 CMOS array sensor. The camera was outfitted with an external K&F series-B variable neutral-density (ND) filter that served to reduce the photon flux sufficiently so the array sensor operated within its linear regime. The filter was adjustable over the range $2 \leqslant \text{ND} \leqslant 400$, corresponding to a radiant-flux transmittance adjustable over the range $0.0025 \leqslant \mathcal{T} \leqslant 0.5$. This is an *open system*.
5. Both the camera and the eye are responsive to photon flux and incorporate time exposure/integration mechanisms. However, the two instruments differ substantially and each operates in its own intrinsic color dimensions, which results in *device-dependent color imaging*. Nevertheless, for the comparison at hand, the photographs taken with the camera are expected to qualitatively track the reflected integrated radiant flux viewed by the eye.

The *analysis* that permitted the brightness ratios to be estimated proceeds as follows:

1. Since P_0 was the same for all three sources, and since the optical rays traversing the eye followed paths that were essentially independent of wavelength, the radiance L at the fovea was the same for all three sources.
2. The foveal cones transformed the radiometric radiance L into the wavelength-dependent photometric luminance $L_V(\lambda_0)$. Since all three sources were monochromatic, these quantities could be related by $L_V(\lambda_0) = 683\, L \cdot V(\lambda_0)$ lm/W, where $V(\lambda_0)$ is the photopic luminous efficiency function, as provided in (8.8-5) and in Fig. 8.9-1.
3. As set forth in Table 8.9-1, the light from the green, red, and violet laser pointers had photopic luminous efficiencies given by $V(532\text{ nm}) = 0.8804$, $V(650\text{ nm}) = 0.1070$, and $V(405\text{ nm}) = 0.0008$, respectively. In accordance with (8.8-9), therefore, under ideal conditions the light from the green laser pointer should be a factor of $\sqrt[3]{0.8804/0.1070} = 2$ brighter than that from the red laser pointer, and a factor of $\sqrt[3]{0.8804/0.0008} = 10$ brighter than that from the violet one. Only qualitative comparisons are warranted, however, since this calculation does not account for the processing of the individual laser-spot images by the iPhone 13 Pro camera software, nor by the sRGB Photoshop software used to format the images, nor by the sRGB monitor display software, each of which may include its own gamma correction.
4. Examination of the photographs presented below the curve in Fig. 8.9-1 reveals that it is indeed plausible to suggest that the green laser spot is brighter than the red laser spot, which in turn is brighter than the violet laser spot. The high brightness of the light from the green laser pointer testifies as to why it is preferred over the red laser pointer, and especially over the rarely used violet one.

BIBLIOGRAPHY

Visual System

D. A. Atchison, *Optics of the Human Eye*, CRC Press/Taylor & Francis, 2nd ed. 2023.

L. A. Remington and D. Goodwin, *Clinical Anatomy and Physiology of the Visual System*, Elsevier, 4th ed. 2022.

S. Strong, *Introduction to Visual Optics: A Light Approach*, Elsevier, 2022.

G. W. Schwartz, *Retinal Computation*, Academic/Elsevier, 2021.

S. Grossberg, *Conscious Mind Resonant Brain*, Oxford University Press, 2021.

M. Yanoff and J. S. Duker, *Ophthalmology*, Elsevier, 5th ed. 2019.

P. Artal, ed., *Handbook of Visual Optics: Fundamentals and Eye Optics*, vol. 1, CRC Press/Taylor & Francis, 2017.

M. Ramamurthy and V. Lakshminarayanan, Human Vision and Perception, in R. Karlicek, C.-C. Sun, G. Zissis, and R. Ma, eds., *Handbook of Advanced Lighting Technology*, Springer Nature, pp. 757–784, 2017.

M. Pircher and R. J. Zawadzki, Review of Adaptive Optics OCT (AO-OCT): Principles and Applications for Retinal Imaging, *Biomedical Optics Express*, vol. 8, pp. 2536–2562, 2017.

D. H. Sliney, What is Light? The Visible Spectrum and Beyond, *Eye* (*London*), vol. 30, no. 2, pp. 222–229, 2016.

J. S. Werner and L. M. Chalupa, *The New Visual Neurosciences*, MIT Press, 2013.

W. S. Geisler and M. Banks, Visual Performance, in M. Bass, C. DeCusatis, J. Enoch, V. Lakshminarayanan, G. Li, C. Macdonald, V. Mahajan, and E. van Stryland, eds., *Handbook of Optics: Volume III – Vision and Vision Optics*, 3rd ed., McGraw–Hill, Chap. 2, 2010.

W. N. Charman, Optics of the Eye, in M. Bass, C. DeCusatis, J. Enoch, V. Lakshminarayanan, G. Li, C. Macdonald, V. Mahajan, and E. van Stryland, eds., *Handbook of Optics: Volume III – Vision and Vision Optics*, 3rd ed., McGraw–Hill, Chap. 1, 2010.

R. W. Rodieck, *The First Steps in Seeing*, Sinauer Associates, 1998.

B. A. Wandell, *Foundations of Vision*, Sinauer Associates, 1995.

D. H. Sliney, R. T. Wangemann, J. K. Franks, and M. L. Wolbarsht, Visual Sensitivity of the Eye to Infrared Laser Radiation, *Journal of the Optical Society of America*, vol. 66, pp. 339–341, 1976.

Trichromacy, Color Opponency, and Color Vision

K. A. Jameson, T. A. Satalich, K. C. Joe, V. A. Bochko, S. R. Atilano, and M. C. Kenney, *Human Color Vision and Tetrachromacy*, Cambridge University Press, 2020.

C. W. Tyler, Is Human Color Perception Complementary or Opponent?, *Investigative Ophthalmology & Visual Science (Abstract)*, vol. 61, no. 7, p. 2332, 2020.

M. K. Parthasarathy and V. Lakshminarayanan, Color Vision and Color Spaces, *Optics & Photonics News*, vol. 30, no. 1, pp. 44–51, 2019.

S. K. Shevell and P. R. Martin, Color Opponency: Tutorial, *Journal of the Optical Society of America A*, vol. 34, pp. 1099–1108, 2017.

L. Sawides, A. de Castro, and S. A. Burns, The Organization of the Cone Photoreceptor Mosaic Measured in the Living Human Retina, *Vision Research*, vol. 132, pp. 34–44, 2017.

N. Daw, *How Vision Works: The Physiological Mechanisms Behind What We See*, Oxford University Press, 2012.

A. Stockman and D. H. Brainard, Color Vision Mechanisms, in M. Bass, C. DeCusatis, J. Enoch, V. Lakshminarayanan, G. Li, C. Macdonald, V. Mahajan, and E. van Stryland, eds., *Handbook of Optics: Volume III – Vision and Vision Optics*, 3rd ed., McGraw–Hill, Chap. 11, 2010.

D. Mustafi, A. H. Engel, and K. Palczewski, Structure of Cone Photoreceptors, *Progress in Retinal and Eye Research*, vol. 28, pp. 289–302, 2009.

S. G. Solomon and P. Lennie, The Machinery of Colour Vision, *Nature Reviews Neuroscience*, vol. 8, pp. 276–286, 2007.

L. T. Sharpe, A. Stockman, W. Jagla, and H. Jägle, A Luminous Efficiency Function, $V^*(\lambda)$, for Daylight Adaptation, *Journal of Vision*, vol. 5, pp. 948–968, 2005; A Luminous Efficiency Function, $V^*_{\mathrm{D65}}(\lambda)$, for Daylight Adaptation: A Correction, *Color Research and Application*, vol. 36, pp. 42–46, 2011.

P. K. Kaiser and R. M. Boynton, *Human Color Vision*, Optical Society of America, 2nd ed. 1996.

G. Buchsbaum and A. Gottschalk, Trichromacy, Opponent Colours Coding and Optimum Colour Information Transmission in the Retina, *Proceedings of the Royal Society of London Series B (Biological Sciences)*, vol. 220, pp. 89–113, 1983.

W. D. Wright, *The Rays are Not Coloured: Essays on the Science of Vision and Colour*, Hilger, 1967.

L. M. Hurvich and D. Jameson, An Opponent-Process Theory of Color Vision, *Psychological Review*, vol. 64, pp. 384–404, 1957.

W. D. Wright, *Researches on Normal and Defective Colour Vision*, Kimpton, 1946.

Radiometry, Photometry, and Brightness

R. M. Bunch, *Optical Systems Design Detection Essentials: Radiometry, Photometry, Colorimetry, Noise, and Measurements*, IOP Publishing, 2021.

W. R. McCluney, *Introduction to Radiometry and Photometry*, Artech House, 2nd ed. 2014.

M. Bukshtab, *Applied Photometry, Radiometry, and Measurements of Optical Losses*, Springer, 2012.

B. G. Grant, *Field Guide to Radiometry*, SPIE Optical Engineering Press, 2011.

A. V. Arecchi, T. Massadi, and R. J. Koshel, *Field Guide to Illumination*, SPIE Press, 2007.

R. B. Barlow, Jr., Brightness Sensation and the Neural Coding of Light Intensity, in S. J. Bolanowski, Jr. and G. A. Gescheider, eds., *Ratio Scaling of Psychological Magnitude: In Honor of the Memory of S. S. Stevens*, Erlbaum, pp. 163–182, 1991.

J. C. Stevens and S. S. Stevens, Brightness Function: Effects of Adaptation, *Journal of the Optical Society of America*, vol. 53, pp. 375–385, 1963.

G. Ekman, H. Eisler, and T. Künnapas, Brightness Scales for Monochromatic Light, *Scandinavian Journal of Psychology*, vol. 1, pp. 41–48, 1960.

Historical Accounts and Seminal Publications

B. R. Masters, A History of Human Color Vision from Newton to Maxwell, *Optics & Photonics News*, vol. 22, no. 1, pp. 43–47, 2011.

B. R. Masters, Hermann von Helmholtz: A 19th Century Renaissance Man, *Optics & Photonics News*, vol. 21, no. 3, pp. 34–39, 2010.

J. D. Mollon, The Origins of Modern Color Science, in S. K. Shevell, ed., *The Science of Color*, Optical Society of America/Elsevier, pp. 1–39, 2nd ed. 2003.

S. S. Stevens, On the Psychophysical Law, *Psychological Review*, vol. 64, pp. 153–181, 1957.

A. Wood and F. Oldham, *Thomas Young, Natural Philosopher (1773–1829)*, Cambridge University Press, 1954.

E. Hering, *Zur Lehre vom Lichtsinne*, Grundzüge einer Theorie des Farbensinnes (Sechste Mittheilungen an die Kaiserliche Akademie der Wissenschaften in Wien 1874), pp. 107–141. Druck und Verlag von Carl Gerold's Sohn (Wien), Zweiter unveränderter Abdruck, 1878 [Translation: *Outlines of a Theory of the Light Sense*, L. M. Hurvich and D. Jameson, translators, Harvard University Press, 1964].

H. L. F. von Helmholtz, Physiologische Optik, in G. Karsten, ed., *Handbuch der physiologischen Optik*, Volume 9, *Allgemeine Encyklopädie der Physik*, pp. 1–874, Leopold Voss (Leipzig), 1867 [Translation: *Handbook of Physiological Optics*, N. Wade, ed. and J. P. C. Southall, translator, Thoemmes Continuum (Bristol), 2000].

J. C. Maxwell, On the Theory of Compound Colours, and the Relations of the Colours of the Spectrum, *Philosophical Transactions of the Royal Society of London*, vol. 150, pp. 57–84, 1860.

T. Young, The Bakerian Lecture. On the Theory of Light and Colours, *Philosophical Transactions of the Royal Society of London*, vol. 92, pp. 12–48, 1802.

G. Palmer, *Theory of Colours and Vision*, S. Leacroft (London), 1777 [Reprinted in D. L. MacAdam, ed., *Sources of Color Science: Selected and Edited by David L. MacAdam*, MIT Press, 1970].

I. Newton, *Opticks: or A Treatise of the Reflections, Refractions, Inflections & Colours of Light*, Samuel Smith and Banjamin Walford, Printers to the Royal Society, 1st ed. 1704; 4th ed. 1730, Dover, reissued 1979.

CHAPTER

9

COLORIMETRY

9.1 COLOR MATCHING AND MIXING 268

9.2 GRASSMANN'S LAWS 270

9.3 COMPLEMENTARY AND METAMERIC COLORS 271

9.4 COLOR APPEARANCE 273

9.5 COLOR SPACES AND COLOR SOLIDS 276

9.6 CHROMATICITY DIAGRAMS 285

9.7 COLOR TEMPERATURE 291

9.8 CORRELATED COLOR TEMPERATURE 296

9.9 COLOR RENDERING INDEX 299

Hermann Grassmann (1809–1877), left, a German polymath, studied the perception of mixtures of light of different colors. The empirical laws that emerged from his work serve as the underpinnings of colorimetry. **W. David Wright (1906–1997)**, center, and **John Guild (1889–1979)**, right, were British scientists who independently carried out series of color-matching experiments in the late 1920s. With the help of Grassmann's laws, their results became the basis of XYZ color space, a colorimetry system adopted by the Commission Internationale de l'Éclairage (CIE) in 1931 that is still in wide use today.

LED Lighting: Devices and Colorimetry. Malvin Carl Teich.
Google Books. Published 2024.

Colorimetry is the science and technology of color and its measurement. It is concerned with the spectral properties of stimuli and how they are interpreted as color at the retina and at higher neural waystations in the visual system. Understanding the principles of colorimetry enables color to be precisely controlled for a wide variety of applications, ranging from textile and paint manufacture, to color-reproduction processes such as printing and displays, to LED lighting.

Basic colorimetry, also called **tristimulus colorimetry**, developed as a technical field in the mid-nineteenth century. It describes the approaches, procedures, and outcomes of color matching experiments, and focuses on the perception of colors and their mixtures. It enables an observer to ascertain *when* two colors look alike, while circumventing the far more complex question of *what* they look like. Although limited to the comparison of stimuli that are temporally and spatially identical, and are viewed under identical conditions, basic colorimetry provides a versatile platform for evaluating the nature and use of color under a wide variety of circumstances.

It has long been understood, however, that color perception also involves phenomena that lie beyond the purview of basic colorimetry. These include time-dependent and space-dependent features, examples of which are the adaptation state of the observer and the color of the region surrounding the area of interest in a scene, respectively. Color perception is also dependent on the intensity and spectral properties of the source of light, and involves memory and nonlinearity (Sec. 8.8). As codified by Wyszecki,[†] these features lie within the domain of **advanced colorimetry**, where consideration is given to *what* colors look like in natural and complex settings, both relative to each other and under different viewing conditions. Advanced colorimetry is generally cast in the form of **color appearance models**, which began to appear the mid-twentieth century. The ultimate goal of advanced colorimetry is to permit all perceptual attributes of color to be predicted, under the broadest range of viewing conditions. This is a tall order that requires models of substantial complexity.

We begin in Sec. 9.1 with the notion of color matching, along with the concepts of additive and subtractive elementary color mixing. This is followed in Sec. 9.2 by a narration of Grassmann's celebrated four laws of color mixing, which are used principally in basic colorimetry and are approximate. Grassmann's second and third laws, which relate to complementary and metameric colors, respectively, are elucidated in Sec. 9.3. Relative color appearance and color appearance phenomena that lie outside the confines of basic colorimetry, such as Hering's observations pertaining to opponent colors, are introduced in Sec. 9.4.

Color spaces are colorimetry systems that rely on specific sets of stimulus-based primaries that offer particular features. A number of color spaces associated with both basic and advanced colorimetry are examined in Sec. 9.5. Grassmann's first and fourth laws, which characterize the essential elements of color and the linearity of luminance perception, respectively, provide important contributions to this section. Chromaticity diagrams, which are 2D planar constructs derived from 3D color spaces, greatly facilitate the visualization and interpretation of colorimetric data, as explained in Sec. 9.6.

Color temperature, a widely used measure in illumination engineering, characterizes the color of a thermal source of light in a concise manner, as detailed in Sec. 9.7. Correlated color temperature is an analogous measure used to characterize a nonthermal source of light whose color resembles that of a thermal source, as described in Sec. 9.8. Finally, the color rendering index, discussed in Sec. 9.9, is a measure that indexes how well a light source illuminating an object renders its color.

[†] G. Wyszecki, Current Developments in Colorimetry, in *COLOUR73: Survey Lectures and Abstracts of the Papers Presented at the Second Congress of the International Colour Association Held at the University of York 2–6 July 1973*, Hilger (London), pp. 21–51, 1973.

The CIE. The *Commission Internationale de l'Éclairage*, the abbreviation for which is CIE in French, is known as the *Internationale Beleuchtungskommission* in German and the *International Commission on Illumination* in English. The CIE was established in August 1913 — a brief history recounting its creation is provided in the caption of Fig. 9.0-1. Based in Vienna, it is an authoritative source in the fields of vision, color measurement of light, photobiology, lighting technology, interior lighting, exterior lighting, and digital imaging. To this day, the CIE remains the international arbiter on light and illumination, and develops the standards on color and lighting.

The CIE holds conferences and symposia, and annually publishes dozens of Technical Reports (TRs), Technical Notes (TNs), International Standards (ISs), and Position Statements. Of its eight divisions, the two that are most closely allied with the material considered in this book are Division 1 (*Vision and Color*) and Division 3 (*Interior Environment and Lighting Design*):

The mandate of Division 1 is to "study visual responses to light and to establish standards of response functions, models and procedures of specification relevant to photometry, colorimetry, color rendering, visual performance and visual assessment of light and lighting." The mandate of Division 3 is to "study and evaluate visual factors which influence the satisfaction of the occupants of a building with their environment, and their interaction with thermal and acoustical aspects, and to provide guidance on relevant design criteria for both natural and man-made lighting; as well as to study design techniques, including relevant calculations, for the interior lighting of buildings; incorporating these findings and those of other CIE Divisions into lighting guides for interiors in general, for particular types of interiors and for specific problems in interior lighting practice."

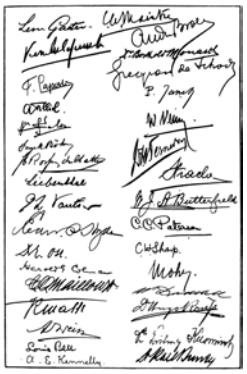


Figure 9.0-1 The Commission Internationale de l'Éclairage (CIE) was built on the foundation of its predecessor, the Commission Internationale de Photométrie (CIP), which had been established in Paris in 1900 to set standards for the measurement of light emitted by incandescent gas lamps, which were widely used for street lighting in the late nineteenth century. At the CIP meeting held in Berlin in August 1913, the CIE was established as its successor. The new commission was given a broad mandate to consider all issues related to the lighting industry and to the science appertenant thereto, and to forge international agreements on lighting. A collection of the delegates' signatures collected at the Esplanade Hotel conference dinner during the 1913 meeting is displayed at left. (Adapted from J. W. T. Walsh and A. M. Marsden, *History of the CIE: 1913–1988*, CIE Report No. 82-1990, p. 5, photocopy ed. 1999.)

A listing of some of the milestones attained by the CIE that are germane to the topics reviewed in this chapter is provided below:

- *1924*: The CIE established the photopic luminous efficiency function $V(\lambda_0)$ (Fig. 8.5-3), which represents the relative effectiveness of light of different wavelengths in stimulating the photopic visual system. An updated version for daylight adaptation, denoted $V^*(\lambda_0)$, dates from 2005.
- *1931*: Based on a 1922 colorimetry report from the Optical Society of America (OSA), together with the scientific and technological developments of the following decade, the CIE arrived at: a definition of the CIE 1931 2° standard colorimetric observer; its corresponding color matching functions; the CIE 1931 RGB and XYZ color spaces; and the standard illuminants A, B, and C.
- *1951*: The CIE established the scotopic luminous efficiency function $V'(\lambda_0)$, which is analogous to the photopic luminous efficiency function $V(\lambda_0)$ set forth by the CIE in 1924.
- *1960*: The CIE introduced the CIE 1960 UCS (uniform color space) for the calculation of the correlated color temperature (CCT). Although this color space has been superseded by 1976 CIELUV and CIELAB for most applications, it continues to be used for the determination of the CCT since it turns out to be more uniform for nominally white chromaticities.
- *1964*: The 10° standard colorimetric observer, and its corresponding color matching functions, were added to the CIE repertoire, along with the standard daylight illuminant D_{65}.

- *1965*: In response to the widespread use of fluorescent lamps, the CIE recommended use of the color rendering index (CRI) as a metric. The CRI is computed using the otherwise-obsolete 1964 CIEUVW color space.
- *1976*: The CIE introduced the CIELUV and CIELAB uniform color spaces with the goal of improving perceptual uniformity; these were the initial formal color appearance models in the CIE colorimetry-systems repertoire and have stood the test of time.
- *2007*: The CIE initiated consideration of LED lighting with studies of LED measurements and the color-rendering properties of metameric-white LEDs.
- *2015*: The CIE introduced a new cone-fundamental-based CIE colorimetry system with color matching functions that accommodate observer age and stimulus field size.
- *2016*: The CIE introduced the CAM16 color appearance model as a successor to CIECAM02. The accompanying color space, CAM16-UCS, is under consideration for adoption.

9.1 COLOR MATCHING AND MIXING

Color Matching

Color matching is an experimental technique in which an observer is asked to view two adjacent colored lights (visual stimuli), and to make adjustments on one of them until the two look alike. The experiments are usually conducted by using a circular split screen, called a *bipartite field*, with a diameter of 2° (so that the image falls within the fovea) or 10°, although other diameters are used as well (Fig. 8.4-3). A *test light* is projected onto one of the hemispheres, say the upper one, while a *comparison light* that is adjustable by the observer is projected onto the other. The adjustable light comprises a mixture of three primary lights with fixed spectral colors but adjustable luminances. The observer modifies the luminance of each of the three primaries until a match to the test patch is attained. The manner in which the procedure is conducted is illustrated in Fig. 9.1-1 and detailed in the figure caption. The operating and adaptation conditions for the test and comparison lights are required to be similar.

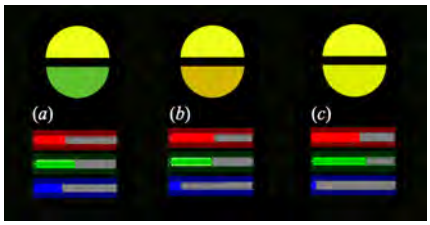


Figure 9.1-1 Maximum-saturation method of color matching using red, green, and blue primaries. The upper and lower hemispheres are the test and comparison patches, respectively. The sliders alter the relative contributions of the primaries to the comparison patch. (*a*) Starting point. (*b*) Color matching is improved by increasing the red and green, and reducing the blue. (*c*) Color matching is optimized by further increasing the red and green, and reducing the blue.

Using this procedure in conjunction with a particular set of primaries, it is possible to obtain a color match most, but not all, of the time. It turns out that some test colors cannot be mimicked by a combination of the three primary lights; this situation arises when the test color is too saturated to be matched by the primaries or, stated differently, is out of their gamut. All is not lost, however, since in those cases, the experimental protocol is altered so that one of the primaries is removed from the comparison patch and its variable intensity is instead added to the test patch, which serves to desaturate it and bring it within the gamut. A match between the modified test color and the two remaining primary colors can then always be found. The intensity of the primary light required to be added to the test color to achieve a match is considered to be subtracted from the comparison color, i.e., to be negative. Allowing a primary color to have a 'negative intensity' over some range of color matches allows all test colors to be matched, whatever the choice of the primaries. Indeed, real primaries always give

rise to negative values because three such primary lights cannot severally and uniquely stimulate the three types of cones since their sensitivities overlap throughout the visible spectrum. There is no unique way of defining the three primaries; red, green, and blue are often used because they are convenient and offer matching over a wide range of test colors.

For monochromatic test patches that range over all visible wavelengths, the results of the color matching procedure can be succinctly summarized by plotting the intensities of each of the primaries required to effect a color match over all wavelengths. The triad of plots of this form, one for each primary, are known as the **color matching functions** for those particular primaries.

Color matching allows an observer to ascertain* WHEN *two colors look alike, while avoiding the subjective task of having to describe* WHAT *they look like.

Color matches vary with retinal size and position. However, they generally survive changes in context and adaptation, provided that the changes are applied equally to both lights. Moreover, a color match that satisfies a particular normal observer will generally satisfy all normal observers. Much of the behavioral data that contribute to our understanding of how color vision operates, from Grassmann's laws discussed below to the chromaticity diagrams presented in Sec. 9.5, are built upon color matching experiments. Grassmann's laws inform us that polychromatic test patches can also be used in color matching experiments.

Elementary Color Mixing

The superposition of two or more beams of light is described by **additive color mixing**. This phenomenon is most clearly demonstrated by the superposition of red, green, and blue (RGB) spectral lights in equal proportions, as illustrated in Fig. 9.1-2. In the absence of light the result is black, as is apparent at the periphery of the illuminated screen, whereas the superposition of all three primary lights yields white (or gray), as appears at the center of the screen. White, black, and their neutral gray intermediaries are referred to as **achromatic colors**.

Moreover, as is clear in Fig. 9.1-2, the superposition of any two primary lights in equal proportion yields a particular secondary light: red plus green yields yellow; green plus blue yields cyan; and blue plus red yields magenta. The secondaries are lighter in shade than the primaries. Additive color mixing is sometimes illustrated in the form of temporal color mixing by viewing a spinning multicolored disk, an approach promulgated by Maxwell. Color mixing finds use in applications ranging from OLED displays (Sec. 7.6) to LED lighting (Sec. 11.3). More generally, the superposition of RGB spectral lights in arbitrary proportions is displayed in the RGB color solid provided in Figs. 9.5-2(a),(b).

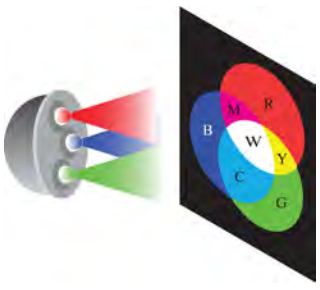


Figure 9.1-2 Additive color mixing in an **RGB color system**. A device that generates light with a selectable color can be constructed from LEDs that emit in the Red (R), Green (G), and Blue (B). When projected in equal proportions, the overlapping light beams exhibit the following colors:

R + G → Y (Yellow)
G + B → C (Cyan)
B + R → M (Magenta)
R + G + B → W (White)

In **subtractive color mixing**, illustrated in Fig. 9.1-3, the primaries are usually cyan, magenta, and yellow (CMY) pigments or transparencies. An important application of subtractive color mixing is printing, in which an external source of white light is reflected from printer's ink on white paper. The absence of pigment yields white (white light reflected from white paper), whereas the presence of all three pigments in equal proportions yields black (all external white light is absorbed by the pigments and none is reflected). The mixing of any two pigments in equal proportion gives rise to reflected light with a secondary color: cyan plus magenta yields blue; magenta plus yellow yields red; and yellow plus cyan yields green. The secondaries are darker in shade than the primaries. The result of mixing CMY primaries in arbitrary proportions is illustrated in the CMY color solid portrayed in Figs. 9.5-2(c),(d).

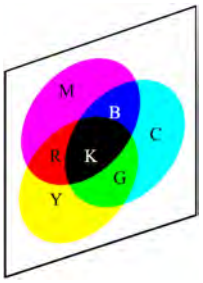


Figure 9.1-3 Subtractive color mixing in a **CMY color system**. Inks that reflect light with a selectable color can be achieved by mixing Cyan (C), magenta (M), and yellow (Y) pigments on a white background. When mixed in equal proportion, the reflected light exhibits the following colors:

C + M → B (Blue)
M + Y → R (Red)
Y + C → G (Green)
C + M + Y → K (Black)

9.2 GRASSMANN'S LAWS

Hermann Grassmann, a German polymath, studied how mixtures of light that simultaneously impinge on the same region of the retina are perceived. Building on the work of Newton, Helmholtz, and Maxwell, he extended the scope of elementary additive color mixing introduced in the previous section. His empirical laws, which are approximate, serve as the underpinnings of colorimetry and comprise a set of principles that describe how the addition of colors in various proportions is perceived. Grassmann's laws, which have stood the test of time, are the precursors to our contemporary understanding of color spaces and chromaticity diagrams, and inform our understanding of human color vision.

Interpretations of Grassmann's Laws. An examination of the literature pertaining to Grassmann's laws reveals that different authors interpret them very differently. Some authors even frame his laws in terms of a set of axioms, although Grassmann never presented them in that form. The rendition of Grassmann's laws provided below closely hews to the original text of his 1853 article, published in German in *Annalen der Physik*.[†]

[†] H. Grassmann, Zur Theorie der Farbenmischung, *Annalen der Physik*, vol. 165(5), pp. 69–84, 1853.

Grassmann's Laws of Color Mixing

Hermann Grassmann, in an article entitled *Zur Theorie der Farbenmischung* that appeared in *Annalen der Physik* in 1853, set forth four empirical laws of additive color mixing. His work provided a roadmap for describing how mixtures of colored lights are perceived. Using modern terminology, Grassmann's four laws can be cast in the following form:

1. *Elements of Color.* Three elements are necessary and sufficient to specify a color: (a) **hue**, (b) **saturation**, and (c) **luminance**. (Grassmann referred to these elements, respectively, as: (a) *der Farbenton*, (b) *die Intensität des beigemischten farblosen Lichtes*, and (c) *die Intensität der Farbe*.)
2. *Complementary Colors.* Every color has an associated **complementary color**, such that the mixing of their lights can give rise to achromatic (white or gray) light.
3. *Metameric Colors.* **Metamers** are lights that have matching colors (identical hue and saturation) but non-matching power spectral densities. Mixing two metamers yields a third metamer with the same hue and saturation.
4. *Linearity in Luminance.* The total luminance of a mixture of light is the sum of the luminances of the constituent lights, signifying that color mixing is **linear in luminance**.

Grassmann's first law laid the foundations for the definition of the iconic CIE 1931 XYZ color space considered in Sec. 9.5. Grassmann's fourth law, which enunciated that color mixing is linear in luminance, ensures that a panoply of alternative color representations exist and that they are all linear transforms of each other, a result that is extensively used in Sec. 9.5. The linearity in luminance (and its variants) in the visual system parallels the linearity in radiance (and its variants) in photodetectors.

9.3 COMPLEMENTARY AND METAMERIC COLORS

We now proceed to discuss in greater detail Grassmann's second and third laws, which pertain to complementary and metameric colors, respectively.

Complementary Colors

Every color has a **complementary color**, which, when the two are mixed gives rise to a desaturation of the stronger of the two or, when mixed in equal proportion to an achromatic white or gray. In the RGB color model, the complements of the primaries red, green, and blue (RGB) are the secondaries cyan, magenta, and yellow (CMY), respectively. White light can be generated as the combination of two, rather than three, complementary colors because the secondary color complementary to the primary consists of the other two primaries. Hence, all three primaries are effectively present, which enables white to be produced; mixing blue and yellow, for example, gives rise to white (or gray) because yellow is itself a combination of red and green. The mixing of complementary colors is illustrated in Examples 9.3-1 and 9.3-2, while the mixing of noncomplementary colors is portrayed in Example 9.3-3 (image colors were generated using sRGB in Adobe Photoshop®).

EXAMPLE 9.3-1. ***Mixing Blue and Yellow.*** Blue and yellow are complementary colors. As is apparent in Fig. 9.3-1, when the proportion of blue exceeds that of yellow (B>Y), the result is bluish; conversely, when the yellow exceeds the blue (Y>B), the result is yellowish. Panels B>Y and Y>B, which are desaturated versions of B and Y, respectively, are themselves complementary colors. When mixed in equal proportions (B=Y), the outcome is fully desaturated, achromatic white. When displayed at a reduced luminance, as sketched in Fig. 9.3-2, B>Y and Y>B remain complementary colors and the B=Y mixture yields fully desaturated achromatic gray rather than white. Still, complementary colors presented adjacently to each other offer the strongest contrast.

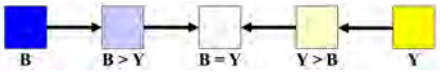


Figure 9.3-1 Mixing the complementary colors blue and yellow in different proportions.

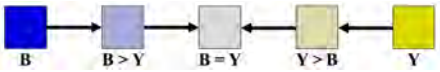


Figure 9.3-2 Mixing the complementary colors blue and yellow at reduced luminance.

EXAMPLE 9.3-2. ***Mixing Red and Cyan.*** imilarly, red and cyan are complementary colors. As is clear in Fig. 9.3-3, when the proportion of red exceeds that of cyan (R>C), the result is reddish; conversely, when the cyan exceeds the red (Y>B), the result is pale cyan. Panels R>C and C>R, which are desaturated versions of R and C, respectively, are themselves complementary colors. Again, when mixed in equal proportions (R=C), the outcome is fully desaturated, achromatic white. When displayed at reduced luminance, as sketched in Fig. 9.3-4, R>C and C>R remain complementary colors and the R=C mixture yields fully desaturated achromatic gray rather than white.

Figure 9.3-3 Mixing the complementary colors red and cyan in different proportions.

Figure 9.3-4 Mixing the complementary colors red and cyan at reduced luminance.

EXAMPLE 9.3-3. ***Mixing Red and Blue.*** In contrast to the outcome of mixing complementary colors, as displayed in Examples. 9.3-1 and 9.3-2, mixing noncomplementary colors such as red and blue does not lead to desaturation, and mixing in equal proportions does not lead to an achromatic result (as there is no green present). As depicted in Fig. 9.3-5, an equal proportion of red and blue (R=B) yields magenta. When displayed at reduced luminance, as sketched in Fig. 9.3-6, the R=B mixture remains magenta.

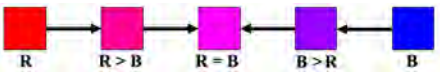


Figure 9.3-5 Mixing the noncomplementary colors red and blue in different proportions.

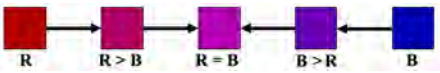


Figure 9.3-6 Mixing the noncomplementary colors red and blue at reduced luminance.

There is a distinction between complementary and opponent colors in visual perception. As illustrated in Example 9.3-2, the color complementary to red is cyan, whereas the color opponent to red is green, as discussed in Sec. 8.6. Indeed, mixing red and green yields yellow, while mixing red and cyan yields white (since cyan itself is a mixture of green and blue) (Fig. 9.1-2).

Metameric Colors

Sources of light with distinct power spectral densities $S_\lambda(\lambda_0)$ that are nevertheless perceived to be identical with respect to hue and saturation are known as **metamers**. Yellow light appears identical to the observer, as an example, whether its spectrum is a monochromatic line of wavelength of $\lambda_0 = 590$ nm, or two monochromatic lines, one in the red and another in the green. When perceptually indistinguishable, such spectra are said to give rise to **metameric colors**.

As another example, light with a uniform wavelength-based power spectral density $S_\lambda(\lambda_0)$ appears white to the eye (Example 9.6-3). But as demonstrated in Figs. 9.3-1 and 9.3-3, white light can also be produced by the additive mixing of blue and yellow, or red and cyan, which have spectra that consist of individual narrow lines. All three of these sources of white light are indistinguishable to the eye despite the fact that they are easily distinguished by a spectrophotometer. White is an achromatic color whose perception is induced by light with particular spectra. This feature of human vision proves important in the design of LED lighting, where **metameric white light** plays a key role, as discussed in Chapter 10.

> WHITE LIGHT *has a uniform power spectral density. Light that does not have a uniform power spectral density, but is nevertheless perceived as white, is called* METAMERIC WHITE LIGHT**.**

It will become clear in Sec. 9.5 that the origin of metamerism resides in the fact that the visual percept of a particular color is created by a triplet of numbers known as the **tristimulus values** — but identical tristimulus values can be generated by light with very different spectral properties. Furthermore, the mixing of two metamers results in a third metamer, as provided in Grassmann's third law, because their tristimulus values are conserved in additive color mixing.

Variations on the theme of metamerism exist. **Illuminant metamerism** refers to a situation in which the colors of two objects perceived by an observer match for one illuminant, but fail to match for a different illuminant. **Observer metamerism** refers to a situation in which the colors of two objects match for one observer, but fail to match for another observer. A measure of metamerism is provided by the color rendering index (CRI) discussed in Sec. 9.9, which serves to assess the difference in sample and reference spectral reflectances using a standardized set of different colored reflectors.

9.4 COLOR APPEARANCE

Describing the appearance of a color is a complex enterprise that relies on a collection of terms, measures, and abstractions. As a point of departure, it is useful to first set forth a number of definitions, which we draw in large part from the second edition of the CIE International Lighting Vocabulary:

Definitions of Color Appearance Terms and Measures

- *Color*. A characteristic of visual perception described by the attributes hue, brightness (or lightness), and colorfulness (or saturation or chroma). Color depends not only on the spectral features of the stimulus, but also on the colors that surround it and on the state and experience of the observer.
- *Color Gamut*. A volume in a color space, or more commonly an area in a chromaticity diagram, that defines a range of achievable colors under a given set of viewing conditions.
- *Unrelated Colors*. Colors perceived as belonging to an area or an object seen in isolation from other colors.
- *Related Colors*. Colors perceived as belonging to an area or an object seen in relation to other colors.
- *Achromatic Colors*. Colors such as white, black, and gray, that are devoid of hue.
- *Brightness*. An attribute of visual perception according to which an area appears to emit, transmit, or reflect, more or less light (Sec. 8.8).
- *Lightness*. The brightness of an area judged relative to the brightness of a similarly illuminated reference white area. Lightness is relative brightness. It is approximately constant across changes in luminance level.
- *Hue*. An attribute of visual perception according to which an area appears to be similar to one of the colors red, yellow, green, and blue, or to a combination of adjacent pairs of these colors when considered in a closed ring.
- *Colorfulness*. An attribute of visual perception according to which the perceived color of an area appears to be chromatic to a lesser or greater degree. Colorfulness describes the intensity of the hue in a particular color sample. It increases with increasing luminance level.
- *Saturation*. The colorfulness of an area judged relative to its own brightness.
- *Chroma*. The colorfulness of an area judged relative to the brightness of a similarly illuminated reference white area. Chroma is relative colorfulness. It is approximately constant across changes in luminance level.

Absolute and Relative Color Appearance. Color appearance can be described in terms of absolute measures (basic colorimetry), or in terms of relative measures that are normalized to accommodate changes in the viewing and illumination conditions (advanced colorimetry):

- *Absolute Color Appearance* is described by the attributes **hue**, **colorfulness** (**saturation**), and **brightness**. Unrelated colors exhibit only these perceptual attributes. In the context of Grassmann's empirical laws for self-luminous sources (Sec. 9.2), these attributes are referred to as **hue**, **saturation**, and **luminance**.
- *Relative Color Appearance* is described by the attributes **hue**, **chroma**, and **lightness**. These attributes correspond to the absolute attributes listed above, but are normalized to allow for changes in illumination and viewing conditions. Related colors exhibit these perceptual attributes in addition to those listed above for unrelated colors.

As discussed in Sec. 9.1, color matching can be described on the basis of three spectral colors with adjustable luminances. A proper description of color appearance, on the other hand, generally requires five perceptual parameters: brightness, lightness, hue, colorfulness, and saturation (chroma can be derived from these). For related colors, the description simplifies and it suffices to consider three relative color appearance attributes: hue, chroma, and lightness.

Color Appearance Phenomena. While basic colorimetry is suitable for describing color matching in a broad variety of contexts, its applicability is subject to a standard set of constraints. Color-vision effects that cannot be explained by basic colorimetry because of a *violation* of one or more of these constraints fall in the domain of advanced colorimetry and are known as **color appearance phenomena**. Violations often involve the viewing field and/or features such as background color, illumination color, luminance level, adaptation level, object structure, and visual-system nonlinearity. These violations often serve as the bases for *optical illusions*. Representative color appearance phenomena include the following:

- Chromatic adaptation.
- Simultaneous contrast exhibited using different backgrounds.
- Change of hue with luminance (Bezold–Brücke effect).
- Change of hue with colorimetric purity (Abney effect).
- Increase of brightness with saturation (Helmholtz-Kohlrausch effect).
- Increase of colorfulness with luminance (Hunt effect).
- Increase of contrast with luminance (Stevens effect).
- Increase of emissive image contrast with surround luminance (Bartleson-Breneman effect).

Chromatic adaptation, the ability of the human visual system to preserve the appearance of the colors of an object under different illumination colors, is one of the most important of the color appearance phenomena. This behavior was first considered by Johannes von Kries, who, like his contemporary Max Planck (p. 61), was a student of Helmholtz (p. 234). In 1902, von Kries suggested that each type of cone photoreceptor adapts independently to its illumination, with its gain determined by the particular scene under view.[†] This notion has stood the test of time; most modern models of chromatic adaptation rely on von Kries' approach but use different nonlinear adaptation functions and/or incorporate modifications that leave his basic structure intact. Edwin Land's **retinex theory**, for example, is a version of the von Kries model in which the usual spectral effects are augmented by spatial effects that make use of the average response over a particular region of the scene to normalize the response at a given point. The retinex approach is successful in explaining color variations attendant to changes in the background of the stimulus.

As indicated in the introduction to this chapter, the origin of formal color appearance models in CIE colorimetry systems can be traced back to the creation of the 1976 CIELUV and 1976 CIELAB uniform color spaces (Sec. 9.5), in which attempts were made to quantitatively predict the relative color appearance attributes hue, chroma, and lightness. In the same way that the two-stage zone model set forth in Sec. 8.6 explains how cone-opponent theory can be joined with trichromatic theory to elucidate color discrimination, various expanded three-stage zone models can serve as stepping stones toward elucidating color appearance phenomena. Given the complexity of the underlying neural system, however, it is ambitious to expect this from a mechanistic approach.

Most recently, the many advances that have been achieved over the years have resulted in CAM16, the current CIE CAM standard set forth in 2016, and its associated uniform color space CAM16-UCS (Sec. 9.5). The CAM16 model incorporates the perceptual attributes brightness, lightness, hue, colorfulness, saturation, and chroma.

It is important to recognize that some color appearance phenomena, including chro-

[†] J. von Kries, Theoretische Studien ueber die Umstimmung des Sehorgans (Theoretical Studies on the Retuning of the Visual Organ), in *Festschrift der Albrecht-Ludwigs-Universität in Freiburg zum fünfzigjärigen Regierungs-Jubiläum Seiner Königlichen Hoheit des Grossherzogs Friedrich*, C. A. Wagner's Universitäts-Buchdruckerei, pp. 143–158, 1902 [Translation: Chromatic Adaptation, in D. L. MacAdam, ed., *Sources of Color Science: Selected and Edited by David L. MacAdam*, pp. 109–119, MIT Press, 1970].

matic adaptation, have substantial cognitive, as well as sensory, aspects. Previous experience with particular sources of light, viewing environments, and objects plays an important role in color interpretation. A shadow falling across an object is ignored, for example, as is the curvature of the object.

Image appearance models, more complex forms of color appearance models, incorporate various aspects of temporal vision and spatial vision, and allow for the measurement of image differences. Whereas color appearance models consider attributes such as hue, colorfulness, chroma, brightness, and lightness, image appearance models also consider attributes such as contrast, graininess, sharpness, and resolution.

9.5 COLOR SPACES AND COLOR SOLIDS

A **color space** is a mathematical structure resembling a 3D vector space that enables color to be specified, created, and visualized. In its simplest conception, its basis vectors represent a set of physical primary colors such as red, green, and blue. The color space supports a palette of perceived colors, such as those associated with the tristimulus values determined via color matching experiments conducted in the laboratory, using the primary lights specified.

It can be said that the era of modern basic colorimetry began with the RGB color space conceptualized by the CIE in 1931 on the basis of the extensive maximum-saturation color matching data collected by W. David Wright and John Guild (p. 265) in the late 1920s. The CIE 1931 RGB color space relies on monochromatic $\mathfrak{RGB}$ primaries of wavelengths 700, 546.1, and 435.8 nm, respectively. Since RGB color models admit additive color mixing and offer a wide range of colors, and since red, green, and blue LEDs are widely available, RGB color spaces are widely used in LED lighting. Several variants of these color spaces are detailed at the end of this section.

A color space need not make use of physical lights as primaries, however, nor need it display a palette of supported colors arising from color matching experiments. Experimental color-matching data can be mathematically transformed into color spaces whose primary lights are physically unrealizable (imaginary) and therefore not visible, and whose palette displays visual attributes such as hue (color), saturation (colorfulness), and luminance (brightness). The basis vectors are then linked to a set of color matching functions.

Color Solids. **Color solids**, sometimes called **color volumes**, are 3D analogs of 2D color wheels that display the panoply of colors (gamut) supported by different color spaces. They depict how various color-appearance features, such as hue, chroma, and lightness, relate to one another. In the early 1900s, the artist Albert Munsell developed a roughly spherical color solid that allowed colors to be visually and spatially organized so that they were intuitively understandable and useful for various applications. Munsell's construct is still in use today, as will be seen in Fig. 9.9-1) as an example.

Color Systems. While a color space offers a mathematical platform for representing colors, a **color system** couples a color space with a medium and technological implementation for the supported colors (e.g., lighting, display, printing). The Munsell Color System, for example, augmented his initial theoretical framework by including practical implementations such as color samples and color atlases.

Color Spaces Abound. Different color spaces make use of different primaries (degrees of freedom) and offer different salutary features. In this section, our attention

will be principally directed to five commonly used color spaces labeled: LMS, CIE 1931 XYZ, 1976 CIELUV, CAM16-UCS, and sRGB. The first is the LMS color space associated with the S-, M-, and L-cones in the human retina, each of which is endowed with its own spectral sensitivity curve (Sec. 8.5). The primaries that give rise to the cone fundamentals $\bar{s}(\lambda_0)$, $\overline{m}(\lambda_0)$, and $\bar{l}(\lambda_0)$ as color matching functions should hypothetically stimulate each of the three types of cones severally and uniquely. Since the three cone sensitivity curves overlap significantly, however, this is not possible, which results in imaginary primaries. In practice, this color space is principally used today for modeling chromatic adaptation and color-vision deficiencies.

The second, CIE 1931 XYZ, was mathematically constructed from CIE 1931 RGB. It was adopted by the CIE (Commission Internationale de l'Éclairage) in 1931, and is the grandfather of all color spaces. The CIE selected imaginary primaries that rendered the associated color matching functions everywhere nonnegative. This space separates hue, saturation, and luminance, the elements of color inherent in Grassmann's first law (Sec. 9.2), and accommodates the full gamut of perceptible colors. It has remained the *de facto* standard among color spaces for nearly a century and continues to enjoy widespread use. Grassmann's fourth law (Sec. 9.2) teaches that color mixing in trichromatic photopic vision is approximately linear so that one color space can be readily transformed to another.

The third color space, 1976 CIELUV, was created as an early color appearance model with the additional advantage of improved uniformity, signifying that a specified Euclidian distance anywhere within the space represents the same degree of perceived color difference. The fourth, CAM16-UCS is an updated color space developed in 2016 that offers superior uniformity and increased subtlety. The use of a uniform color spaces (UCS) facilitates the interpretation of color mixing data and the determination of color differences.

Finally, the fifth is the family of RGB color spaces born of the digital era, which are commonly used for display, illumination, and printing. These color spaces typically focus on attributes of the stimulus rather than on the processing carried out in the retina and visual system. However, like all color spaces, color solids, and color systems, they too are ultimately accountable to the biology of color perception. Three related and widely used color models, CMY, HSV, and HSL, are briefly mentioned at the close of the section and selected color solids are displayed.

Definitions, Notation, and Significance of Different Fonts

- *Primary Lights.* A set of three independent lights (real or imaginary) used to match a test patch (actually or hypothetically). Red, green, and blue primary lights in the RGB color model are denoted $\mathfrak{R}$, $\mathfrak{G}$, and $\mathfrak{B}$, respectively, and yield white light when added together in appropriate proportions.
- *Color Matching Functions.* Three functions of wavelength, denoted $\bar{r}(\lambda_0)$, $\bar{g}(\lambda_0)$, and $\bar{b}(\lambda_0)$, characterize the human observer by specifying the intensities of the primary lights required to match spectral colors at every wavelength.
- *Tristimulus Values.* For monochromatic light, the three intensities of the color matching functions, denoted R, G, and B, that represent a particular spectral color; for polychromatic light, the three inner products of the incident spectral density and the color matching functions.
- *Color Space.* Akin to a 3D vector space, but often limited to a unit cube. An RGB color space, with basis vectors labeled RGB, displays all supported colors associated with the tristimulus values RGB. A color space is specified either by its stimulus-based primaries or by its color matching functions.
- *Normalized Tristimulus Values.* The tristimulus values R, G, and B, each normalized by division by the sum $R + G + B$, and denoted r, g, and b, respectively, so that $b = 1 - r - g$.
- *Chromaticity Diagram.* A projection of the normalized tristimulus values onto a 2D plane parameterized by the dimensionless chromaticity coordinates r and g, as discussed in Sec. 9.6.

Summary. The usual notation for color spaces takes the following form: An* RGB *color space is associated with: 1) the primaries $\mathfrak{RGB}$; 2) the color matching functions $\bar{r}(\lambda_0)\bar{g}(\lambda_0)\bar{b}(\lambda_0)$; 3) the tristimulus values RGB; and 4) the normalized tristimulus values (chromaticity coordinates) rgb.

LMS *Color Space*

The cone fundamentals, designated as $\bar{s}(\lambda_0)$, $\overline{m}(\lambda_0)$, and $\bar{l}(\lambda_0)$ and displayed in Fig. 8.5-1(b), serve as color matching functions for the LMS color space. A particular spectral color along the abscissa of the plot is mapped to a perceived color via the set of ordinates of the color matching functions, which are called the **LMS tristimulus values**. For light of arbitrary spectrum, the tristimulus values are determined by computing the inner products of the wavelength-based spectral density incident on the eye and the color matching functions. The three-dimensional space spanning all LMS tristimulus values is a cube that represents all perceived colors and is known as the **LMS color space**. The primaries that give rise to the cone fundamentals are three imaginary lights that hypothetically stimulate the three types of cones severally and uniquely.

Although LMS holds a special place in the pantheon of color spaces by virtue of its biological basis for trichromats, it is not unique and is only occasionally used (principally for modeling chromatic adaptation and color blindness). Now that the LMS cone fundamentals have been determined with exceptional accuracy, however, the CIE is engaged in the *ab initio* establishment of a comprehensive new colorimetry system based on the cone fundamentals. Such a system is far more intuitive than one based on imaginary primaries.

As suggested by Grassmann's fourth law, the conversion of one color space to another takes the form of a matrix transformation of their tristimulus values

and, by extension, of their color matching functions.

CIE 1931 XYZ Color Space

Long before the cone spectral sensitivity curves depicted in Fig. 8.5-1(a) were measured, a consensus had been reached that it was important to establish a single color space as a standard to advance commerce and industry. As indicated in the introduction to this chapter, the CIE convened a meeting in Cambridge, England in 1931 with the express purpose of doing so and the **XYZ color space** emerged as that standard. As mentioned above, this space separates hue, saturation, and luminance, in accordance with Grassmann's first law (Sec. 9.2), by drawing on imaginary primaries labeled $\mathfrak{X}$, $\mathfrak{Y}$, and $\mathfrak{Z}$. It accommodates the full gamut of perceptible colors. The definition of this color space, like all others, specifies a well-defined source of illumination, lighting conditions, and viewing geometry, all of which affect color appearance (Sec. 9.4).

Color-Matching Functions. The **color-matching functions** for the standard observer are displayed in Fig. 9.5-1. They were selected by the CIE so that the three functions, $\bar{x}(\lambda_0)$, $\bar{y}(\lambda_0)$, and $\bar{z}(\lambda_0)$, were everywhere nonnegative and had equal areas. The peak value of $\bar{y}(\lambda_0)$ was chosen to be unity. The three color matching functions overlap significantly and the primaries are imaginary.

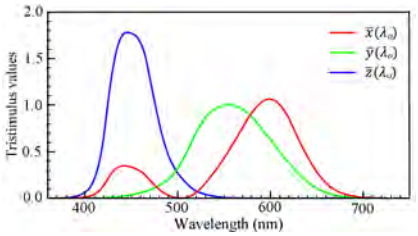


Figure 9.5-1 CIE (Commission Internationale de l'Éclairage) 1931 standardized set of color-matching functions: $\bar{x}(\lambda_0)$, $\bar{y}(\lambda_0)$, and $\bar{z}(\lambda_0)$. The perceived color of a source of light of arbitrary spectral density $S_\lambda(\lambda_0)$ is specified in terms of the tristimulus values X, Y, and Z computed from $S_\lambda(\lambda_0)$ and the three color-matching functions.

An important feature, and a particular convenience, of this space is that the central color matching function $\bar{y}(\lambda_0)$, portrayed as the green curve in Fig. 9.5-1, was defined to be identical to the photopic luminous efficiency function $V(\lambda_0)$ displayed in Fig. 8.5-3. Incorporating luminance information into the color-matching functions formed a useful bridge between the CIE 1924 photometry (light-measurement) system and the CIE 1931 colorimetry (color-measurement) system.

Tristimulus Values. For a monochromatic stimulus, the tristimulus values XYZ can be inferred from the ordinate values in Fig. 9.5-1, much as with the LMS color space. For light of arbitrary wavelength-based power spectral density $S_\lambda(\lambda_0)$ entering the eye, the **integrated tristimulus values** are calculated by constructing inner products of the spectral density and the color-matching functions:

$$X = k\int_{\lambda_1}^{\lambda_2} S_\lambda(\lambda_0)\,\bar{x}(\lambda_0)\,d\lambda_0\,,\quad Y = k\int_{\lambda_1}^{\lambda_2} S_\lambda(\lambda_0)\,\bar{y}(\lambda_0)\,d\lambda_0\,,\quad Z = k\int_{\lambda_1}^{\lambda_2} S_\lambda(\lambda_0)\,\bar{z}(\lambda_0)\,d\lambda_0\,. \tag{9.5-1}$$

The inner products representing the integrated tristimulus values effectively serve to project an infinite-dimensional spectral-density space onto a 3D color-response space. This compressed version of the spectrum embodies Grassmann's proportionality and additivity laws, integrated over wavelength.

The integrals in (9.5-1) have the same form as that for the luminous flux P_{v} in photometry reported in (8.8-1), the key distinction being that here the three color matching functions replace the photopic luminous efficiency function $V(\lambda_0)$. In both cases, if the light entering the eye consists of a mixture of components, the overall spectral density is the sum of the spectral densities associated with the individual components.

Reflected and Transmitted Light. If the light entering the eye has been reflected from a colored object (such as the artwork displayed in Fig. 2.7-5), its spectral density is the product of that of the light incident on the object $S_{\mathrm{in}}(\lambda_0)$ and the intensity reflectance of the object $\mathcal{R}(\lambda_0)$, i.e., $S_\lambda(\lambda_0) = S_{\mathrm{in}}(\lambda_0)\,\mathcal{R}(\lambda_0)$. On the other hand, if the light entering the eye has been transmitted through a colored object or a colorant (such as stained glass or printer's ink on paper), its spectral density is the product of that of the light incident on the object $S_{\mathrm{in}}(\lambda_0)$ and the intensity transmittance of the object $\mathcal{T}(\lambda_0)$, so that $S_\lambda(\lambda_0) = S_{\mathrm{in}}(\lambda_0)\,\mathcal{T}(\lambda_0)$.

Relative Colorimetry. In **relative colorimetry**, a subclass of basic colorimetry, the normalization constant in (9.5-1) is often set to

$$k \equiv 1 \Big/ \int_{\lambda_1}^{\lambda_2} S_\lambda(\lambda_0)\,\bar{y}(\lambda_0)\,d\lambda_0\,, \tag{9.5-2}$$

which renders the XYZ tristimulus values dimensionless and fixes Y at unity. Each of the tristimulus values then falls within, or close to, the interval [0,1] and the XYZ color space is represented by something close to a unit cube. Some researchers instead fix the normalization constant at

$$k' = 100\,k\,. \tag{9.5-3}$$

The maximum value, which is $Y = 1$ or $Y = 100$ depending on the normalization chosen, then represents the brightest possible white that can be attained.

The normalized tristimulus values that represent a perceived color in XYZ color space are established by calculating the inner products of the color-matching functions and the power spectral density of the light entering the eye. For light consisting of a mixture of components, the spectral densities and tristimulus values of the components are summed, in accordance with Grassmann's laws.

Absolute Colorimetry. In **absolute colorimetry**, another subclass of basic colorimetry, the tristimulus values in (9.5-1) are usually expressed in the form of inner products between the color matching functions and a *normalized version of the power spectral density*, S_λ/P_0. Again, reconciliation of the modern and earlier definitions of the candela requires the prefactor of 683 lm/W. Given that the relations provided in (9.5-1) involve integrations only over wavelength, they can be cast in analogous forms for other spatial variants of the spectral radiant flux, such as the spectral radiant intensity, spectral irradiance, or spectral radiance. In particular, since the color matching function $\bar{y}(\lambda_0)$ and the photopic luminous efficiency function $V(\lambda_0)$ are identical, i.e.,

$$\bar{y}(\lambda_0) \equiv V(\lambda_0)\,, \tag{9.5-4}$$

the tristimulus value Y_{abs} is most conveniently expressed in terms of the spectral radiance L_λ. Referring to (8.8-4) reveals that Y_{abs} is then identical to the luminance L_{v}:

$$Y_{\mathrm{abs}} = 683 \int_{380}^{780} L_\lambda(\lambda_0)\,\bar{y}(\lambda_0)\,d\lambda_0 \equiv L_{\mathrm{v}}\,. \tag{9.5-5}$$

Tristimulus Value Y_{abs}

In absolute colorimetry, the tristimulus value Y_{abs}, with units of cd/m^2, is the sole repository of luminance information. In relative colorimetry, Y is instead fixed at $\mathbf{1}$ or $\mathbf{100}$, depending on the normalization selected, which is arbitrary.

Experimental Colorimetry. The experimental tristimulus values for an arbitrary source of light are determined by making use of an instrument known as a *tristimulus colorimeter*.

Normalized Tristimulus Values. Further normalization of the XYZ tristimulus values capacitates a convenient projection onto a plane. This mapping is achieved by dividing each tristimulus value in (9.5-1) by the sum of the three, which yields **normalized tristimulus values** xyz given by

$$x = \frac{X}{X+Y+Z}\,, \quad \text{(9.5-6a)} \qquad y = \frac{Y}{X+Y+Z}\,, \quad \text{(9.5-6b)} \qquad z = \frac{Z}{X+Y+Z}\,. \quad \text{(9.5-6c)}$$

In this case, the normalization procedures used earlier in connection with relative and absolute colorimetry are superfluous since all of the constants k in (9.5-1) cancel. The result is two independent, dimensionless parameters, x and y; the third parameter z is redundant since

$$x + y + z = 1 \quad \text{so that} \quad z = 1 - x - y. \tag{9.5-7}$$

Since the dependence on luminance is removed in the course of carrying out this normalization procedure, the (unnormalized) tristimulus value Y displayed in (9.5-5), representing the luminance, is carried along separately.

xyY Color Space. The **xyY color space** is designed as a partially normalized cross between the (unnormalized) CIE 1931 XYZ color space and the associated (normalized) tristimulus values xyz. It carries the full complement of information contained in XYZ, although in slightly different form. The dimensionless, normalized tristimulus values x and y represent chromaticity information (hue and saturation) while the orthogonal, unnormalized tristimulus value Y represents luminance. The xyY color space is the antecedent of the xy chromaticity diagram described in Sec. 9.6.

EXAMPLE 9.5-1. ***Numerical Computation of Tristimulus Values for Arbitrary Spectra.*** In the context of the XYZ color space, light with an arbitrary, continuous power spectral density $S_\lambda(\lambda_0)$ gives rise to the tristimulus values specified in (9.5-1). For light consisting of a mixture of components, the overall spectral density is the sum of the spectral densities associated with the individual components. In practice, the overall spectral density is measured with a spectrometer, the output of which is generally restricted to a set of discrete, normalized, numerical values $S(\lambda_i)$, called the **sampled spectrum**, at free-space wavelengths λ_i spaced 1 nm apart over the range 360–830 nm, as suggested by current CIE recommendations. Tables for the **sampled color-matching functions** $\bar{x}(\lambda_i)$, $\bar{y}(\lambda_i)$, and $\bar{z}(\lambda_i)$, similarly spaced 1 nm apart over the same wavelength range, are readily available on the web. The tristimulus values may then be numerically calculated via the following discretized forms of (9.5-1),

$$X \approx k \sum_{\lambda_i=360}^{830} S(\lambda_i)\,\bar{x}(\lambda_i)\,, \qquad Y \approx k \sum_{\lambda_i=360}^{830} S(\lambda_i)\,\bar{y}(\lambda_i)\,, \qquad Z \approx k \sum_{\lambda_i=360}^{830} S(\lambda_i)\,\bar{z}(\lambda_i)\,, \tag{9.5-8}$$

where $k \approx 1/\sum_{\lambda_i=360}^{\lambda_i=830} S(\lambda_i)\,\bar{y}(\lambda_i)$. The summands of X, Y, and Z are then readily computed with the help of a spreadsheet, such as EXCEL, and their sums yield estimates of the relative inner products X, Y, and Z.

EXAMPLE 9.5-2. ***Numerical Computation of xyY Values for Arbitrary Spectra.*** The numerical computation of the tristimulus values XYZ for a source of arbitrary spectrum is described in Example 9.5-1. The chromaticity coordinates x and y associated with the CIE 1931 xyY color space are determined from the X, Y, and Z values calculated in (9.5-8) by making use of (9.5-6a) and (9.5-6b). The luminance Y is obtained directly from (9.5-8).

EXAMPLE 9.5-3. ***Transformation from* LMS *to* XYZ *Color Space.*** The LMS color-matching functions displayed in Fig. 8.5-1(b) differ markedly from the XYZ color-matching functions portrayed in Fig. 9.5-1. In accordance with Grassmann's fourth law, however, a simple linear transformation, encoded in the form of a 3×3 matrix, moves one space into the other. A commonly used version of this matrix, suitable for the CIE 1931 2° standard colorimetric observer, takes the form

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 1.94735 & -1.41445 & 0.36476 \\ 0.68990 & 0.34832 & 0 \\ 0 & 0 & 1.9349 \end{bmatrix} \begin{bmatrix} L \\ M \\ S \end{bmatrix}. \tag{9.5-9}$$

Examination of (9.5-9) reveals that X is a mixture of L, M, and S whereas Z represents S. The luminance Y is a mixture of L and M in the relative proportions 0.69 and 0.35; these values are in rough accord with the curves portrayed in Fig. 8.5-1(b), which characterize the foveal cone mosaic displayed in Fig. 8.4-4 (as described in the figure caption thereof).

1976 CIELUV *Uniform Color Space*

One factor that limits the usefulness of the CIE 1931 XYZ and xyY color spaces considered above is that they are not visually uniform. In an effort to improve perceptual uniformity, in 1976 the CIE created the L*u*v* color space, commonly abbreviated CIE 1976 CIELUV, by employing coordinates that are nonlinear functions of the tristimulus values XYZ. The intent was to create a uniform color space (UCS) in which the Euclidian distance between any two points approximately represents a measure of their perceived color difference in terms of **hue**, **chroma**, and **lightness**. In particular, lightness L^* in the CIELUV color space is proportional to the cube root of the tristimulus value Y representing luminance, as in the brightness–luminance relation provided in (8.8-7) and (8.8-8).

CIELUV replaced CIE 1960 UCS, an earlier uniform color space, by virtue of its generally superior perceptual uniformity. Still, CIELUV turns out to be only moderately more perceptually uniform than CIE 1931 XYZ, which is one of the reasons that the latter endures. CIELUV is often called upon for computer-graphics applications and for self-luminous colored emitters, and is readily implemented from XYZ via a matrix transformation.

CAM16-UCS *Uniform Color Space*

Color appearance models (CAMs) must have the capacity, at a minimum, to predict the relative color appearance attributes of hue, chroma, and lightness (Sec. 9.4). In particular, they must accommodate **chromatic adaptation**, which represents the ability of the human visual system to preserve the appearance of the colors of an object under different colors of illumination. This aspect of color perception is implemented by a chromatic-adaptation transform (CAT). The **CAM16 color appearance model**, which dates from 2016, is a successor to CIECAM02 that incorporates a number of updates and improvements. CAM16 is slightly different from CIECAM16, which is expressly designed for color management systems. The prediction of brightness and colorfulness, as well as luminance-dependent effects such as the Stevens effect, requires models of greater complexity. CAMs make use of interval scales for hue but rely on ratio scales for colorfulness, saturation, chroma, brightness, and lightness.

In principle, the 1976 CIELUV color space considered earlier is a CAM since the source and stimulus chromaticities provided at its input yield predictions for hue, chroma, and lightness at its output. In practice, however, CIELUV is not a viable CAM since its (subtractive) chromatic-adaptation transform is physiologically unrealistic and its performance in predicting color differences is poor.

CAM16 is accompanied by a color space called **CAM16-UCS**. This is the updated, current version of a succession of CIE color spaces of increasing subtlety and uniformity that have been developed since the advent of 1976 CIELUV. Extensive studies have demonstrated that CAM16-UCS is highly reliable for predicting color differences for both the CIE 2° and 10° observers. Based on a large number of individual datasets, its performance has been determined to be superior to that of other color-difference formulas.[†]

Accurate color-difference determination is essential for assessing metamerism (Sec. 9.3), and, as will be seen in the sequel, for determining correlated color temperature (Sec. 9.8) and color rendering index (Sec. 9.9). CAMs and their associated color spaces have specific constraints regarding viewing conditions, however, so it is important to assess their usefulness for particular LED lighting applications.

RGB *Color Spaces*

The RGB color model is frequently used to achieve additive color mixing since red, green, and blue are convenient primaries and their addition generates a wide range of colors that include a significant portion of the gamut of human vision. We examine a number of additive RGB color spaces based on the RGB color model.

CIE 1931 RGB Color Space. The earliest RGB color space was constructed by the CIE in 1931 on the basis of the extensive color-matching experiments conducted in the late 1920s by the British scientists W. David Wright and John Guild (p. 265), working at Imperial College and at the National Physical Laboratory, respectively.[‡] Both series of experiments made use of multiple observers. Some of the experiments relied on monochromatic primaries at various wavelengths and intensities, along the lines of the approach portrayed in Fig. 9.1-1, while others used broadband primaries. Ultimately, the monochromatic $\mathfrak{RGB}$ primary wavelengths chosen for the CIE standard were 700, 546.1, and 435.8 nm, respectively; these were convenient because they made use of the strong green and blue emission lines from a Hg-vapor discharge. The associated color matching functions, $\bar{r}\bar{g}\bar{b}$, were used to define the **1931 CIE standard colorimetric observer**.

However, the use of these primaries led one of the CIE 1931 RGB color matching functions to be negative over a region of wavelengths. This is related to the fact that monochromatic primaries cannot uniquely and severally stimulate the three types of cones in the human retina because of the substantial overlap in their spectral sensitivity curves (Fig. 8.5-1). Moreover, none of the color matching functions provided an explicit representation for luminance, which was desirable. These two deficits were rectified by the mathematical construction of CIE 1931 XYZ, which was created concomitantly with CIE 1931 RGB. These two CIE 1931 color spaces offered the first quantitative linkages between the visible wavelengths of the electromagnetic spectrum and the physiological perception of color. The transformation from RGB to XYZ (Example 9.5-4) results in a tristimulus value Y that is proportional to the luminance L_{v}, as desired.

[†] M. R. Luo, Q. Xu, M. Pointer, M. Melgosa, G. Cui, C. Li, K. Xiao, and M. Huang, A Comprehensive Test of Colour-Difference Formulae and Uniform Colour Spaces Using Available Visual Datasets, *Color Research and Application*, DOI:10.1002/col.22844, 2023.

[‡] W. D. Wright, A Re-Determination of the Trichromatic Coefficients of the Spectral Colours, *Transactions of the Optical Society (London)*, vol. 30, pp. 141–164, 1929; J. Guild, The Colorimetric Properties of the Spectrum, *Philosophical Transactions of the Royal Society of London*, vol. A230, pp. 149–187, 1931.

EXAMPLE 9.5-4. ***Transformation from CIE 1931* RGB *to CIE 1931* XYZ *Color Space.*** The matrix that encodes the transformation from CIE 1931 RGB to CIE 1931 XYZ color space is expressible as

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.49000 & 0.31000 & 0.20000 \\ 0.17697 & 0.81240 & 0.01063 \\ 0 & 0.01000 & 0.99000 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \tag{9.5-10}$$

The tristimulus value X is a mixture of R, G, and B that is chosen to be nonnegative; Y, which represents the luminance, is given by $Y \approx 0.18R+0.81G+0.01B$, indicating that B plays a negligible role; and $Z \approx B$. Implementing this transformation trades the color matching function $\bar{r}(\lambda_0)$, which is negative over a region of wavelengths, for the color matching function $\bar{x}(\lambda_0)$, which is always nonnegative (Fig. 9.5-1), but is bimodal and is associated with an imaginary primary.

RGB Color-Space Variants. Advances in digital electronics and photonics in the late twentieth and early twenty-first centuries fostered the development of many specialized RGB color spaces for various applications. Examples include displays for computer monitors, smartphones, and television receivers, and, of course, LED lighting. The specification of color for OLED displays (Sec. 7.6) and LED lighting (Sec. 11.3) is established by the excitations applied to physical red, green, and blue light emitters. Backlit LCD displays also produce red, green, and blue light.

RGB color spaces can be represented in the form of unit cubes, in which the R, G, and B coordinates span the interval [0,1]. These spaces are constructed using particular sets of $\mathfrak{RGB}$ primaries, with cyan, magenta, and yellow serving as secondaries. The color solid representing one such space is displayed in Figs. 9.5-2(a),(b) from perspectives focusing on the coordinates (1,1,1) and (0,0,0), which are white and black, respectively. A multitude of RGB color spaces that make use of different primaries are in common use. RGB color spaces can be **device-dependent** or **device-independent**, designations indicating whether the resultant color does or does not depend on the system used for display.

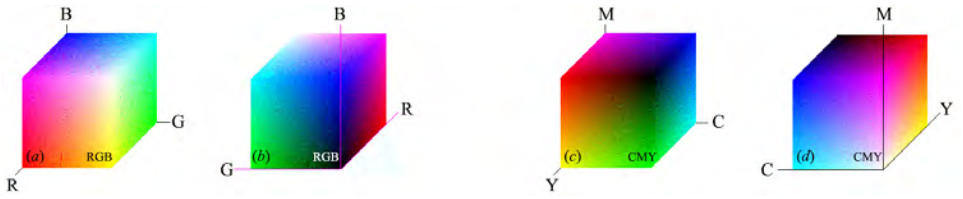


Figure 9.5-2 (a) RGB color solid from a perspective that focuses on (1,1,1), which is white. (b) RGB color space from a perspective that focuses on (0,0,0), which is black. (c) CMY color solid from a perspective that focuses on (1,1,1), which is black. (d) CMY color solid from a perspective that focuses on (0,0,0), which is white.

CMY Color Space. The color solid representing the CMYK color space, which is generally used in color printing, is also represented by a unit cube with coordinates that span the interval [0,1]. The primary colors are $\mathfrak{CMY}$ (cyan, magenta, and yellow), and the secondaries are red, green, and blue. Unlike RGB, which is additive, this color space is subtractive. The color solid is displayed in Figs. 9.5-2(c),(d) from perspectives focusing on the coordinates (1,1,1) and (0,0,0), which are black and white, respectively. The coordinate (1,1,1) is black since C, M, and Y allow only R, G, and B to be transmitted, respectively, so no residual color remains at (1,1,1). In principle, C, M, and Y are sufficient to attain black, but in practice it turns out that mixing equal components of

these three inks yields dark brown instead, so black (K) is usually added as an adjunct. CMYK color spaces are usually device-dependent.

Device-Centric to Human-Centric Transformations. Display colors are established by specifying the voltages or currents to be applied to individual photonic devices or pixels to generate red, green, and blue light. As discussed in Sec. 9.2, however, human color perception is more suitably described by hue, saturation, and luminance. For applications such as computer graphics, a Cartesian RGB color space (such as sRGB) is readily transformed to a human-centric configuration such as HSV or HSL, which are often referred to as color models. The colorimetric properties of the transformed space are related to those of the color space from which it is derived. HSV and HSL color models are briefly considered in turn. They are often preferred by those working in the visual arts, for whom hue and saturation are intuitive, and who traditionally create color using **tints** (mixtures with white), **shades** (mixtures with black), and **tones** (mixtures with both).

HSV Color Model. The HSV (hue, saturation, and value) color model, also known as HSB (hue, saturation, and brightness), serves as an alternative to RGB because it is easier to visualize and is device-independent. A color solid frequently used to represent this color model is the *hexcone* displayed in Fig. 9.5-3. The primary and secondary colors (red, yellow, green, cyan, blue, and magenta) are represented at the six vertices of its hexagonal base, which appears at the top of the figure. HSV is also often represented in the form of a right circular cylinder.

Figure 9.5-3 Hexcone color solid representing the HSV color model. The three HSV variables (hue, saturation, and value) are represented by azimuthal angle (0° is defined as red, 120° is green, and 240° is blue); radial distance; and height; respectively. Saturation stretches from zero at the hexcone axis to unity at its surface. Value, representing brightness or relative luminance, extends along the hexcone axis, from zero at the apex to unity at the base.

HSL Color Model. HSL (hue, saturation, and lightness) is closely related to HSV, with lightness replacing brightness. A perfectly light color in HSL is pure white whereas a perfectly bright color in HSV results from shining a white light on a colored object. HSL serves as another alternative to RGB. Its color solid is cast in the form of a *double hexcone* or a right circular cylinder.

9.6 CHROMATICITY DIAGRAMS

The color spaces considered in Sec. 9.5 are 3D configurations that portray the collection of supported colors representing the tristimulus values of their coordinates. However, visualizing these colors is challenging. One strategy for improving the visualization and interpretation of colorimetric data is to make use of chromaticity diagrams that reduce the 3D data to 2D planar images by removing luminance information. In this section, we describe the chromaticity diagrams associated with the color spaces discussed in Sec. 9.5.

We begin by displaying early 2D chromaticity diagrams constructed by Newton (p. 1) in 1730, Maxwell (p. 24) in 1860, and Helmholtz (p. 234) in 1855. Following the introduction of these early constructs, we proceed to describe the CIE 1931 xy

chromaticity diagram that is ubiquitous today, as well as the CIE 1976 $u'v'$ diagram. Finally, we present a number of RGC diagrams, which are customarily portrayed on the xy-diagram template.

Early Chromaticity Diagrams. Chromaticity diagrams dating from 1730 (Newton's color circle), 1860 (Maxwell's color triangle), and 1855 (Helmholtz' color horseshoe) are schematized in Fig. 9.6-1.

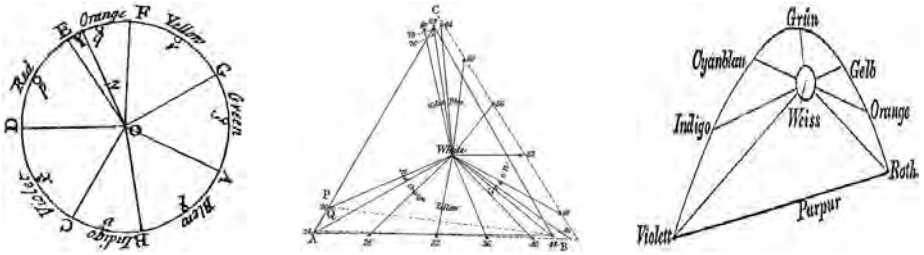


Figure 9.6-1 Early chromaticity diagrams. *Left*: Newton's color circle (1730). The circumference of the circle is divided into arcs labeled with the spectral colors, whose centers are designated p, q, r, s, t, v, and x. *Indigo* is now seldom used as a color name and *blew* is now blue. The center of the circle at o is presumed to be white. Nonspectral colors, such as the one designated at point z, are described by their distance from o and from the corresponding spectral color labeled Y. *Center*: Maxwell's color triangle (1860). Red, green, and blue primary lights associated with Maxwell's trichromatic theory are located at the corners of the triangle, and white is at the center. Mixing the primaries in various proportions yields the colors represented along the edges of the triangle as well as in its interior. *Right*: Helmholtz' chromaticity diagram (1855) constructed using his measurements of complementary colors. In modern German, roth is written as rot. Helmholtz' diagram closely resembles modern chromaticity diagrams in shape, and even includes a line of purples, as will become apparent in the next section.

CIE 1931 xy Chromaticity Diagram

As discussed in Sec. 9.5, the CIE 1931 XYZ color space, together with its close cousin, the xyY color space, serve as the *linguae francae* of color spaces because of their convenience and widespread use. The tristimulus values XYZ, normalized to their relative values xy via (9.5-6), carry the chromaticity information (hue and saturation), while Y represents the luminance, which is principally governed by the external level of illumination.

The representation of chromaticity in this color space is therefore reduced from three dimensions to two; the collection of all perceptual colors is represented in a 2D plane. The xy **chromaticity diagram**, where x and y are the (dimensionless) chromaticity coordinates, is displayed in Fig. 9.6-2, and its properties are delineated below. Inasmuch as Grassmann's empirical laws (Sec. 9.2) played a central role in the development of the CIE 1931 XYZ color space, they are *de facto* incorporated into the fabric of the xy chromaticity diagram.

Properties. The xy chromaticity diagram has the following properties:

- It encompasses the **full gamut of color vision**, comprising some 4000 gradations that consist of 200 hues, each with 20 saturation levels. Light of a particular color is specified in terms of its (x, y) chromaticity coordinates.
- The outer curved boundary represents the fully saturated *locus of spectral colors*, whose associated wavelengths (in nm) are indicated at the periphery.

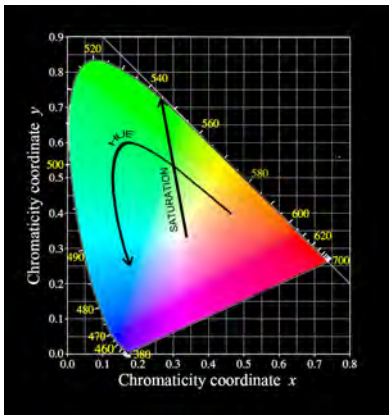


Figure 9.6-2 The xy chromaticity diagram associated with the CIE 1931 xyY color space for a standard observer, under specified conditions of illumination. The chromaticity coordinates are denoted x and y. In rough analogy with a polar coordinate system, hue (color) and saturation (colorfulness) are represented in the polar and radial directions, respectively, with the origin at the white center. The luminance (brightness) information associated with Y is not represented in the diagram. The unusual horseshoe shape of the image derives from the form of the XYZ color matching functions. Were we to extend the thin white line tangent to the diagram beyond the boundaries of the figure, it would intersect the abscissa at $x = 1.0$ and the ordinate at $y = 1.0$.

- The straight edge at the bottom of the figure is called the **line of purples**. Its colors are fully saturated but nonspectral, i.e., they have no counterparts in monochromatic light and can be generated only by mixing red and blue.
- Any color, including white, that lies on a straight line between any two points in the diagram can be generated by mixing the colors at the endpoints of that line. The relative weights of the two endpoint contributions required to generate a particular color depend on factors beyond the geometrical distance of the desired color along the line (Example 9.6-1).
- All colors lying within a triangle in the diagram can be generated by mixing the three colors represented by the vertices of that triangle (Example 9.6-2). Similarly, all colors within any simple polygon can be generated by mixing the colors at its vertices.
- White light with a uniform wavelength-based spectral density (**equal-energy white**) has the chromaticity coordinates $(x, y) = (1/3, 1/3)$ (Example 9.6-3).
- Less saturated colors appear in the interior of the diagram, with white toward the center. Mixing a spectrally pure color with white leads to a color with the same hue but different saturation. For example, pure red (100% saturated) mixed with white leads to pink ($< 100\%$ saturated), and ultimately to white (0% saturated).
- **Complementary colors** lie on opposite sides of white, and along every line that passes through white.
- The **dominant wavelength** for a color in the interior of the diagram is established by drawing a straight line through the white point and that color, and then determining where on the boundary the extension of that line intersects the locus of spectral colors. The **color purity** is defined as the distance from the white point to the color divided by the distance from the white point to the boundary. Spectral colors, which lie on the boundary, have a color purity of unity.
- A nonspectral color that lies within a triangle whose vertices are at the white point and at the two bottom angular corners of the diagram are conventionally identified by the dominant wavelength of its complement.
- Spectral colors such as orange can be converted to brown by reducing the luminance of the red and green components, which is tantamount to mixing it with black.

EXAMPLE 9.6-1. ***Chromaticity Coordinates for a Mixture of Two Colors.*** We consider the mixing of two sources of light with peak wavelengths λ_1 and λ_2; power spectral densities $S_1(\lambda_0)$ and $S_2(\lambda_0)$; and chromaticity coordinates (x_1, y_1) and (x_2, y_2), respectively. We assume that the spectral widths of the two components are much narrower than the XYZ color matching functions

$\bar{x}(\lambda_0)$, $\bar{y}(\lambda_0)$, and $\bar{z}(\lambda_0)$, so that their spectral densities can be approximated by delta functions. The overall power spectral density can then be written as

$$S_\lambda(\lambda_0) \approx P_1\delta(\lambda - \lambda_1) + P_2\delta(\lambda - \lambda_2), \tag{9.6-1}$$

where P_1 and P_2 represent the optical powers (radiant flux) emitted by the two sources, respectively. Inserting this expression into (9.5-1) then yields, with the help of the sifting property of the delta functions in the integrands,

$$X \approx k[P_1\bar{x}(\lambda_1) + P_2\bar{x}(\lambda_2)] \tag{9.6-2a}$$

$$Y \approx k[P_1\bar{y}(\lambda_1) + P_2\bar{y}(\lambda_2)] \tag{9.6-2b}$$

$$Z \approx k[P_1\bar{z}(\lambda_1) + P_2\bar{z}(\lambda_2)]\,. \tag{9.6-2c}$$

Using (9.5-6a) and (9.5-6b), along with the definitions

$$K_1 = P_1\,[\bar{x}(\lambda_1) + \bar{y}(\lambda_1) + \bar{z}(\lambda_1)] \tag{9.6-3a}$$

$$K_2 = P_2\,[\bar{x}(\lambda_2) + \bar{y}(\lambda_2) + \bar{z}(\lambda_2)] \tag{9.6-3b}$$

then provides

$$x = \frac{x_1K_1 + x_2K_2}{K_1 + K_2} \qquad \text{and} \qquad y = \frac{y_1K_1 + y_2K_2}{K_1 + K_2}\,. \tag{9.6-4}$$

We conclude that the chromaticity coordinates for the combined light are linear combinations of the chromaticity coordinates for the individual sources, suitably weighted by K_1 and K_2, which in turn are governed by the relative radiant flux and values of the color matching functions at the peak wavelengths of the individual sources. The two sources can be located anywhere within the chromaticity diagram; they need not be on its boundary nor does the line connecting them need to transect the white point.

It follows from (9.6-4) that the chromaticity coordinates of the mixture fall along the straight line connecting the chromaticity coordinates of the individual sources. This approach is used, for example, in determining the chromaticity coordinates for a white phosphor-conversion (PC) LED (Example 10.5-3). It is also invaluable for establishing the chromaticity coordinates for sources of finite linewidth (Sec. 9.6).

EXAMPLE 9.6-2. ***Chromaticity Coordinates for a Mixture of Three Colors.*** A generalization of Example 9.6-1 considers the mixing of three light sources that have peak wavelengths λ_1, λ_2, and λ_3; power spectral densities $S_1(\lambda_0)$, $S_2(\lambda_0)$, and $S_3(\lambda_0)$; and chromaticity coordinates (x_1, y_1), (x_2, y_2), (x_3, y_3), respectively. When the widths of the individual spectral components are much narrower than the XYZ color matching functions $\bar{x}(\lambda_0)$, $\bar{y}(\lambda_0)$, and $\bar{z}(\lambda_0)$, the spectral densities can be approximated by delta functions. The overall spectral density can then be written as

$$S_\lambda(\lambda_0) \approx P_1\delta(\lambda - \lambda_1) + P_2\delta(\lambda - \lambda_2) + P_3\delta(\lambda - \lambda_3), \tag{9.6-5}$$

where the quantities $P_{1,2,3}$ represent the radiant flux of the three sources. With the help of the sifting property of the delta function, inserting (9.6-5) for $S_\lambda(\lambda_0)$ into (9.5-1) then provides

$$X \approx k[P_1\bar{x}(\lambda_1) + P_2\bar{x}(\lambda_2) + P_3\bar{x}(\lambda_3)] \tag{9.6-6a}$$

$$Y \approx k[P_1\bar{y}(\lambda_1) + P_2\bar{y}(\lambda_2) + P_3\bar{y}(\lambda_3)] \tag{9.6-6b}$$

$$Z \approx k[P_1\bar{z}(\lambda_1) + P_2\bar{z}(\lambda_2) + P_3\bar{z}(\lambda_3)]\,. \tag{9.6-6c}$$

Now, using (9.5-6a) and (9.5-6b), together with the definitions

$$K_1 = P_1\bar{x}(\lambda_1) + P_1\bar{y}(\lambda_1) + P_1\bar{z}(\lambda_1) \tag{9.6-7a}$$

$$K_2 = P_2\bar{x}(\lambda_2) + P_2\bar{y}(\lambda_2) + P_2\bar{z}(\lambda_2) \tag{9.6-7b}$$

$$K_3 = P_3\bar{x}(\lambda_3) + P_3\bar{y}(\lambda_3) + P_3\bar{z}(\lambda_3), \tag{9.6-7c}$$

leads to

$$x = \frac{x_1K_1 + x_2K_2 + x_3K_3}{K_1 + K_2 + K_3} \qquad \text{and} \qquad y = \frac{y_1K_1 + y_2K_2 + y_3K_3}{K_1 + K_2 + K_3}\,. \tag{9.6-8}$$

Again, the chromaticity coordinates of the mixed light are linear combinations of the chromaticity coordinates of the individual sources, suitably weighted by factors that depend on the values of the color matching functions at the three peak wavelengths, and on their relative radiant flux. It follows from (9.6-8) that the chromaticity coordinates of the mixture lie within a triangle on the diagram whose vertices are located at the coordinates associated with the three constituent sources. Equation (9.6-8) reduces to (9.6-4) when only two colors are mixed. A generalization of this example leads to the result that the chromaticity coordinates of a mixture of multiple colors lies within the simple polygon on the chromaticity diagram whose vertices are located at the coordinates associated with the constituent sources.

EXAMPLE 9.6-3. ***Chromaticity Coordinates for Spectrally Uniform White Light.*** Light with a uniform wavelength-based power spectral density $S_\lambda(\lambda_0)$, often called **equal-energy white** and referred to as the standard CIE colorimetric illuminant E, appears white to the eye. Because the color matching functions $\bar{x}(\lambda_0)$, $\bar{y}(\lambda_0)$, and $\bar{z}(\lambda_0)$ have equal areas, the three integrals represented in (9.5-1) are identical, whereupon $X = Y = Z$. In accordance with (9.5-6), this leads to $x = y = z = 1/3$. We conclude that uniform-spectral-density white light is represented by the chromaticity coordinates $(x, y) = (1/3, 1/3)$, which is known as the **white point** of the chromaticity diagram.

CIE 1976 $u'v'$ Chromaticity Diagram

An examination of the xy chromaticity diagram presented in Fig. 9.6-2 reveals that an unexpectedly large portion of its area is occupied by green. In an attempt to redress this anomaly, which is a manifestation of the extensive nonuniformity throughout the diagram, in 1976 the CIE introduced the CIELUV (or CIE L*u*v*) color space (Sec. 9.5). The goal was to create a uniform color space (UCS) in which the Euclidian distance between any two points in the associated $u'v'$ chromaticity diagram would approximately represent their perceived color difference in terms of **hue, chroma, and lightness**.

The $u'v'$ chromaticity diagram associated with the CIELUV color space is displayed in Fig. 9.6-3. Its chromaticity coordinates (u', v') are readily obtained from the coordinates (x, y) associated with CIE 1931 xyY via the mapping specified in Example 9.6-4. Although 1976 CIELUV replaced the earlier CIE 1960 UCS by virtue of its generally superior perceptual uniformity, the uv chromaticity diagram associated with the latter is nevertheless considered to be more suitable for determining correlated color temperature (Sec. 9.8).

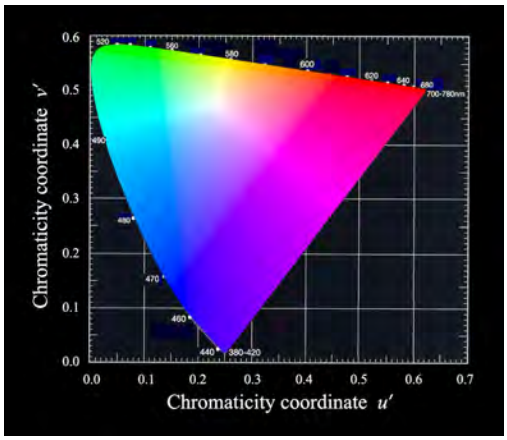


Figure 9.6-3 The $u'v'$ chromaticity diagram associated with the 1976 CIELUV (CIE 1976 UCS) color space for a standard observer. Although designated as a uniform color space (UCS), CIELUV is only moderately more perceptually uniform than CIE 1931 XYZ. Nevertheless, comparing the 1976 $u'v'$ diagram displayed here with the 1931 xy chromaticity diagram portrayed in Fig. 9.6-2 reveals that the proportions of blue and red are suitably enlarged in 1976 $u'v'$.

EXAMPLE 9.6-4. ***Mapping (x,y) to (u′,v′) Chromaticity Coordinates.*** The mapping of (x, y) to (u', v') chromaticity coordinates takes the following form:

$$u' = \frac{4x}{-2x + 12y + 3}, \qquad v' = \frac{9y}{-2x + 12y + 3}. \tag{9.6-9}$$

RGB *Chromaticity Diagrams*

Chromaticity diagrams that illustrate the ranges of reproducible colors for several RGB color spaces are traced out on the CIE 1931 xy diagram presented in Fig. 9.6-4. The ranges of perceptible colors are represented by their gamuts, i.e, by the colors enclosed within their 2D triangles. We portray three chromaticity diagrams commonly used for color displays. The standard RGB (sRGB) diagram, introduced by Hewlett-Packard and Microsoft in 1996 for use with digital devices, along with the Apple RGB diagram used for Apple devices, are seen to have gamuts that are somewhat limited, particularly in the blue and green. Adobe RGB has a larger gamut and is designed to accommodate most colors available with CMYK color printers. Adobe wide-gamut RGB (not plotted in Fig. 9.6-4) makes use of pure spectral primaries with wavelengths 700, 525, and 450 nm, and offers a substantially enhanced gamut.

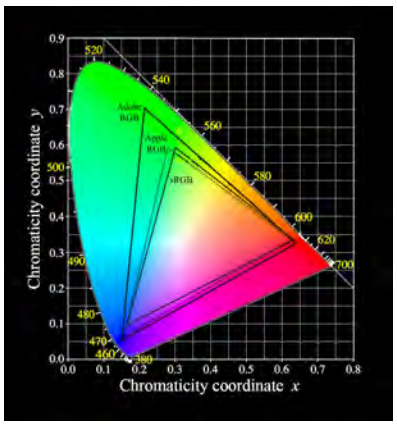


Figure 9.6-4 RGB chromaticity diagrams traced out on the CIE 1931 xy diagram. Diagrams for three device-dependent RGB color spaces are presented: sRGB, Apple RGB, and Adobe RGB. A universal set of $\mathfrak{RGB}$ primaries does not exist so each of the color spaces defined in the figure has its own set, which are identified by the chromaticity coordinates of the vertices of their triangles. The gamut of reproducible colors associated with each triangle is a subset of the gamut of the xy diagram, the properties of which were considered in the discussion surrounding Fig. 9.6-2. Color-mixing LEDs with individually addressable red, green, and blue dies define triangles that designate the available gamut for LED lighting with tunable hue and saturation (Sec. 11.3).

Although all three of these RGB diagrams offer a substantial range of colors, they are clearly subsets of the xy chromaticity diagram, which accommodates the full range of human color perception. Enveloping the full gamut using three primaries entails using a triangle whose vertices lie outside the region of real colors, thereby corresponding to imaginary primaries. ProPhoto RGB is one such example.

RGB color spaces are widely used in connection with LED lighting. As will be elucidated in Chapter 10, individual red, green, and blue LEDs enable the creation of RGB color systems that accommodate the generation of light with tunable hue, saturation, and luminance. Discrete LEDs of many different colors are commercially available, as are additive color-mixing LEDs that contain individually addressable red, green, blue, and white dies within a single LED package (Secs. 11.2 and 11.3).

Chromaticity Coordinates for Sources of Finite Linewidth. Although the emission spectrum associated with an individual LED is relatively narrow, it is not vanishingly small. Thermal broadening results in a spectral width $\Delta\lambda \approx 1.45\,\lambda_p^2\,kT$, as

provided in (6.4-14) and (6.4-17), and as illustrated in Example 7.2-1. Other mechanisms, such as alloy and various forms of inhomogeneous broadening, also play a role when present. The XYZ tristimulus values for LED light, which are established by making use of (9.5-1), depend on the overall power spectral density $S_\lambda(\lambda_0)$ of the light entering the eye. For a red LED, the tristimulus values typically give rise to xy chromaticity coordinates that lie at a point along the red boundary of the chromaticity diagram, while for a blue LED the coordinates usually lie at a point interior, but quite close to, the boundary. For a green LED, on the other hand, the coordinates typically lie in the interior of the diagram.

These distinct outcomes can be understood on the basis of the local curvature of the chromaticity diagram. Figure 9.6-2 and Example 9.6-1 reveal that mixing colors corresponding to any two points on the diagram results in a color that lies on the straight line connecting those two points. Since the boundary of the chromaticity diagram in the red region is essentially a straight line, the assembly of red spectral components that comprise the broadened spectrum of a red LED all lie along this same straight boundary, and therefore so too do the chromaticity coordinates of mixtures of all pairs of spectral components. The same reasoning applies to the yellow photoluminescence generated in a white phosphor-conversion (PC) LED, as discussed in Example 10.5-3.

In contrast, the assembly of green spectral components comprising the broadened spectrum of a green LED lie along the sharply curved green boundary of the chromaticity diagram. If a subset of three green spectral components drawn from the broadened spectrum are considered as the vertices of a triangle, the chromaticity coordinates for their mixture lie within that triangle, as explained in Example 9.6-2. By induction, the chromaticity coordinates for the full collection of spectral components that comprise the broadened spectrum of a green LED lie in the interior of the xy diagram. The results for a blue LED are intermediate between those for red and green LEDs.

9.7 COLOR TEMPERATURE

The color of the light emitted by a source in thermal equilibrium depends solely on its thermodynamic temperature T. This temperature therefore serves as a convenient shorthand for defining its color, which is referred to as the **color temperature (CT)** and is specified in kelvins. Color temperature is defined only for thermal light.

Color temperature endures as a measure in illumination engineering because humans prefer thermal-light illumination. This preference likely arose because the principal source of light at the surface of the earth, sunlight, has a spectrum that closely follows the blackbody radiation law with $T \approx 5800$ K (Sec. 4.7). Indeed, with the help of Wien's law (4.7-14), it is readily shown that this temperature corresponds to a peak wavelength $\lambda_p \approx 500$ nm, which is close to the peak wavelengths of the trichromatic scotopic and photopic luminous efficiency functions, $V'(\lambda_0)$ and $V(\lambda_0)$, respectively (Sec. 8.5). It can be plausibly argued that the comfort of thermal light, such as that from an incandescent lamp, is linked to the blackbody spectrum.

We begin by describing the spectral characteristics of thermal light and then explain how color temperature is represented on the chromaticity diagram. We proceed by considering color temperature in the context of other temperature measures, namely thermodynamic, thermographic, and biological temperature. We conclude with a few words regarding the significance of the terms *warm white light* and *cool white light*.

Spectral Radiance, Irradiance, and Density for Thermal Light

The wavelength-based spectral radiance $L_\lambda(\lambda, T)$ of a blackbody source of temperature T, which represents its power per unit wavelength per unit projected area per unit solid

angle, takes the form

$$L_\lambda(\lambda,T) = \frac{2hc^2}{\lambda^5}\,\frac{1}{\exp(hc/\lambda kT)-1}\,. \tag{9.7-1}$$

Spectral Radiance (Blackbody Source)

Blackbody radiation is isotropic and its spectral radiance L_λ (W·nm^{-1}·m^{-2}·sr^{-1}) is related to its wavelength-based spectral energy density ϱ_λ specified in (4.7-10) (J·nm^{-1}·m^{-3}) via $L_\lambda(\lambda,T) = (c/4\pi)\,\varrho_\lambda(\lambda,T)$, where c is the speed of light and T is the thermodynamic temperature of the source. Also closely related are the spectral irradiance $I_\lambda(\lambda,T)$ and the wavelength-based power spectral density $S_\lambda(\lambda,T)$ illustrated in Fig. 2.7-5.

These four measures are interrelated by the following expressions:

- Spectral Radiance: $L_\lambda(\lambda,T) = (c/4\pi)\,\varrho_\lambda(\lambda,T)$ (W·nm^{-1}·m^{-2}·sr^{-1}) (9.7-2a)
- Spectral Irradiance: $I_\lambda(\lambda,T) = (c/4)\,\varrho_\lambda(\lambda,T)$ (W·nm^{-1}·m^{-2}) (9.7-2b)
- Spectral Density: $S_\lambda(\lambda,T) = (cA_{\text{eff}}/4)\,\varrho_\lambda(\lambda,T)$ (W·nm^{-1}), (9.7-2c)

where A_{eff} represents the effective projected area. Since the four measures tabulated in (9.7-2) are mutually proportional, i.e., $S_\lambda(\lambda,T) \propto I_\lambda(\lambda,T) \propto L_\lambda(\lambda,T) \propto \varrho_\lambda(\lambda,T)$, their spectral dependencies are identical. It is customary to plot the spectral radiance $L_\lambda(\lambda,T)$, as displayed in Fig. 9.7-1(a), but the curves represent all four quantities, with a simple change of scale on the ordinate. Each curve in this figure, corresponding to a specified thermodynamic temperature T(K), is a smooth, single-peaked function of the free-space wavelength λ_0 that extends from the ultraviolet (UV) to the mid-infrared.

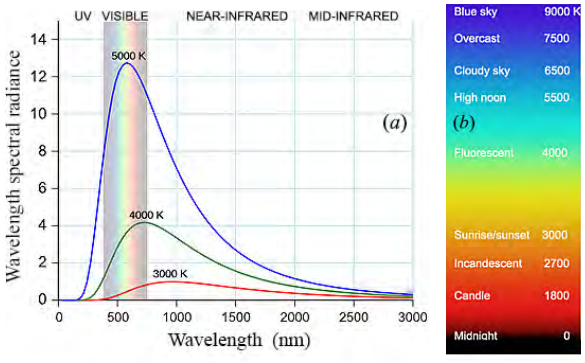


Figure 9.7-1 (a) The spectral radiance $L_\lambda(\lambda,T)$ of a blackbody (or graybody) radiator is the power radiated per unit wavelength per unit projected area per unit solid angle (kW·nm^{-1}·m^{-2}·sr^{-1}). The curves are parameterized by T (K) and plotted against λ_0. The spectral behavior, and therefore the shapes, of $S_\lambda \propto I_\lambda \propto L_\lambda \propto \varrho_\lambda$ are the same; only their scales differ. (b) Objects whose color temperatures stretch from low (red) to high (blue) values.

Chromaticity Coordinates and the Planckian Locus

Although some sense of the color of thermal light at various temperatures can be gleaned from Fig. 9.7-1(a), a reliable assessment of the color involves coupling the spectral density to the visual system's color matching functions. This is achieved by generating the **Planckian locus**, a curve on the chromaticity diagram that comprises the collection of chromaticity coordinates that correspond to thermal radiation at different thermodynamic temperatures. The procedure for calculating the Planckian locus is as follows:

1. Begin with the power spectral density of thermal light $S_\lambda(\lambda, T)$ provided in (9.7-2c), which is proportional to the spectral energy density $\varrho_\lambda(\lambda, T)$ set forth in (4.7-10).
2. Use (9.5-1) to form inner products of the spectral density $S_\lambda(\lambda, T)$ with the color matching functions $\bar{x}(\lambda, T), \bar{y}(\lambda, T), \bar{z}(\lambda, T)$, which lead to the tristumulus values $X(T), Y(T), Z(T)$.
3. Use (9.5-6) to normalize the tristimulus values, which provides $x(T)$ and $y(T)$ as functions of temperature.
4. Plot $x(T)$ and $y(T)$ on the chromaticity diagram for a collection of values of T, labeling them with the temperature. This is the Planckian locus.

The Planckian locus, which tracks the path of the color temperature for the radiation emitted by blackbodies and graybodies at thermodynamic temperature T (K), is represented by the black curve imposed on the xy chromaticity diagram in Fig. 9.7-2, which is based on the one in Fig. 9.6-2. The color temperature transitions from deep red at a low thermodynamic temperature ($T \approx 1000$ K), to orange, to yellow, to yellowish-white, then to white, and ultimately to bluish-white and blue at a sufficiently high thermodynamic temperature ($T \approx 10000$ K). While the shift in color with increasing thermodynamic temperature can be qualitatively understood by observing that the spectral radiance curves in Fig. 9.7-1(a) broaden with increasing temperature, thereby increasing the relative proportions of yellow and blue, the Planckian locus provides a quantitative accounting.

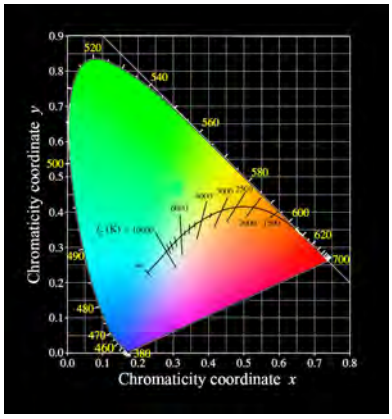


Figure 9.7-2 The xy chromaticity diagram associated with the CIE 1931 xyY color space, as displayed in Fig. 9.6-2. The black curve plotted on the figure is the Planckian locus, which traces the path of colors associated with the radiation emitted by a blackbody or graybody as the temperature T (K) changes. The straight lines that transect the Planckian locus are representative loci of constant correlated color temperature T_c (K), as explained in Sec. 9.8. These loci are more clearly portrayed in Fig. 9.8-1, which displays them on the uv chromaticity diagram associated with the CIE 1960 color space.

The exemplars of color exhibited in Fig. 9.7-1(b) represent a collection of objects whose color temperatures stretch from low (red) to high (blue). Colors such as green and violet, which are well away from the Planckian locus, are never elicited by thermal radiation, regardless of how high or how low their temperature.

Incandescent Light

From the late 1800s until the early 2000s, artificial lighting was principally provided by sources of incandescent thermal light, the generation and properties of which have been detailed in Sec. 4.8 The quintessential incandescent thermal source is the old-fashioned glass light bulb containing a thin tungsten filament that is ohmically heated by an electric current. Tungsten is used for such lamps because, of all metals, it has the lowest vapor pressure (1 Pa at 3477 K) and highest melting point (3695 K). In some devices, thermal losses and material evaporation are limited by filling the bulb with a protective noble gas and winding the filament into a compact coil. The emissivity of a heated tungsten filament has a value $\varepsilon \approx 0.44$ across the visible region of the

spectrum (Table 4.8-1), so it radiates as a graybody and is well-characterized by the Planck radiation formula (4.7-10).

Tungsten lamps are typically operated at temperatures between 2600 and 3300 K; below 2600 K the light is overly reddish in color with barely a hint of white, and above 3300 K the filament is liable to melt. As is evident from the curve in Fig. 9.7-1(a) labeled 3000 K, as well as from Fig. 9.7-1(b) and from the Planckian locus on Fig. 9.7-2, incandescent emission from tungsten is limited to the reddish end of the visible spectrum. Tungsten incandescent sources therefore cannot produce bright white light. They have lifespans of some 1500 hours and emit only about 5% of the energy they consume as visible light (with a corresponding wall-plug luminous efficiency $\eta_{\text{WPC}} \approx 2\%$), the remainder being dissipated as heat in the form of infrared radiation. Halogen lamps have lifespans of about 4000 hours and emit about 7% of the energy they consume as visible light. In spite of their drawbacks, tungsten lamps have nevertheless continued to serve as a point of reference in illumination engineering because of their long history, ease of construction, low cost, and ideal color rendering quality.

While incandescent artificial lighting is appealing to the eye, its color is limited to shades of reddish-white and its generation is hampered by low efficiency.

Temperature: Color, Thermodynamic, Thermographic, and Biological

The concept of color temperature is clarified by juxtaposing it with three other temperature measures, all of which represent energy in one form or another:

Temperature Measures: Energy Forms and Measurement Modalities

- *Color Temperature*. Spectral density of thermal radiant energy in the visible region, measured by retinal cones and interpreted by the visual system.
- *Thermodynamic Temperature*. Average internal energy of a system, measured by a thermometer.
- *Thermographic Temperature*. Power per unit area of thermal radiant energy in the infrared region, measured by an infrared photodetector array.
- *Biological Temperature*. Average thermal energy of ambient air molecules, measured by skin thermoreceptors.

We consider each of these temperature measures in turn.

Color Temperature. A source of thermal light of color temperature T emits visible light with the color of a blackbody (or graybody) radiator in thermal equilibrium at the thermodynamic temperature T. Color temperature is readily visualized along the Planckian locus on the chromaticity diagram (Fig. 9.7-2).

Thermodynamic Temperature. As described in Sec. 4.1, the kinetic theory of gases, in conjunction with Newton's laws of motion and the ideal gas law, allow a relation to be forged between thermodynamic temperature and the average internal energy of the system in which it is measured. Thermodynamic temperature is the most fundamental, and the common denominator, of all temperature measures.

Thermographic Temperature. As discussed in Sec. 4.8, thermography relies on an infrared photodetector array to register the power per unit area emitted by the elements of a thermal image. The thermographic temperature of these elements usually lies somewhere in the range 10 K $\leqslant T \leqslant$ 4000 K, roughly corresponding to emission in the

wavelength range $300\ \mu\text{m} \geqslant \lambda_0 \geqslant 0.7\ \mu\text{m}$. The principal distinctions between color temperature and thermographic temperature are as follows:

- The notion of color temperature, which involves the human visual system, is restricted to the visual region of the spectrum whereas thermographic temperatures are typically centered in the infrared.
- The visual system establishes the color temperature of an object by resolving the spectrum of the incoming thermal light while ignoring its luminance; the infrared photodetector array determines the thermographic temperature of an object by resolving the irradiance of the arriving thermal radiation while ignoring its spectrum.
- The determination of the color temperature of a source relies principally on its spectral radiance (9.7-1) and on the spectral response characteristics of the human visual system. The color temperature is represented on the chromaticity diagram. The accuracy with which the color temperature of a thermal source can be determined depends on how closely its spectrum adheres to the ideal blackbody or graybody form specified in (9.7-1).
- The determination of the thermographic temperatures of the pixels in a specimen relies on the radiated infrared power per unit area (4.8-1) and on the power-resolving capabilities of the pixels in the array detector. The calculations rely principally on the Stefan–Boltzmann law. The accuracy with which the effective thermographic temperature can be determined depends on the degree to which the local emissivity (4.8-2) is known.
- Wien's law (4.7-14), which provides the peak emission wavelength as a function of temperature, provides qualitative guidance for determining both color temperature and thermographic temperature.
- The false-color palette used to display thermodynamic temperature in thermography is chosen arbitrarily and bears no relation to color temperature. The coldest portions of the image are usually, but not always, portrayed as black or violet and the warmest portions as red or white. This is the mapping used in Fig. 4.8-2(a),(b), for example, but the opposite color convention is used in Fig. 4.8-2(c).

Biological Temperature. Over a limited range of thermodynamic temperatures (15–45 °C = 288–318 K), ambient skin temperature T is gauged by thermoreceptors that are sensitive to thermal energy. The percepts of increased and reduced temperature at the skin, relative to its nominal value, are engendered by two distinct sensory modalities mediated by thermoreceptors with different properties. Increased and reduced skin temperature are assigned the semantic labels "warm" and "cold," respectively.

In contradistinction to photoreceptors, which are all localized in the eye, both forms of somatosensory receptors are distributed in a punctate configuration over the entire surface of the skin. Discrete skin zones, each ≈ 1 mm in diameter, contain one or the other type of thermoreceptor that registers an increase or a decrease of temperature relative to normal skin temperature ($T_s \approx 34$ °C). Both warmth and cold thermoreceptors comprise the bare nerve endings of dorsal-root-ganglion primary afferent fibers endowed with temperature-sensitive ion channels that give rise to nerve-fiber action potentials. Different versions of these ion channels, which are responsive to various ranges of temperature, result in action potentials that transmit the temperature information to the central nervous system via the anterolateral system.

Cutaneous warmth receptors are activated at temperatures above normal skin temperature, in the range $34 \leqslant T \leqslant 45$ °C. The afferent nerve-fiber discharge rate λ_w in this temperature region is expressible as $\lambda_w \approx \lambda_s + k_w(T - T_s)$, where λ_s is the receptor spontaneous discharge rate in the absence of a thermal stimulus and k_w is a constant. The psychophysical magnitude estimate of warmth tracks the afferent discharge rate and behaves as $W \propto (T - T_s)$. Cutaneous cold receptors, whose behavior mirrors that of

warmth receptors, are activated at temperatures below normal skin temperature, in the range $34 \geqslant T \geqslant 15$ °C. The afferent nerve-fiber discharge rate λ_c in this temperature region is expressible as $\lambda_c \approx \lambda_s + k_c(T_s - T)$, where k_c is a constant. Again, the psychophysical magnitude estimation of cold follows the afferent discharge rate, and behaves as $C \propto (T_s - T)$.

Warm White Light and Cool White Light. The adjectives warm and cool are undefined for concepts that do not rely on thermodynamic temperature. In social discourse, for example, a "warm reception" indicates an amicable greeting, whereas a "cool reception" signifies a less-than-hospitable greeting. In contemporary conversation, in contrast, a "heated discussion" is an unwelcome event whereas a "cool encounter" is a welcome event. These countervailing examples illustrate the arbitrariness of the terms warm and cool in settings unrelated to thermodynamic temperature.

A more salient example, in the context of illumination, relates to the use of the adjectives warm and cool for characterizing the white light emanating from a thermal light source. The light emitted by a heated tungsten filament at a (relatively cool) thermodynamic temperature of 2700 K, which is reddish-yellow in color, is termed *warm white light*, while that emitted by a cloudy sky at a (relatively warm) thermodynamic temperature of 6500 K, which is bluish-white in color, is referred to as *cool white light*. The use of these assignations appears to date from the late eighteenth century. It is worthy of mention, perhaps, that the former is conducive to relaxation and inclination toward sleep (especially when it is also dim), while the latter is said to promote alertness and high performance (especially when it is also bright).

WARM (COOL) white light is emitted by a thermal source at a COOL (WARM) thermodynamic temperature. However, this apparent contradiction has no significance because the use of the terms warm and cool in connection with the character of light is a semantic choice that bears no relation to the use of these same terms in connection with thermodynamic temperature.

9.8 CORRELATED COLOR TEMPERATURE

In the early 2000s, an accumulation of advances in LED technology revealed that the advantages of LED lighting were incontrovertible and incandescent lighting finally yielded its preeminent position. LED lighting has the capability of generating light of any color, including the full range of whites at all color temperatures, and it does so with high efficiency. One of the principal topics considered in Chapter 10 is the design of metameric-white LEDs with output characteristics that mimic those of thermal light.

The question arises as to whether the notion of color temperature, which is useful for describing the color of thermal light, has an analog for nonthermal light. The answer turns out to be in the affirmative — but only for nonthermal emissions whose colors closely resemble those of thermal sources, i.e., for ources whose chromaticity coordinates lie sufficiently close to the Planckian locus on the chromaticity diagram (Fig. 9.7-2). This measure is called the **correlated color temperature (CCT)** and is denoted T_c (K). The CCT is the color temperature of the thermal source whose perceived color most closely matches that of the nonthermal source under consideration, under the same illumination and viewing conditions. This one-dimensional metric is often used as a proxy for the color of a nonthermal source of light because of its simplicity and convenience. Only light sources that are approximately white, such as metameric-white LEDs, metal halide lamps, and fluorescent lamps, are properly characterized by a CCT.

Color temperature and correlated color temperature provide simple characterizations of the color of the light emitted by thermal and nonthermal radiators, respectively. If a source of nonthermal light has chromaticity coordinates that lie near the Planckian locus, its correlated color temperature is taken to be the color temperature of the closest point on the Planckian locus.

CIE 1960 UCS

The CIE 1960 UCS (uniform color space) was expressly introduced to enable the CCT to be quantified. This color space has been largely superseded by CIE 1976 CIELUV, which in general has superior uniformity (Secs. 9.5 and 9.6). However, CIE 1960 UCS remains the preferred color space for calculating the CCT because it is more uniform for nominally white chromaticities. A simple matrix encodes the transformation from CIE 1931 XYZ to CIE 1960 UCS.

The associated uv chromaticity diagram, a portion of which is portrayed in Fig. 9.8-1, resembles the CIE 1976 $u'v'$ diagram shown in Fig. 9.6-3 more closely than it does the CIE 1931 xy version presented in Fig. 9.7-2, but it is distinct from both. Also called the **MacAdam chromaticity diagram**, it exhibits CCT isotherms that perpendicularly transect its Planckian locus over the range $1000 \leqslant T_c \leqslant 10000$ K. Following the steps used to calculate the color temperature for a thermal source (Sec. 9.7), the CCT of a source of light is established by using its power spectral density to determine the tristimulus values in the CIE 1960 UCS space, converting these to uv chromaticity coordinates, and then identifying the closest isotherm. While the minimum-distance calculation to determine the CCT is carried out within CIE 1960 UCS, CCTs can be displayed on any chromaticity diagram, such as CIE 1931 XYZ (Fig. 9.7-2). Representative values of the CCT for various sources used in LED lighting are displayed in Table 11.9-1.

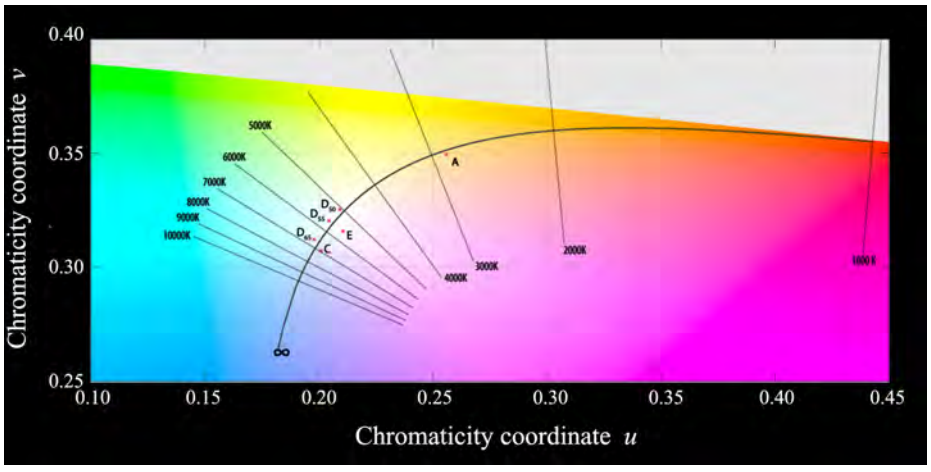


Figure 9.8-1 Expanded view of the uv chromaticity diagram for the CIE 1960 UCS (uniform color space), also called the MacAdam chromaticity diagram, in the vicinity of the Planckian locus (curved line). The straight lines that perpendicularly transect the Planckian locus, labeled in kelvins (K), represent correlated color temperature (CCT) isotherms. The (u, v) chromaticity coordinates for a number of CIE standard illuminants are indicated: A (incandescent light with $T_c \approx 2856$ K), E (equal-energy with $T_c \approx 5460$ K), and several variations on daylight (D_{50}, D_{55}, and D_{65}). The CCT for an arbitrary light source is determined by finding the isotherm closest to its chromaticity coordinates via interpolation.

Properties. The portion of the uv chromaticity diagram in the vicinity of the Planckian locus, displayed in Fig. 9.8-1, has the following properties:

- The loci of constant correlated color temperature T_c(K) are represented by straight lines that perpendicularly transect the Planckian locus. However, these loci are not perpendicular to the Planckian locus in the xy chromaticity diagram (Fig. 9.7-2).
- The chromaticity coordinates for several standard illuminants codified by the CIE are displayed: Illuminant series A, D, and E represent average incandescent light ($T_c \approx 2856$ K), variations on daylight, and equal-energy illumination ($T_c \approx 5460$ K), respectively. In 2018. the CIE introduced standard illuminants representing different types of LEDs with $2700 \leqslant T_c \leqslant 6600$.
- The illuminant D_{65}, which roughly corresponds to average midday light with $T_c = 6500$ K (the origin of the subscript 65), is represented by the chromaticity coordinates $(u, v) = (0.198, 0.312)$.
- The (u, v) coordinates (Fig. 9.8-1) differ from the (x, y) coordinates (Fig. 9.7-2) because their color matching functions are different. For the illuminant D_{65}, for example, $(u, v) = (0.198, 0.312)$ while $(x, y) = (0.31271, 0.32902)$.

Following the convention used with thermal light, nonthermal sources with color temperatures 2700 $\lesssim T_c \lesssim$ 3500 K (yellowish) are usually considered to be* warm white *or* soft white*, those with 3500 $\lesssim T_c \lesssim$ 5000 K (yellowish-white) are said to be* neutral white *or* bright white*, and those with 5000 $\lesssim T_c \lesssim$ 7500 K (bluish-white) are considered to be* cool white *or* daylight*, although there is considerable latitude in the way these ranges are defined.

EXAMPLE 9.8-1. ***Mapping (x,y) and (u′,v′) to (u,v) Chromaticity Coordinates.*** The mapping of (x, y) to (u, v) chromaticity coordinates takes the following form:

$$u = \frac{4x}{-2x + 12y + 3}, \qquad v = \frac{6y}{-2x + 12y + 3}. \tag{9.8-1}$$

A comparison with the mapping of (x, y) to (u', v') chromaticity coordinates provided in Example 9.6-4 leads to

$$u = u', \qquad v = 2v'/3. \tag{9.8-2}$$

EXAMPLE 9.8-2. ***Determination of the CCT from the xy Chromaticity Coordinates.*** A concise polynomial expression is available for estimating the correlated color temperature T_c of a source from its CIE 1931 xyY chromaticity coordinates (x, y). The relationship results from the fact that the CCT isotherms converge toward the bottom of the xy chromaticity diagram (Fig. 9.7-2). The most commonly used version of this approximation, which is suitable for daylight and incandescent sources with CCTs in the range $2000 \lesssim T_c \lesssim 12500$ K, takes the form[†]

$$T_c \approx -449\,\zeta^3 + 3525\,\zeta^2 - 6823.3\,\zeta + 5520.33 \tag{9.8-3a}$$

where

$$\zeta = (x - 0.3320) / (y - 0.1858). \tag{9.8-3b}$$

The typical error associated with this procedure, $\leqslant 2$ K, is remarkably small. We provide two numerical examples:

(a) *Equal-energy (spectrally uniform) white light*: The chromaticity coordinates for the standard CIE colorimetric illuminant E are $(x, y) = (1/3, 1/3)$, as established in Example 9.6-3. Inserting these coordinates into (9.8-3b) yields $\zeta = 0.0088136$, which, when used in (9.8-3a)

[†] C. S. McCamy, Correlated Color Temperature as an Explicit Function of Chromaticity Coordinates, *Color Research and Application*, vol. 17, pp. 142–144, 1992; Erratum: vol. 18, p. 150, 1993.

returns $T_c \approx 5460$ K. An approximation to equal-energy white light can be generated by a phosphor-conversion LED that makes use of a specially blended collection of phosphors (Example 10.4-3).

(b) *The standard CIE colorimetric illuminant* D_{65}: As indicated earlier, the chromaticity coordinates for this illuminant are $(x, y) = (0.3127, 0.3290)$. Inserting these coordinates into (9.8-3b) yields $\zeta = -0.13469$, which, when used in (9.8-3a) returns $T_c \approx 6504$ K.

The McCamy method can be used to estimate the correlated color temperature for a white phosphor-conversion LED (Example 10.5-4). Formulas that are applicable over broader ranges of the CCT are also available, but their use entails increased complexity.

9.9 COLOR RENDERING INDEX

The **color rendering index (CRI)** is a widely used measure designed to represent how well a light source illuminating an object renders its color. The CRI is to be contrasted with the correlated color temperature (CCT), which characterizes the perceived color of the light source itself (Sec. 9.8). The CRI, like the CCT, is used principally for light sources that are approximately white, and both are often used together for assessing the quality of color rendering.

It was the advent of fluorescent lamps that led the CIE to consider color rendering and to recommend the CRI as a metric in 1965. The CRI is determined by comparing the colors of the light emitted by the test source with those emitted from a blackbody source of the same CCT, upon reflection from a set of Munsell color samples. Ideally, the luminous flux of the two sources is the same when carrying out the comparison. The eight *standard Munsell samples*, which are considered to be representative of everyday colors such as foliage and sky blue, are labeled R1–R8. Subsequently added to the mix are seven *special Munsell samples* that include saturated red, green, and blue and are labeled R9–R15; these are more challenging for most light sources to render accurately. The 15 standard and special Munsell samples attendant to the determination of the CRI are presented in Fig. 9.9-1.

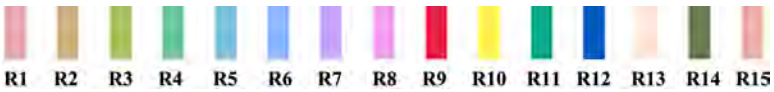


Figure 9.9-1 Munsell color samples R1–R15 involved in determining the CRI.

Specifically, the CRI is computed by shifting and averaging the Euclidian distances between the test and blackbody colors reflected from the samples, using the 1964 CIEUVW color space, which continues to be used for this purpose but is otherwise obsolete. A source that allows objects to be seen as they would appear under illumination by daylight, or by an incandescent source of the same CCT, is assigned the maximum value: CRI = 100.

Representative values of the CRI for various sources used in LED lighting are provided in Table 11.9-1. Roughly speaking, a value CRI < 75 signifies that the colors of the illuminated objects appear unnatural and the use of such sources is not recommended for indoor lighting. By virtue of its definition, the CRI has also been used as a measure for metamerism (Sec. 9.3).

Despite its widespread use, a substantial body of evidence reveals that the CRI often misassesses human judgments of naturalness and overall color preference, particularly

for LED lighting. As a result, a number of other color-rendering measures have been suggested as alternates over the years. Nevertheless, no single metric developed so far has proved capable of adequately capturing the multidimensional features of color rendering, although efforts continue to develop such a metric. In the meantime, despite its deficiencies, the CRI continues to be widely used for assessing the suitability of light sources for particular color applications, as do any number of *ad hoc* combinations of nonstandard measures.

EXAMPLE 9.9-1. ***LER, CCT, and CRI for Uniform-Spectral-Density White Light.*** As discussed in Example 9.6-3, equal-energy white light has a uniform wavelength-based power spectral density $S_\lambda(\lambda_0)$. It is represented by the white point on the chromaticity diagram, $(x, y) = (1/3, 1/3)$, and appears white to the eye. Making use of (8.8-1) and (8.9-1), the LER defined in Sec. 8.9 can be written as

$$\eta_{\text{LER}} = \frac{P_V}{P_0} = 683 \int_{\lambda_{\text{MIN}}}^{\lambda_{\text{MAX}}} \left[\frac{S_\lambda(\lambda_0)}{P_0}\right] V(\lambda_0)\, d\lambda_0\,, \tag{9.9-1}$$

where the integration extends from the lower cutoff of the spectral density λ_{MIN} to its upper cutoff λ_{MAX}, and the optical power P_0 represents the integrated spectral density.

In practice, the integration is usually approximated by employing the **sampled photopic luminous efficiency function** $\{V(\lambda_i)\}$. This is a set of real numbers that represents the sampled values of $V(\lambda_0)$ at discrete, consecutive, free-space wavelengths λ_i, spaced 1 nm apart, that range from 380 to 780 nm. These data are available on the web (as detailed in footnote a of Table 9.9-1); the values are readily imported into a spreadsheet such as EXCEL. Equation (9.9-1) can be written in discretized form as

$$\eta_{\text{LER}} \approx \frac{683}{\sum_{\lambda_{\text{MIN}}}^{\lambda_{\text{MAX}}} S(\lambda_i)} \sum_{\lambda_i=\lambda_{\text{MIN}}}^{\lambda_{\text{MAX}}} V(\lambda_i) = \frac{683}{(\lambda_{\text{MAX}} - \lambda_{\text{MIN}} + 1)} \sum_{\lambda_i=\lambda_{\text{MIN}}}^{\lambda_{\text{MAX}}} V(\lambda_i)\,, \tag{9.9-2}$$

where the summations extend over the selected values of λ_{MIN} and λ_{MAX}. The denominator in the prefactor, $(\lambda_{\text{MAX}} - \lambda_{\text{MIN}} + 1)$, represents the number of data samples and provides the normalization that insures that the optical power is the same, whatever the values chosen for λ_{MIN} and λ_{MAX}. Representative values for the LER are posted in Table 9.9-1 for selected bandwidths of the uniform spectral density. In the particular case when $\lambda_{\text{MIN}} = 380$ nm and $\lambda_{\text{MAX}} = 780$ nm, (9.9-2) sums to $\eta_{\text{LER}} \approx 182$ lm/W.

Examining the entries in Table 9.9-1, it is apparent that as the uniform-spectral-density bandwidth $\lambda_{\text{MAX}} - \lambda_{\text{MIN}}$ decreases, the LER increases monotonically. This behavior emerges because narrower bandwidths are localized toward the center of the photopic luminous efficiency function, where its values are largest. Indeed, in the limit when the spectral band is narrowed to straddle 555 nm and allow only 1 nm width to either side, the LER approaches 683 lm/W (bottom row of Table 9.9-1); this is its maximum possible value, as specified in (8.9-3). LER values for truncated Planckian spectral densities behave similarly, returning CCT and CRI values close to those displayed in rows 2–5 of Table 9.9-1 for matching bandwidths, a conclusion that is applicable for color temperatures ranging from 2500 to 8000 K.

Numerical studies also confirm that the color rendering index (CRI) becomes unacceptably small ($\leqslant 70$) when the uniform-spectral-density source is truncated such that $\lambda_{\text{MIN}} > 453$ lm and $\lambda_{\text{MAX}} < 663$ nm, corresponding to $\lambda_{\text{MAX}} - \lambda_{\text{MIN}} < 210$ nm (Table 9.9-1). The elimination of segments of the short- and long-wavelength spectral components carries over to the light reflected from an illuminated object, which is then deficient in blue and red and thus exhibits impaired color rendering. In the limit when the spectral band straddles 555 nm with a mere 1 nm of spectral width on either side, the light is essentially monochromatic and yellowish-green in color. Since it is then far from white, a CCT is not defined and the CRI $= 0$.

EXAMPLE 9.9-2. ***LER and CRI for Spectrally Matched Light.*** As a final example in this section, consider a source of light whose spectral density $S_\lambda(\lambda_0)$ is chosen to precisely match the photopic luminous efficiency function $V(\lambda_0)$. This stimulus choice is expected to maximize the luminous efficacy of radiation since the light entering the eye then has a wavelength distribution that corresponds identically to the superposed wavelength spectral sensitivities of the S-, M-, and L-cones in the retina (Sec. 8.5).

Table 9.9-1 Representative values of the luminous efficacy of radiation (LER) for selected values of the uniform spectral-density bandwidth. Successive columns in the table represent, respectively: the lower spectral-density cutoff λ_{MIN} (nm), the upper spectral-density cutoff λ_{MAX} (nm), the uniform-spectral-density bandwidth $(\lambda_{\text{MAX}} - \lambda_{\text{MIN}})$ (nm), the calculated LER (lm/W), the calculated correlated color temperature (CCT) (K), and the calculated color-rendering index (CRI).

λ_{MIN}	λ_{MAX}	$\lambda_{\text{MAX}} - \lambda_{\text{MIN}}$	LER[a]	CCT[b]	CRI[b,c]
380	780	400	182	5460	
406	697	291	250	5440	96
413	687	274	265	5415	97
422	677	255	284	5324	98
453	663	210	341	4418	72
475	650	175	399	–	↓
505	605	100	559	–	|
530	580	50	652	–	↓
554	556	2	683	–	0

[a]The LER was calculated using the CIE 1924 2° data for $V(\lambda_i)$ in conjunction with (9.9-2). These data have been tabulated by G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, Wiley, 2nd ed. 1982, Table I(3.3.1), and can be downloaded from `http://www.cvrl.org/database/text/lum/vl.htm`. Though the 1924 photometric data is used extensively, it is widely understood that it underestimate visual-system photopic sensitivity at short wavelengths. As discussed in Sec. 8.5, an updated version for daylight adaptation, denoted $V^*(\lambda_0)$, was set forth in 2005 (see footnote on p. 246).

[b]The CCT and CRI are generally used for light that is approximately white. The CCT and CRI data displayed in rows 2–5 are drawn from T. W. Murphy, Maximum Spectral Luminous Efficacy of White Light, *Journal of Applied Physics*, vol. 111, 104909, 2012.

[c]The slight local increase in the CRI as the largest bandwidths are narrowed results from the trimming of the blue and red spectral components in the uniform spectral density that exceed those in the Planckian density.

The calculation is carried out by forging a discretized version of (9.9-1), and making use of the sampled spectrum $S(\lambda_i)$ introduced in Example 9.5-1 and the sampled photopic luminous efficiency function $V(\lambda_i)$ employed in Example 9.9-1. Spectrally matched light means that $S(\lambda_i) \equiv V(\lambda_i)$, so that (9.9-1) can be approximated as

$$\eta_{\text{LER}} \approx \frac{683}{\sum_{\lambda_{\text{MIN}}}^{\lambda_{\text{MAX}}} S(\lambda_i)} \sum_{\lambda_i=\lambda_{\text{MIN}}}^{\lambda_{\text{MAX}}} S(\lambda_i)\, V(\lambda_i) = \frac{683}{\sum_{\lambda_{\text{MIN}}}^{\lambda_{\text{MAX}}} V(\lambda_i)} \sum_{\lambda_i=\lambda_{\text{MIN}}}^{\lambda_{\text{MAX}}} V^2(\lambda_i)\,, \tag{9.9-3}$$

where the summation extends over the desired range of $\lambda_{\text{MAX}} - \lambda_{\text{MIN}}$. Representative values for the spectrally matched LER are reported in Table 9.9-2 for bandwidth values that are the same as those displayed in Table 9.9-1 for the uniform spectral density, to facilitate comparison.

As is apparent in Table 9.9-2, we have $\eta_{\text{LER}} \approx 493$ lm/W, as determined via (9.9-3), which corresponds to the largest bandwidth $\lambda_{\text{MAX}} - \lambda_{\text{MIN}} = 400$ nm. This value of the luminous efficacy of radiation is substantially larger than that for the uniform spectral density, $\eta_{\text{LER}} \approx 182$ lm/W, which was determined using (9.9-2) and is displayed in Table 9.9-1. Spectral matching of the incident light to the spectral sensitivity of the retinal cones provides an evident advantage in the LER. Further examination of the entries in Table 9.9-2 reveals that the LER remains essentially constant at a value ≈ 493 lm/W as the bandwidth decreases over a substantial range. However, as the selected bandwidth $(\lambda_{\text{MAX}} - \lambda_{\text{MIN}})$ falls below ≈ 210 nm, the LER exhibits a sharper rate of increase. Ultimately, as with uniform-spectral-density light, a very narrow bandwidth straddling 555 nm yields an LER that approaches the maximum allowed value of 683 lm/W.

The color rendering index tells another story, however. Even for the broadest bandwidths, there is a serious disadvantage in using spectrally matched light for illumination. In short, the light is not white, but rather appears yellowish-green to the eye. The spectrum of the light, it turns out, is similar

Table 9.9-2 Representative values of the luminous efficacy of radiation (LER) for a source whose spectral density $S(\lambda_i)$ matches the photopic luminous efficiency function $V(\lambda_i)$. Successive columns in the table represent: lower spectral-density cutoff λ_{MIN} (nm), upper spectral-density cutoff λ_{MAX} (nm), resulting bandwidth $(\lambda_{\text{MAX}} - \lambda_{\text{MIN}})$ (nm), calculated LER (lm/W), and estimated color-rendering index (CRI).

λ_{MIN}	λ_{MAX}	$\lambda_{\text{MAX}} - \lambda_{\text{MIN}}$	LER[a]	CRI[b]
380	780	400	493	25
406	697	291	494	↓
413	687	274	494	|
422	677	255	495	|
453	663	210	500	↓
475	650	175	511	|
505	605	100	581	|
530	580	50	653	↓
554	556	2	683	0

[a] The LER was calculated using the CIE 1924 2° data for $V(\lambda_i)$ in conjunction with (9.9-3). The data were downloaded from `http://www.cvrl.org/database/text/lum/vl.htm`.
[b] Estimated values. The CRI is generally used for light that is approximately white.

to that emitted by a yellow-green LED [Fig. 7.2-3(a)], although somewhat broader. It is therefore not surprising that the CRI is unacceptably low, clocking in at 25 even when the selected bandwidth $\lambda_{\text{MAX}} - \lambda_{\text{MIN}}$ covers the entire 400 nm bandwidth of the visible spectrum (Table 9.9-2). Once again, the diminution of blue and red spectral components in the source carries over to the light reflected from an illuminated object, but in this case the deficiencies are far more severe than those for uniform-spectrum white light and the CRI is correspondingly reduced. In the limit when the spectral band straddles 555 nm with only a 1 nm shoulder on either side, the spectrally matched light is indistinguishable from the uniform-spectrum light of the same bandwidth, as expected. In both cases, the light is then yellowish-green in color and essentially monochromatic, so that a CCT is not defined and the CRI $= 0$.

BIBLIOGRAPHY

Colorimetry

R. Shamey, ed., *Encyclopedia of Color Science and Technology*, Springer, 2023.

A. D. Logvinenko and V. L. Levin, *Foundations of Colour Science: From Colorimetry to Perception*, Wiley, 2023.

CIE (Commission Internationale de l'Éclairage), *International Lighting Vocabulary*, International Commission on Illumination CIE Standard S 017/E:2020, Vienna, Austria, DOI:10.25039/S017.2020, `https://cie.co.at/e-ilv`, 2nd ed. 2020.

R. S. Berns, *Billmeyer and Saltzman's Principles of Color Technology*, Wiley, 4th ed. 2019.

CIE (Commission Internationale de l'Éclairage), Application of CIE 2015 Cone-Fundamental-Based CIE Colorimetry, *LED Professional Review*, no. 60, ISSN:1993-0890X, March/April 2017.

F. Viénot, D. MacLeod, J. D. Mollon, J. D. Moreland, J. Pokorny, L. T. Sharpe, A. Stockman, A. Valberg, J. J. Vos, P. L. Walraven, J. H. Wold, and H. Yaguchi (CIE Technical Committee 1-36), Fundamental Chromaticity Diagram with Physiological Axes – Part 2: Spectral Luminous Efficiency Functions and Chromaticity Diagrams, Technical Report No. 170-2, Commission Internationale de l'Éclairage (International Commission on Illumination), Vienna, Austria, ISBN:9783902842053, 2015.

R. W. G. Hunt and M. R. Pointer, *Measuring Colour*, Wiley, 4th ed. 2011.

D. Malacara, *Color Vision and Colorimetry: Theory and Applications*, SPIE Press, 2nd ed. 2011.

J. Koenderink, *Color for the Sciences*, MIT Press, 2010.

D. H. Brainard and A. Stockman, Colorimetry, in M. Bass, C. DeCusatis, J. Enoch, V. Lakshminarayanan, G. Li, C. Macdonald, V. Mahajan, and E. van Stryland, eds., *Handbook of Optics: Volume III – Vision and Vision Optics*, 3rd ed., McGraw–Hill, Chap. 10, 2010.

M. H. Brill, M. D. Fairchild, H. S. Fairman, K. Houser, R. G. Kuehni, R. Luo, B. Oicherman, C. Oleari, D. Oulton, D. Rich, A. Robertson, J. Schanda, A. W. Tarrant, W. A. Thornton, K. Wenzel, J. Wold, A. Stockman, and F. Vienot (CIE Technical Committee 1-56), Reappraisal of Colour Matching and Grassmann's Laws, Technical Report No. 185, Commission Internationale de l'Éclairage (International Commission on Illumination), Vienna, Austria, ISBN:9783901906787, 2009.

J. Schanda, ed., *Colorimetry: Understanding the CIE System*, Wiley, 2007.

D. MacLeod, J. D. Mollon, J. D. Moreland, Y. Nakano, J. Pokorny, L. T. Sharpe, A. Stockman, A. Valberg, F. Viénot, J. J. Vos, P. L. Walraven, J. H. Wold, H. Scheibner, P. Trezona, and H. Yaguchi (CIE Technical Committee 1-36), Fundamental Chromaticity Diagram with Physiologically Significant Axes – Part 1: Definition of CIE 2006 Cone Fundamentals, Technical Report No. 170, Commission Internationale de l'Éclairage (International Commission on Illumination), Vienna, Austria, ISBN:9783901906466, 2006.

N. Ohta and A. Robertson, *Colorimetry: Fundamentals and Applications*, Wiley, 2005.

A. Stockman, Colorimetry, in T. G. Brown, K. Creath, H. Kogelnik, M. A. Kriss, J. Schmit, and M. J. Weber, eds., *The Optics Encyclopedia: Basic Foundations and Practical Applications*, Wiley–VCH, pp. 207–226, 2004.

S. K. Shevell, ed., *The Science of Color*, Optical Society of America/Elsevier, 2nd ed. 2003.

D. L. MacAdam, *Color Measurement: Theme and Variations*, Springer, 2nd ed. 1985.

S. J. Williamson and H. Z. Cummins, *Light and Color in Nature and Art*, Wiley, 1983.

G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, Wiley, 2nd ed. 1982.

W. D. Wright, *The Measurement of Colour*, Hilger, 1st ed. 1944, 4th ed. 1969.

W. S. Stiles and J. M. Burch, N.P.L. Colour-Matching Investigation: Final Report (1958), *Optica Acta: International Journal of Optics*, vol. 6, pp. 1–26, DOI:10.1080/713826267, 1959.

D. L. MacAdam, Visual Sensitivities to Color Differences in Daylight, *Journal of the Optical Society of America*, vol. 32, pp. 247–274, 1942.

Color Appearance Models

M. R. Luo, Q. Xu, M. Pointer, M. Melgosa, G. Cui, C. Li, K. Xiao, and M. Huang, A Comprehensive Test of Colour-Difference Formulae and Uniform Colour Spaces Using Available Visual Datasets, *Color Research and Application*, vol. 67, p. 020405, DOI:10.1002/col.22844, 2023.

C. Gao, M. R. Luo, M. R. Pointer, and C. Li, Evaluation of Color Difference Prediction with CIECAM16 using CIE 2- and 10-degree Observers, *Journal of Imaging Science and Technology*, DOI:10.2352/J.ImagingSci.Technol.2023.67.2.020405, p. 020405, 2023.

M. Bertalmío, *Vision Models for High Dynamic Range and Wide Colour Gamut Imaging: Techniques and Applications*, Academic/Elsevier, 2020.

A. K. R. Choudhury, *Principles of Colour Appearance and Measurement. Volume 1: Object Appearance, Colour Perception and Instrumental Measurement*, Woodhead/Elsevier, 2014.

M. D. Fairchild, *Color Appearance Models*, Wiley, 3rd ed. 2013.

G. Wyszecki, Color Appearance, in *Handbook of Perception and Human Performance*, Vol. 1, *Sensory Processes and Perception*, Wiley, Chap. 9, pp. 9-1–9-57 (pp. 447–504), 1986.

J. Pokorny and V. C. Smith, Colorimetry and Color Discrimination, in *Handbook of Perception and Human Performance*, Vol. 1, *Sensory Processes and Perception*, Wiley, Chap. 8, pp. 8-1–8-51 (pp. 395–446), 1986.

G. Wyszecki, Current Developments in Colorimetry, in *COLOUR73: Survey Lectures and Abstracts of the Papers Presented at the Second Congress of the International Colour Association Held at the University of York 2–6 July 1973*, Hilger (London), pp. 21–51, 1973.

E. H. Land and J. J. McCann, Lightness and Retinex Theory, *Journal of the Optical Society of America*, vol. 61, pp. 1–11, 1971.

J. von Kries, Theoretische Studien ueber die Umstimmung des Sehorgans (Theoretical Studies on the Retuning of the Visual Organ), in *Festschrift der Albrecht-Ludwigs-Universität in Freiburg zum fünfzigjärigen Regierungs-Jubiläum Seiner Königlichen Hoheit des Grossherzogs Friedrich*, C. A. Wagner's Universitäts-Buchdruckerei, pp. 143–158, 1902 [Translation: Chromatic Adaptation, in D. L. MacAdam, ed., *Sources of Color Science: Selected and Edited by David L. MacAdam*, pp. 109–119, MIT Press, 1970].

Thermodynamic, Thermographic, and Biological Temperature

See also the bibliographies on thermal and statistical physics, blackbody radiation, and infrared detectors and thermography in Chapter 4.

D. Julius, TRP Channels and Pain, *Annual Review of Cell and Developmental Biology*, vol. 29, pp. 355–384, 2013.

D. D. McKemy, W. M. Neuhausser, and D. Julius, Identification of a Cold Receptor Reveals a General Role for TRP Channels in Thermosensation, *Nature*, vol. 416, pp. 52–58, 2002.

M. J. Caterina, M. A. Schumacher, M. Tominaga, T. A. Rosen, J. D. Levine, and D. Julius, The Capsaicin Receptor: A Heat-Activated Ion Channel in the Pain Pathway, *Nature*, vol. 389, pp. 816–824, 1997.

Color Temperature and Correlated Color Temperature

See also the bibliographies on thermal and statistical physics, blackbody radiation, and infrared detectors and thermography in Chapter 4.

C. Li, G. Cui, M. Melgosa, X. Ruan, Y. Zhang, L. Ma, K. Xiao, and M. Ronnier Luo, Accurate Method for Computing Correlated Color Temperature, *Optics Express*, vol. 24, pp. 14066–14078, 2016.

J. Hernández-Andrés, R. L. Lee, Jr., and J. Romero, Calculating Correlated Color Temperatures Across the Entire Gamut of Daylight and Skylight Chromaticities, *Applied Optics*, vol. 38, pp. 5703–5709, 1999.

C. S. McCamy, Correlated Color Temperature as an Explicit Function of Chromaticity Coordinates, *Color Research and Application*, vol. 17, pp. 142–144, 1992; Erratum: vol. 18, p. 150, 1993.

D. B. Judd, Estimation of Chromaticity Differences and Nearest Color Temperature on the Standard 1931 ICI Colorimetric Coordinate System, *Journal of the Optical Society of America*, vol. 26, pp. 421–426, 1936.

R. Davis, A Correlated Color Temperature for Illuminants, *National Bureau of Standards Journal of Research*, vol. 7, pp. 659–681, 1931.

E. P. Hyde, A New Determination of the Selective Radiation from Tantalum (Abstract), *Physical Review*, vol. 32, p. 632, 1911.

Color Rendering

A. Žukauskas and M. S. Shur, Color Rendering Metrics: Status, Methods, and Future Development, in R. Karlicek, C.-C. Sun, G. Zissis, and R. Ma, eds., *Handbook of Advanced Lighting Technology*, Springer Nature, pp. 799–827, 2017.

W. Davis, History of Color Metrics, in R. Karlicek, C.-C. Sun, G. Zissis, and R. Ma, eds., *Handbook of Advanced Lighting Technology*, Springer Nature, pp. 785–797, 2017.

J. Schanda, P. Csuti, F. Szabó, P. Bhusal, and L. Halonen, Introduction to a Study of Preferred Colour Rendering of Light Sources, *Lighting Research and Technology*, vol. 47, pp. 28–35, 2015.

A. David, Color Fidelity of Light Sources Evaluated over Large Sets of Reflectance Samples, *Leukos*, vol. 10, pp. 59–75, 2014.

R. Dangol, M. Islam, M. Hyvärinen, P. Bhusal, M. Puolakka, and L. Halonen, Subjective Preferences and Colour Quality Metrics of LED Light Sources, *Lighting Research and Technology*, vol. 45, pp. 666–688, 2013.

W. Davis and Y. Ohno, Color Quality Scale, *Optical Engineering*, vol. 49, p. 033602, 2010.

D. Nickerson and C. W. Jerome, Color Rendering of Light Sources: CIE Method of Specification and its Application, *Illumination Engineering*, vol. 60, pp. 262–271, 1965.

Historical Accounts and Early Publications

R. Shamey and R. G. Kuehni, *Pioneers of Color Science*, Springer Nature, 2020.

W. D. Wright, Golden Jubilee of Colour in the CIE – The Historical and Experimental Background

to the 1931 CIE System of Colorimetry, in J. Schanda, ed., *Colorimetry: Understanding the CIE System*, Wiley, pp. 9–24, 2007.

J. W. T. Walsh and A. M. Marsden, *History of the CIE: 1913–1988*, CIE Report No. 82-1990, Commission Internationale de l'Éclairage (International Commission on Illumination), Vienna, Austria, ISBN:9783900734190 (Photocopy Edition), 1999.

D. L. MacAdam, ed., *Selected Papers on Colorimetry – Fundamentals*, SPIE Optical Engineering Press (Milestone Series Volume 77), 1993.

D. L. MacAdam, Color Essays, *Journal of the Optical Society of America*, vol. 65, pp. 483–493, 1975.

D. L. MacAdam, ed., *Sources of Color Science: Selected and Edited by David L. MacAdam*, MIT Press, 1970.

CIE (Commission Internationale de l'Éclairage), *Recueil des travaux et compte rendu des séances: huitième session*, Cambridge (September 1931), Cambridge University Press, 1932.

J. Guild, The Colorimetric Properties of the Spectrum, *Philosophical Transactions of the Royal Society of London*, vol. A230, pp. 149–187, 1931.

T. Smith and J. Guild, The C.I.E. Colorimetric Standards and Their Use, *Transactions of the Optical Society (London)*, vol. 33, pp. 73–134, 1931.

W. D. Wright, A Re-Determination of the Trichromatic Coefficients of the Spectral Colours, *Transactions of the Optical Society (London)*, vol. 30, pp. 141–164, 1929.

H. Grassmann, Zur Theorie der Farbenmischung, *Annalen der Physik*, vol. 165(5), pp. 69–84, 1853 [Translation: On the Theory of Compound Colours, *London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, Ser. 4, vol. 7(45), pp. 254–264, 1854].

CHAPTER

10

PHOSPHOR-CONVERSION LEDS

10.1 MONOCHROMATIC AND WHITE LED LIGHT 308

10.2 PHOTOLUMINESCENCE 310

10.3 BROADBAND AND NARROWBAND PHOSPHORS 312

10.4 BLENDED PHOSPHORS 318

10.5 DISCRETE COOL-WHITE PCLEDS 322

10.6 DISCRETE WARM-WHITE PCLEDS 328

10.7 PCLED FILAMENTS 331

10.8 CHIP-ON-BOARD PCLEDS 332

Isamu Akasaki (1929–2021), left, **Hiroshi Amano (born 1960)**, center, and **Shuji Nakamura (born 1954)**, right, shared the Nobel Prize in Physics in 2014 for inventing the blue light-emitting diode that enabled the development of efficient white phosphor-conversion LEDs. These energy-saving devices can be powered by batteries that are in turn charged by solar panels that produce electricity during the day. This has made nighttime light available in remote areas that have limited access to electricity.

LED Lighting: Devices and Colorimetry. Malvin Carl Teich.
Google Books. Published 2024.

The development of white LED lighting has been a cardinal technological achievement since illumination applications most often make use of white light. The principal method for generating metameric white light relies on **phosphor-conversion (PC) LEDs**, which are both practical and economical, largely because of their simplicity. PCLEDs operate on the basis of photoluminescence and, in their simplest form, require only a single LED and a single phosphor. This chapter relies extensively on the developments provided in Chapter 7, which recounts the operation and characteristics of LEDs, and on Chapters 8 and 9, which are dedicated to the fundamental principles underlying human color vision and colorimetry, respectively. This chapter should be viewed as a precursor to Chapter 11, which is devoted to LED lighting.

The combination of royal blue and yellow, when blended in suitable proportions, appears white to the eye, as may be understood from Example 9.3-1. This superposition, which produces metameric white light as described in Sec. 9.3, can be conveniently created by directing a portion of the light generated by a blue LED to a phosphor that emits yellow light. In principle, other complementary-color pairs, such as red and cyan (Example 9.3-2) or green and magenta, could also be used to produce metameric white light. However, blue/yellow is a judicious choice for several reasons:

- Blue LEDs fabricated from InGaN are highly efficient.
- The energy of blue photons is sufficient to excite phosphors of all colors.
- The yellow-phosphor emission wavelength has high photopic luminous efficiency.

The implementation of white LED lighting therefore relied heavily on the development of the blue LED. While it was understood early on that InGaN would generate blue light if it could be cast in the form of a forward-biased p–n junction diode, implementing this solution proved unexpectedly difficult from a technical point-of-view. It demanded not only the ability to grow high-quality InGaN in the form of heterojunctions and quantum wells, but also the capability of converting n–type InGaN, the default, into p–type material. The fabrication of an efficient blue InGaN LED, which was ultimately perfected in the early 1990s, relied on numerous advances in crystal growth and materials science that were the product of the dogged persistence of Isamu Akasaki, Hiroshi Amano, and Shuji Nakamura, over a period of many years.[†] Their achievements led the way to the efficient white phosphor-conversion LED and earned them the Nobel Prize in Physics in 2014 "for the invention of efficient blue light-emitting diodes which has enabled bright and energy-saving white light sources" (p. 306).

We begin the chapter by reviewing the salient features of monochromatic and metameric-white LED light in the context of human trichromatic vision (Sec. 10.1). A description of photoluminescence and the measures used to quantify it follows (Sec. 10.2). The physics and characteristics of a number of commonly used broadband and narrowband phosphors are examined in Sec. 10.3, and the properties of several salient phosphor blends are detailed in Sec. 10.4. The classic example of the generation of metameric cool-white light using a discrete phosphor-conversion LED that relies on the illumination of a yellow phosphor by light from a blue LED is related in Sec. 10.5. The use of a judiciously chosen phosphor blend leads to the generation of metameric warm-white light instead, as described in Sec. 10.6. Finally, Sec. 10.7 is dedicated to describing the operation of phosphor-conversion LED filaments, such as those used in white retrofit lamps, while Sec. 10.8 is devoted to reviewing chip-on-board (COB) phosphor-conversion LEDs, which generate high luminous flux and are ubiquitous.

[†] K. Itoh, T. Kawamoto, H. Amano, K. Hiramatsu, and I. Akasaki, Metalorganic Vapor Phase Epitaxial Growth and Properties of GaN/$Al_{0.1}Ga_{0.9}N$ Layered Structures, *Japanese Journal of Applied Physics*, vol. 30, no. 9R, p. 1924, 1991; S. Nakamura, T. Mukai, and M. Senoh, Candela-Class High-Brightness InGaN/AlGaN Double-Heterostructure Blue-Light-Emitting Diodes, *Applied Physics Letters*, vol. 64, pp. 1687–1689, 1994.

10.1 MONOCHROMATIC AND WHITE LED LIGHT

The structures and operating principles of electroluminescent LEDs with active regions comprising different type of semiconductor materials were reviewed in Chapters 5–7. In particular, the features and properties of the most prominent types of LEDs were discussed in the following sections:

- Multiquantum-Well LEDs (LEDs or MQWLEDs): Secs. 5.7, 6.5, 7.3, and 7.4.
- Quantum-Dot & White Quantum-Dot LEDs (QLEDs & WQLEDs): Secs. 5.8, 6.6, and 7.5.
- MicroLEDs (μLEDs): Sec. 7.3.
- Organic & White Organic LEDs (SMOLEDs, PLEDs, & WOLEDs): Secs. 5.9, 7.6, and 11.7.
- Perovskite & White Perovskite LEDs (PeLEDs & PeWLEDs): Secs. 5.9 and 7.7.

Tables 7.4-1 and 7.6-1 reveal that multiquantum-well LEDs (which we refer to as either MQWLEDs or simply as LEDs) generate far more light than do the other categories of light-emitting diodes of comparable area itemized above, at least in the current state of their technological development. Specifically, RGB MQWLEDs reliably exhibit superior external quantum efficiency, power-conversion efficiency, radiant flux, luminous flux, wall-plug luminous efficacy, and wall-plug luminous efficiency. Moreover, the performance of MQWLEDs scales with device area so that high luminous flux and efficacy are the norm — these devices are no longer characterized as low-, medium-, or high-power, as they were in the early days.

MQWLEDs are therefore the preferred choice for use in LED lighting today, and this chapter and the next are devoted to exploring their use in this capacity. Some of the other classes of LEDs listed above are currently undergoing extensive development, however, and hold substantial promise for the LED lighting technologies of tomorrow. In this connection, it is important to note that the use of multiple chips operating at low current, in place of a single chip that operates at high current, can mitigate efficiency droop, increase device lifespan, and simplify thermal management.

Monochromatic LED Light

An individual electroluminescent light-emitting diode (ELLED) emits narrowband light over a limited range of wavelengths that is principally established by the bandgap wavelength λ_g of the material from which it is fashioned (Sec. 6.4 and Fig. 7.2-3). Stand-alone, single-die devices that are not integrated into a larger system, commonly called **discrete LEDs**, are available with an endless array of central wavelengths. The light generated by these devices is termed **quasi-monochromatic**, signifying that its spectral width is sufficiently narrow, relative to its central wavelength, that its behavior is effectively monochromatic in the context at hand. Indeed, the prefix *quasi* is often omitted as a shorthand and ELLED light is simply referred to as **monochromatic**. Stated differently, LED electroluminescence is partially coherent but, in the context of LED lighting, it can be considered to be coherent (Sec. 2.7).

Human Color Gamut

By virtue of its monochromaticity, LED light emitted at a particular visible wavelength maps to a specific location along the outer curved boundary of the CIE 1931 chromaticity diagram portrayed in Fig. 9.6-2. Such light is perceived by the visual system as a spectral color, of which the eye can distinguish about 200. Hence, a collection of some 200 LEDs can in theory be used to evoke the perception of any hue along the locus of fully saturated spectral colors. As explained in Sec. 9.6, adding the light emitted from a properly chosen auxiliary source of adjustable luminance can then be used to create a broad range of desaturated colors, including white. Indeed, human trichromats are estimated to be able to discern 2 million color gradations, comprising 200 hue, 20 saturation, and 500 luminance gradations (Example 8.7-3). In principle, therefore, a collection of LEDs operating at different visible wavelengths, together with associated

desaturating auxiliary sources, can evoke any perceptible color within the gamut of human color vision.

Fortunately, the trichromatic nature of the human visual system offers a dramatic shortcut for accessing the color gamut. Mixing the light from only three monochromatic sources, each with fixed coordinates on the chromaticity diagram and with adjustable luminance, provides access to the full palette of colors enclosed by the triangle whose vertices are located at those chromaticity coordinates (Sec. 9.6). This feature of human color vision is repeatedly invoked in this and the following chapter.

The trichromatic character of the human visual system makes it possible for a suitably chosen triad of monochromatic LED lights of fixed hues and adjustable luminances to dial up a color of arbitrary hue, saturation, and luminance.

White LED Light

White is by far the most important color for general-purpose illumination, as well as for numerous applications such as high-quality projection. Yet, white is essentially the antithesis of the spectral colors generated by electroluminescent LEDs: as opposed to a fully saturated spectral color located on the outer rim of the chromaticity diagram (Fig. 9.6-2), white is a fully desaturated achromatic color found near the center of the diagram. True white light has a uniform power-spectral density, as elucidated in Sec. 9.3 and Example 9.6-3, whereas metameric white light is *perceived* as white by the visual system despite the fact that its power-spectral density is nonuniform.

Light-emitting diodes do not generate true white light, which has a uniform power-spectral density. Nonetheless, monochromatic LEDs are widely used for generating* METAMERIC WHITE LIGHT, *which is* perceived *as white despite its nonuniform power-spectral density. When we speak of white LED light, it is understood that what we really mean is metameric-white LED light.

Three general methods exist by means of which monochromatic LED light can be converted into metameric white light:

Phosphor-Conversion Devices. The *first method*, which is nearly universally used because of its simplicity, effectiveness, and low cost, employs **phosphor-conversion devices** that contain one or more discrete PCLEDs. In its simplest implementation, as is understood from Example 9.3-1, a PCLED comprising a royal blue LED and a yellow phosphor can generate metameric cool-white light via photoluminescence. A red phosphor can be added to the blend to create a PCLED that generates warmer metameric white light with a reduced correlated color temperature (CCT, Sec. 9.8) and an increased color rendering index (CRI, Sec. 9.9). Narrowband red phosphors offer higher luminous flux and efficacy than broadband ones since their photoluminescence spectra do not extend into the (invisible) near-infrared region, which allows their radiant flux to be fully utilized (Secs. 10.3 and 10.4).

This chapter is devoted to elucidating the operation and use of phosphor-conversion LEDs. In particular, we examine:

- Discrete PCLEDs (Secs. 10.5 and 10.6). These stand-alone devices can be conceptualized as 0D (zero-dimensional) sources.
- PCLED filaments (Sec. 10.7). These chains of discrete PCLEDs can be conceptualized as 1D sources.
- Chip-on-board (COB) PCLEDs (Sec. 10.8). These arrays of discrete PCLEDs can be conceptualized as 2D sources.

The correlated color temperature of the metameric white light emitted by PCLED devices can range from cool-white to warm-white, depending on the choice of phosphor:

- Metameric cool-white light can be generated by making use of a broadband-yellow phosphor (Sec. 10.5).
- Metameric warm-white light can be generated by using a phosphor blend that incorporates a broadband-red phosphor (Sec. 10.6).
- Metameric warm-white light can be more efficiently generated by employing a phosphor blend that incorporates a narrowband-red phosphor (Sec. 10.6).

Hundreds (if not thousands) of commercially available LEDs are available for generating monochromatic light of arbitrary wavelengths. Also available in the marketplace are hundreds of PCLEDs that emit colored light and metameric white light of arbitrary correlated color temperature (Sec. 11.2).

Two other methods exist for converting monochromatic LED light into metameric white light, as will be discussed in Chapter 11, but they are not often used:

Additive Color Mixing. The *second method* for generating metameric white light, known as **additive color mixing**, relies on superposing the light generated by several LEDs of different colors. This approach has the merit of being able to generate white light more efficiently, in principle, and it offers color-tunable LED lighting, but it is also subject to a number of difficulties, as described in Sec. 11.3.

Hybrid Approach. The *third method* for generating metameric white light, often referred to as the **hybrid approach**, makes use of two or more LEDs of different colors (e.g., blue and red), in conjunction with one or more phosphors, as explained in Sec. 11.5. Although the hybrid method was successfully used for the efficient generation of metameric white light early on, its complexity and cost resulted in its being abandoned soon thereafter in favor of the phosphor-conversion approach. Today, the hybrid approach is widely used for the generation of light of tunable color.

10.2 PHOTOLUMINESCENCE

We begin with an examination of photoluminescence, the process that underlies the operation of PCLEDs.

Luminescence

Thermal excitation, studied in Chapter 4, and current injection, considered in Chapter 6, are but two examples of how a material system can be excited to a higher energy level and then emit light as it subsequently decays to the ground state. Light can also be emitted as a consequence of excitation by other forms of energy, including beams of photons, electrons, ionizing particles, and sound; as well as by the release of energy by chemical and biological reactions. The resulting radiation process is then described by the umbrella term **luminescence**, and the atomic or molecular entity that emits such light is known as a **luminophore**. Luminescence generated with a time lag of ps to μs following excitation is also called **fluorescence**, whereas luminescence delayed by ms or longer after excitation is also referred to as **phosphorescence**.

The molecular entity responsible for fluorescence is termed a* FLUOROPHORE*, whereas the entity underlying phosphorescence is referred to as a* PHOSPHOR*.

From a photon statistics point-of-view, luminescence is generally well-described by the Neyman Type-A photon-counting distribution rather than by the negative-binomial photon-counting distribution that characterizes thermal light, as specified in (4.2-16).

Photoluminescence

The particular form of luminescence known as **photoluminescence** refers to a process whereby a system excited to a higher energy level by the absorption of a photon subsequently decays to a lower energy level, usually via a combination of radiative and nonradiative transitions. Conservation of energy requires that the emitted photon have energy less than (or equal to) that of the exciting photon, so the luminophore can be said to act as a **downconversion medium**. (Photoluminescence downconversion is to be distinguished from *parametric downconversion*, an energy-conserving process in which a photon splits into two lower energy photons in a nonlinear medium.) Photoluminescence occurs naturally in many substances, including inorganic materials such as diamond and ruby; semiconductors such as CdSe and metal-halide perovskites; noble gases; simple inorganic molecules such as N_2 and CO_2; and aromatic compounds such as dyes. A commonly encountered example of photoluminescence is the phosphorescent glow emitted by certain materials following exposure to ultraviolet (black) light. The photoluminescence emitted by chalcogenide and perovskite colloidal quantum dots of different sizes is illustrated in Figs. 5.8-1(a) and (b), respectively.

Figure 10.2-1 displays a number of idealized schemes by means of which ionic, atomic, and molecular transitions can lead to photoluminescence. The individual energy levels are depicted as horizontal sharp lines, separated by the energy of the pump photon, but in practice they can be energy bands. The solid vertical lines represent radiative transitions and the wiggly lines represent the absorption and emission of photons. Nonradiative downward transitions, depicted by dashed vertical lines, participate in photoluminescence, as sketched in Figs. 10.2-1(a)–(c).

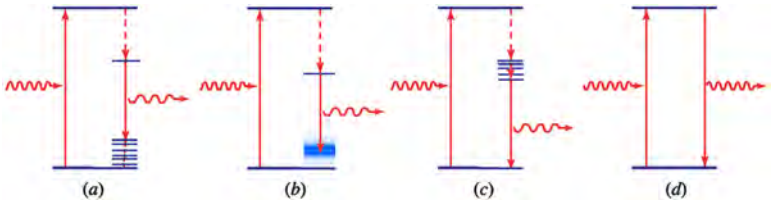


Figure 10.2-1 Generation of photoluminescence from ionic, atomic, or molecular transitions for several idealized energy-level schemes. The upper energy levels are illustrated as sharp horizontal lines, but they can instead be bands. The solid and dashed vertical lines represent radiative and nonradiative transitions, respectively, while the wiggly lines represent absorbed and emitted photons. (a)–(c) Photoluminescence accompanied by various forms of nonradiative decay. In each panel, the photoluminescence quantum yield η_{PLQY} is the probability that a photon entering at left yields a photon exiting at right. The complementary photoluminescence quantum defect $\overline{\eta}_{\mathrm{PLQD}}$ in each panel is the ratio of the length of the vertical solid line at right to that at left. (d) The Rayleigh scattering diagram is shown for comparison; in this case, the photon energy is conserved but its direction of travel is altered.

Photoluminescence Quantum Yield (PLQY)

An oft-used measure for characterizing phosphor-conversion LEDs, as well as organic and perovskite semiconductor materials (Sec. 5.9), is the **photoluminescence quantum yield (PLQY)**. This quantity represents the probability that a photon incident on a photoluminescent material results in the emission of another, lower frequency, photon [Fig. 10.2-1(a)–(c)]. Since a photoluminescent material can relax via both radiative and nonradiative decay following excitation, the PLQY can be expressed as the ratio of the radiative decay rate κ_r to the overall (radiative plus nonradiative) decay rate $\kappa = \kappa_r + \kappa_{nr}$,

$$\eta_{\mathrm{PLQY}} = \frac{\kappa_r}{\kappa} = \frac{\kappa_r}{\kappa_r + \kappa_{nr}}. \tag{10.2-1}$$

The PLQY characterizes the efficiency of photoluminescence emission. It can be maximized by simultaneously making the radiative and nonradiative decay rates as large and as small as possible, respectively. The PLQY is analogous to the internal quantum efficiency (IQE) in a semiconductor, reported in (7.1-5), which is the ratio of the radiative electron–hole recombination coefficient for electroluminescence to the total (radiative plus nonradiative) recombination coefficient, as provided in (5.5-10). Phenomena such as self-absorption and Auger-like effects reduce the PLQY.

Photoluminescence Quantum Defect (PLQD)

Moreover, a photon incident on a photoluminescent material that is successfully converted to a lower-energy photon loses a portion of its energy via nonradiative decay in the course of the conversion [Figs. 10.2-1(a)–(c)]. Sometimes referred to as **Stokes energy loss**, this process lowers the frequency of the converted photon by virtue of (3.2-1). The **photoluminescence quantum defect** η_{PLQD} is defined as the fraction of the energy of the incident photon that is, on average, *lost* in the process of photoluminescence, i.e.,

$$\eta_{\text{PLQD}} = 1 - h\nu_2/h\nu_1 = 1 - \lambda_1/\lambda_2\,, \tag{10.2-2}$$

where ν_2 (λ_2) and ν_1 (λ_1) represent the peak frequency (wavelength) of the photoluminescence and incident light, respectively (Sec. 3.2). The **complementary photoluminescence quantum defect** $\overline{\eta}_{\text{PLQD}}$, defined as the fraction of the incident photon's energy that is *effective* in generating the lower-energy luminescence photon, is therefore given by

$$\overline{\eta}_{\text{PLQD}} \equiv 1 - \eta_{\text{PLQD}} = \frac{h\nu_2}{h\nu_1} = \frac{\lambda_1}{\lambda_2}\,. \tag{10.2-3}$$

10.3 BROADBAND AND NARROWBAND PHOSPHORS

The phosphors that generate photoluminescence for use in LED lighting are usually **dielectric hosts** doped with ions called **activators** that are uniformly dispersed throughout the material with a specified atom concentration (often of the order of 1%). Inorganic phosphors, both broadband and narrowband, are widely used for fabricating PCLEDs that emit white light as well as light of various colors (Sec. 11.2).These phosphors are usually ground into a fine powder before being directly coated on the die of a pump LED, which is typically a blue InGaN device. The photoluminescence results from downward transitions of the valence electrons between particular upper and lower energy levels, as schematized in Fig. 10.2-1(a)–(c).

Desirable features for phosphors include the following:

- High photoluminescence quantum yield.
- Highly saturated photoluminescence.
- Chemical stability.
- Thermal stability.
- High operating temperature stability.
- High humidity stability.
- Stability in the presence of irradiation with high-flux blue light.
- Long lifespan.
- Capability of being directly coated onto a an LED die.
- Narrowband red photoluminescence to minimize unutilized near-infrared light.
- Narrowband photoluminescence in the green and amber to mitigate the electroluminescence green gap and to attain a wide color gamut (WCG).

Atomic Physics of Activator Optical Transitions

We begin by reviewing the principles of atomic physics that underlie the emission of photoluminescence from LED phosphors. This is followed by brief descriptions of a number of widely used broadband and narrowband phosphors. Frequently encountered activators include lanthanide-metal ions (e.g., Ce^{3+}, Eu^{3+}, Eu^{2+}) and transition-metal ions (e.g., Mn^{4+}). Common host materials include garnets, nitrides, fluorometallates, oxynitrides, oxides, halides, sulfides, and uranium-containing compounds, among others, as attested to the entries in the bibliography.

The properties of several salient photoluminescent phosphors that make use of these activators and hosts are presented in Table 10.3-1. The photoluminescence generated by some phosphors is broadband (BB), while that generated by others is narrowband (NB); the boundary between BB and NB is arbitrarily set at $\Delta\lambda_{\text{FWHM}} = 60$ nm. The excited and ground states of the activator ions displayed in Table 10.3-1 are identified by their electron configurations and term symbols. These quantities, which are elucidated below, contain information about the energy levels, transitions, and efficiencies that characterize the generation of light of a particular color in the phosphor.

Table 10.3-1 Selected properties of representative photoluminescent phosphors incorporated in blue LEDs to generate metameric white light. Consecutive columns in the table represent: chemical formulas; designations of the photoluminescence as broadband (BB) or narrowband (NB), along with the colors of the emitted light; electron configurations (and atomic term symbols $^{2S+1}L_J$ or, for Mn^{4+}, molecular term symbols $^{2S+1}M_J$) for excited states (left) and ground states (right); average emission wavelengths $\overline{\lambda}$ (nm); photoluminescence bandwidths $\Delta\lambda_{\text{FWHM}}$ (nm); photoluminescence quantum yields η_{PLQY}; and complementary photoluminescence quantum defects $\overline{\eta}_{\text{PLQD}}$.

PHOSPHOR	Color	Configuration (Term)	$\overline{\lambda}$	$\Delta\lambda_{\text{FWHM}}$	η_{PLQY}	$\overline{\eta}_{\text{PLQD}}$
$Y_3Al_5O_{12}$:Ce^{3+}	BB Yellow	$5d^1\ (^2D) \rightarrow 4f^1\ (^2F_{5/2})$	570	120	0.90	0.78
$Y_3Al_5O_{12}$:Eu^{3+}	BB Red-Orange	$4f^6\ (^5D_0) \rightarrow 4f^6\ (^7F_0)$	615	110	0.90	0.72
$CaAlSiN_3$:Eu^{2+}	BB Red	$4f^6\,5d^1\ (^6P_{7/2}) \rightarrow 4f^7\ (^8S_{7/2})$	650	80	0.90	0.68
K_2SiF_6:Mn^{4+}	NB Red	$3d^3\ (^2E) \rightarrow 3d^3\ (^4A_2)$	631	23	0.90	0.71
Na_2SiF_6:Mn^{4+}	NB Red	$3d^3\ (^2E) \rightarrow 3d^3\ (^4A_2)$	628	16	0.90	0.71
β-SiAlON:Eu^{2+}	NB Green	$4f^6\,5d^1\ (^6P_{7/2}) \rightarrow 4f^7\ (^8S_{7/2})$	540	55	0.90	0.82

Electron Configuration. As is understood from the structure of the periodic table of the elements, the electrons of multielectron atoms reside in shells designated by the principal quantum number $\mathfrak{n}$; each shell can accommodate only a specified number of electrons before being filled. Within each shell are **subshells**, also called **orbitals**, that are designated by the orbital angular-momentum quantum number of its resident electrons, $\ell = 0,\ 1,\ 2,\ 3, \ldots$, or equivalently by *lowercase* letters that represent the vestigial notation bequeathed to us from the early days of observational atomic spectroscopy: $s,\ p,\ d,\ f, \ldots$ (*s*harp, *p*rincipal, *d*iffuse, *f*undamental) $\Leftrightarrow\ \ell = 0,\ 1,\ 2,\ 3, \ldots$. The **electron configuration**, which represents the arrangement of electrons in the shells and subshells, is expressed as a sequence of orbitals of the form $\mathfrak{n}\ell^{\mathfrak{u}}$, where the superscript $\mathfrak{u}$ designates the number of electrons associated with each orbital ℓ.

In accordance with the **Aufbau principle**, electrons fill the orbitals in order of increasing energy. Convention dictates that the electron configurations for filled shells and subshells be omitted from the recitation. An an example, the full electron configuration for the ground state of the Ce^{3+} ion, $1s^2\,2s^2\,2p^6\,3s^2\,3p^6\,3d^{10}\,4s^2\,4p^6\,4d^{10}\,5s^2\,5p^6\,4f^1$, is written as [Xe] $4f^1$, as indicated in Table 10.3-1. This shorthand notation specifies that the electron configuration is that of the configuration for the noble gas xenon [Xe], supplemented by the single valence electron $4f^1$ in the outermost orbital of Ce^{3+}.

Atomic Term Symbol. For the lighter elements, the various angular momenta of the ion (or atom) as a whole obey LS (or Russell-Saunders) coupling and are described by an **atomic term symbol** of the form $^{2S+1}L_J$. The symbol S, which represents half the number of unpaired electrons, is the *total spin angular-momentum* quantum number. The quantity $2S + 1$, the *spin multiplicity* discussed in Sec. 7.6, indicates the spin degeneracy: for the singlet state $S = 0$ and $2S + 1 = 1$ whereas for the triplet state $S = 1$ and $2S + 1 = 3$. The symbol L represents the *total orbital angular-momentum* quantum number, expressed in *uppercase* spectroscopic notation as $L = 0, 1, 2, 3, \ldots \Leftrightarrow S, P, D, F, \ldots$. The symbol J represents the *total overall angular-momentum* quantum number.

As an example, the term symbol for the ground state of Ce^{3+} is determined from the electron configuration outside the core of closed shells as follows: The configuration $4f^1$ specifies that $\mathfrak{n} = 4$, $\ell = 3$, and $\mathfrak{u} = 1$. The value of $\mathfrak{u}$ reveals that there is only a single valence electron, so that $S = 1/2$, $2S+1 = 2$ (a doublet); and $L = \ell = 3$ (denoted F). Finally, LS coupling declares that J stretches from $L + S = 7/2$ to $|L - S| = 5/2$, in intervals of unity. In accordance with **Hund's rule** (which is usually obeyed), if the orbital is less than half-filled the ground state corresponds to the smallest value of J. We conclude that the term symbol for the ground state of Ce^{3+} is $^{2S+1}L_J = {}^2F_{5/2}$, as specified in Table 10.3-1. The heavier elements have stronger spin-orbit interactions and generally follow JJ, rather than LS, coupling.

Lanthanide-Metal Activators. The **lanthanide-series** elements reside in row 6 of the periodic table and comprise the elements from $_{57}$La through $_{71}$Lu. These fifteen elements, plus $_{21}$Sc and $_{39}$Y, are often called the **rare-earth elements** because they were long ago thought to be rare (they are in fact rarely rare). Starting with La, the lanthanide-series elements are constructed by successively adding one electron at a time to the $(\mathfrak{n}\ell =) 4f$ orbital, which, by the vagaries of atomic physics, lies between the filled $5s^2\,5p^6$ and the $5d^1\,6s^2$ subshells. The electron configuration for these elements therefore extends from [Xe] $4f^1\,5d^1\,6s^2$ for the ground state of the neutral $_{58}$Ce atom to [Xe] $4f^{14}\,5d^1\,6s^2$ for the ground state of the neutral $_{71}$Lu atom.

A characteristic feature of the lanthanides is that they readily form triply ionized states by losing one electron from the $5d$ orbital and two from the $6s$ orbital. The electron configurations for the corresponding ions thus stretch from [Xe] $4f^1$ for the ground state of Ce^{3+} to [Xe] $4f^{14}$ for the ground state of Lu^{3+}; the configurations for the core electrons remain the same as those of the parent atoms. As indicated in Table 10.3-1, an optical transition from the ground state to an excited state via the absorption of a blue photon can involve a change in orbitals and term symbols, as exemplified by Ce^{3+} ($4f^1 \rightarrow 5d^1$ and ${}^2F_{5/2} \rightarrow {}^2D$, respectively), or a reorganization among the energy states within an orbital and a change in term symbols, as for Eu^{3+} ($4f^6 \rightarrow 4f^6$ and ${}^7F_0 \rightarrow {}^5D_0$, respectively).

The energy levels and transition cross sections of a phosphor depend on both the activator and on its interaction with the host. The extent to which the activator energy levels are affected by the host medium is principally established by the degree to which the ion's valence electrons are exposed to the host's neighboring lattice atoms. The energy levels of the trivalent lanthanide ion Eu^{3+}, for example, are only weakly influenced by the local fields of the host lattice.

Transition-Metal Activators. In transition-metal ions, in contrast, the valence electrons are not shielded from the host's neighboring lattice atoms so that their energy levels are strongly influenced by the host. The energy levels of the $3d$ electrons of the transition-metal ion Mn^{4+}, for example, are determined in large part by the surrounding electric fields of the host. In particular, each manganese ion in K_2SiF_6:Mn^{4+} is surrounded by an atomic configuration in which the spatially varying potential of the

KSF host is significant. Best represented in terms of **ligand-field theory**, this potential, along with that of the Mn^{4+} nucleus, jointly determine the molecular energy levels of the phosphor via the Schrödinger equation.

Molecular Term Symbol. The combined effects of host ligand fields and activator d-orbital electrons are described by a **molecular term symbol** of the form $^{2S+1}M_J$, where M is called the Mulliken symbol. Molecular term symbols follow many of the same notational conventions as atomic term symbols, although J values are sometimes omitted. The Mulliken symbols A, E, and T (which studiously avoid the atomic designations $SPDF\ldots$) represent nondegenerate, doubly degenerate, and triply degenerate electronic states, respectively. Referring to the entries for KSF and NSF in Table 10.3-1 as examples, $^4A_2$ signifies a quartet ($S = 3/2$) that has a nondegenerate electronic state while ^2E represents a doublet ($S = 1/2$) that has a doubly degenerate electronic state.

Phosphor Broadening Mechanisms. Photoluminescence was described in considerable detail in Sec. 10.2. As illustrated schematically in Fig. 10.2-1, phosphors are subject to an array of broadening mechanisms that include:

- Crystal-field splittings actualized by the crystal symmetry, crystal-field strength, atom coordination, and polarizability of the host medium.
- Inhomogeneous broadening associated with variations in the local environment experienced by the activator (Sec. 4.6).
- Homogeneous broadening associated with excited-state nonradiative vibrational-mode relaxation via multiphonon transitions.
- Defects and impurities in the host.
- Enhanced phonon interactions brought about by higher operating temperatures.

Phosphors emit photoluminescence with a* BROADBAND *or* NARROWBAND *spectral density, depending on the atomic structure of the activator, the nature of the host, and the interaction between the activator's electrons and the host's lattice.

Yellow phosphors are usually used for fabricating discrete cool-white PCLEDs (Sec. 10.5), while phosphor blends that incorporate red, and sometimes also green, phosphors are used to fashion discrete warm-white PCLEDs (Sec. 10.6). Similar phosphors are used in the fabrication of PCLED filaments (Sec. 10.7) and for chip-on-board PCLEDs (Sec. 10.8). Phosphors are also used for creating discrete PCLEDs that emit light of various colors (Sec. 11.2).

Broadband Phosphors

We begin by offering brief descriptions of the properties of selected yellow, red-orange, and red broadband phosphors that are commonly used in the fabrication of white PCLEDs (Table 10.3-1).

Cerium-Doped Yttrium Aluminum Garnet (Broadband Yellow). As a consequence of its high thermal stability, long lifespan, and other desirable features itemized in the introduction to this section, cerium-doped YAG ($Y_3Al_5O_{12}$:Ce^{3+}) has long been used as a yellow photoluminescent phosphor for discrete cool-white PCLEDs (Sec. 10.5). As reported in Table 10.3-1, when stimulated in the blue, YAG:Ce^{3+} emits at an average wavelength $\bar{\lambda} \approx 570$ nm, has a photoluminescence bandwidth of $\Delta\lambda_{\text{FWHM}} \approx 120$ nm, and a typical photoluminescence quantum yield $\eta_{\text{PLQY}} \approx 0.9$. The broadband nature of the photoluminescence is principally a result of nonradiative relaxation involving multiphonon transitions associated with the vibrational modes of the YAG lattice. Since each individual phonon transition carries a random energy,

whose average is $\approx$ 0.05 eV, the emission of a single photoluminescence photon involves tens of phonons of random energies, which results in broadband emission, as illustrated in Fig. 10.2-1(b).

Europium-Doped Yttrium Aluminum Garnet (Broadband Red-Orange). The most commonly used broadband red phosphor in warm-white PCLEDs is europium-doped YAG ($Y_3Al_5O_{12}$:Eu^{3+}), whose properties are reported in Table 10.3-1. This material, which has many of the salutary features itemized at the beginning of this section, efficiently emits red-orange light at $\overline{\lambda} \approx 615$ nm with $\Delta\lambda_{\text{FWHM}} \approx 110$ nm, when stimulated in the blue. Some warm-white PCLEDs also incorporate a green phosphor, such as europium-doped β-silicon aluminum oxynitride (β-SiAlON:Eu^{2+}), or a similar compound, to provide spectral broadening and thereby enhance color rendering quality, as will be discussed in Sec. 10.4.

Europium-Doped Calcium Aluminum Silicon Nitride (Broadband Red). Divalent Eu^{2+} activators have many pathways for $5d \rightarrow 4f$ transitions, and there is no shortage of hosts within which such transitions can take place. Since the $5d$ electrons are exposed to the crystal lattice of the host medium, Eu^{2+}-doped phosphors can exhibit photoluminescence at many wavelengths in the visible. An important example is the material $CaAlSiN_3$:Eu^{2+}, typically doped at an atom-percent level of $\approx 2\%$, which has a broad absorption band centered on the blue. When stimulated by the light from a blue LED, this phosphor efficiently generates red light at $\overline{\lambda} \approx 650$ nm with $\Delta\lambda_{\text{FWHM}} \approx 80$ nm, as shown in Table 10.3-1.

Narrowband Phosphors

We turn now to the properties of selected red and green narrowband phosphors (Table 10.3-1) that are incorporated into phosphor blends for use in warm-white phosphor-conversion devices. The advantages provided by utilizing narrowband phosphors in place of broadband ones are discussed in Secs. 10.4 and 10.6.

Manganese-Doped Potassium Fluorosilicate (Narrowband Red). The narrowband phosphor manganese-doped potassium fluorosilicate (KSF or PFS, K_2SiF_6:Mn^{4+}) is widely used as a red phosphor in PCLEDs because of its many salutary features. In particular, the formulation of KSF introduced by the General Electric Company in 2014, and trademarked GE TriGain™, exhibits superior thermal and chemical stability, and longer lifespan, than earlier versions of this phosphor. When pumped by light from a blue InGaN LED, KSF generates five strong vibronic sidebands surrounding the $^2E \rightarrow {}^4A_2$ zero-phonon line (ZPL) at 623 nm (Table 10.3-1), each with a width of $\approx$ 2 nm. These vibronic molecular transitions have large transition rates because of the odd (*u*ngerade) parity of the (bend and stretch) modes of the MnF_6 octahedral moiety (quantum-mechanically speaking, the transitions are said to be electric-dipole allowed). However, the ZPL at 623 nm, known as the R line, is barely discernable in KSF because of the octahedral inversion symmetry at the activator site, which results in a small transition rate (the transition is electric-dipole forbidden).

Specifically, the Stokes vibronic sidebands are located at $\overline{\lambda} \approx 631$, 635, and 648 nm, while the anti-Stokes sidebands are at 613 and 609 nm; the envelope of these sidebands has a bandwidth $\Delta\lambda_{\text{FWHM}} \approx 23$ nm, as recorded in Table 10.3-1. These sidebands are readily identifiable as the five spikes in the vicinity of 623 nm in Figs. 10.4-2 and 10.4-3. The two curves in Fig. 10.4-2 represent the spectral densities for Cree (green) and Philips (purple) phosphor blends that incorporate KSF and generate metameric white light at $T_c = 3000$ K. The green curve in Fig. 10.4-3 offers an unobstructed view of

the KSF spectrum since the green GE phosphor used in that particular blend is also narrowband.

Other Narrowband Red Phosphors. A compound closely related to KSF that is of particular interest is Na_2SiF_6:Mn^{4+} (NSF), which, by virtue of its strong R line (ZPL) emission and narrow bandwidth, provides slightly greater wall-plug luminous efficacy than K_2SiF_6:Mn^{4+} (KSF). This phosphor is currently being commercialized.

Numerous other alkali and alkaline-earth hexafluorometallate, red-emitting KSF congeners have also been investigated for use as narrowband red phosphors. These include materials in the following classes:

1. Compounds of the form $\mathfrak{A}_2[\mathfrak{S}F_6]$:$Mn^{4+}$, where $\mathfrak{A}$ represents NH_4 or an alkali metal in group I of the periodic table, such as Li, Na, K, Rb, Cs, or combinations thereof; and $\mathfrak{S}$ represents group IV and other elements such as Si, Ge, Sn, Ti, Zr, Al, Ga, In, Sc, Hf, Y, La, Nb, Ta, Bi, Gd, or combinations thereof.
2. Compounds of the form $\mathfrak{E}[\mathfrak{S}F_6]$:$Mn^{4+}$, where $\mathfrak{E}$ represents an alkaline-earth element in group II of the periodic table, such as Mg, Ca, Sr, Ba, Zn, or combinations thereof,
3. Compounds that rely on F_5, F_6, or F_7, and other compounds that have been published, disclosed in patents, or proposed.

Europium-Doped Beta-Silicon Aluminum Oxynitride (Narrowband Green). Much as with the phosphor $CaAlSiN_3$:Eu^{2+}, which generates broadband red light as discussed earlier, the divalent Eu^{2+} activator can undergo $5d \rightarrow 4f$ transitions in a β-SiAlON host. When pumped by a blue LED, the phosphor β-SiAlON:Eu^{2+}, which was developed in 2005 at the National Institute for Materials Science (NIMS) in Tsukuba, Japan, generates narrowband green photoluminescence at $\overline{\lambda} \approx 540$ nm with $\Delta\lambda_{\text{FWHM}} \approx 55$ nm (Table 10.3-1). This green phosphor, or one of its congeners, is often blended with a narrowband red phosphor such as KSF (Example 10.4-2). When ground into fine particles and dispersed in a silicone binder, and then directly deposited on a blue InGaN chip, this phosphor blend serves as a wide-color-gamut source of light with $\mathfrak{RGB}$ primaries (Sec. 9.5).

Other Narrowband Green Phosphors. In recent years, General Electric has developed a number of narrowband green phosphors for on-chip and remote use that are said to offer wider color gamut and greater humidity resistance than β-SiAlON:Eu^{2+}. These proprietary phosphors are currently in the process of commercialization.

Many other narrowband green phosphors are also available, including lanthanum- and transition-ion activated alkali silicates, uranium-based compounds, sulfides, and garnets, examples of which are the following:

1. Lithium-silicate based $Na_vK_xRb_yLi_zCs_w\,(Li_3SiO_4)_4$:$\mathfrak{L}$ with $0 < v < 4$, $0 < x < 4$, $0 < y < 4$, $0 < z < 4$, $0 < w < 4$, and $v + x + y + z + w = 4$, where $\mathfrak{L}$ represents an activator such as Eu^{3+}, Ce^{3+}, Yb^{3+}, or Mn^{4+}, or combinations thereof.
2. Uranium-based $[Ba_{1-a-b}Sr_aCa_b]_x\,[Mg,Zn]_y\,(UO_2)_z\,([P,V]O_4)_{2(x+y+z)/3}$:$Eu^{3+}$ with $0 \leqslant a \leqslant 1$, $0 \leqslant b \leqslant 1$, $0.75 \leqslant x \leqslant 1.25$, $0.75 \leqslant y \leqslant 1.25$, and $0.75 \leqslant z \leqslant 1.25$.
3. Uranium-based $[Ba_{1-a-b}Sr_aCa_b]_p\,(UO_2)_q\,[P,V]_r\,O_{(2p+2q+5r)/2}$:$Eu^{3+}$ with $0 < a \leqslant 1$, $0 < b \leqslant 1$, $2.5 \leqslant p < 3.5$, $1.75 < q \leqslant 2.25$, and $3.5 < r \leqslant 4.5$.
4. Sulfides such as strontium gallium sulfide ($SrGa_2S_4$:Eu) and garnets such as lutetium aluminum garnet (LAG or $Lu_3Al_5O_{12}$:Ce^{3+}), and its congeners.

10.4 BLENDED PHOSPHORS

As mentioned in the introduction to this Chapter, the simplest implementation of a discrete phosphor-conversion LED makes use of a single yellow phosphor (usually YAG:Ce^{3+}) in conjunction with a royal-blue LED die. This combination is suitable for generating cool (daylight) white light ($5000 \lesssim T_c \lesssim 7500$ K), as will be demonstrated in Example 10.5-1. For many applications, however, it is desired to generate neutral white light ($3500 \lesssim T_c \lesssim 5000$ K) or warm (soft) white light ($2700 \lesssim T_c \lesssim 3500$ K). (The manner in which white light sources are categorized in illumination engineering has been spelled out in Secs. 9.7 and 9.8.) The latter range of CCTs mimics the color temperature of classic tungsten incandescent lamps and is thus often favored for home illumination at eventide, at least in chillier climes. A practical, and widely used, method for generating neutral- and warm-white light, while concomitantly attaining a high value of the CRI, relies on the use of blended phosphors, which we consider in this section. The phosphor thickness can be adjusted across the lateral extent of the device to achieve the desired spatial light distribution.

Use of Narrowband vs. Broadband Phosphors in Blends. It will be demonstrated in Sec. 10.6 for discrete warm-white PCLEDs, and in Sec. 10.8 for warm-white COB PCLEDs, that incorporating a narrowband red phosphor in the blend, rather than a broadband one, provides an enhancement in device performance. Specifically, it will be shown in Tables 10.6-1 and 10.8-1 that the use of KSF, rather than YAG:Eu^{3+}, leads to enhanced values of the luminous flux $P_{\mathtt{V}}$, wall-plug luminous efficacy $\eta_{\mathtt{WPE}}$, and wall-plug luminous efficiency $\eta_{\mathtt{WPC}}$.

The origin of this advantage may be understood by examining Fig. 10.4-1, which displays the spectral densities for blue-excited, 3000-K Cree blends that incorporate red phosphors that are broadband (red curve) and narrowband (green curve). Also illustrated is the photopic luminous efficiency function first presented in Fig. 8.5-3. Denoted $V(\lambda_0)$ and plotted as the gray curve in Fig. 10.4-1, this function specifies the overall photopic sensitivity of the human visual system as a function of the free-space wavelength λ_0. All three curves are normalized to their maximum values.

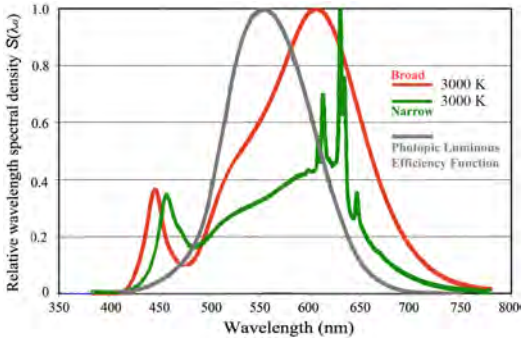


Figure 10.4-1 Spectral densities for Cree metameric white phosphor blends that incorporate broadband and narrowband red phosphors are indicated by the red and green curves, respectively. The photopic luminous efficiency function $V(\lambda_0)$ is plotted in gray. The three curves are normalized to their peak values. (Broadband data adapted from Cree data sheet CLD-DS199-REV7 for XLamp® XHP35.2, 2023; narrowband data adapted from Cree data sheet CLD-DS334-REV1 for XLamp® XHP35.2 Pro9™, 2023.)

The performance enhancement stems from the smaller photoluminescence bandwidth of narrowband KSF (K_2SiF_6:Mn^{4+}) relative to that of broadband YAG:Eu^{3+} ($Y_3Al_5O_{12}$:Eu^{3+}): $\Delta\lambda_{\mathrm{FWHM}} \approx 23$ nm for KSF:Mn^{4+} as opposed to ≈ 110 nm for YAG:Eu^{3+} (Table 10.3-1). It is evident in Fig. 10.4-1 that the preponderance of the KSF contribution to the spectral density (the five narrow spikes of the green curve) falls in a wavelength region where $V(\lambda_0)$ (the gray curve) is appreciable. In contrast, a sizable portion of the YAG:Eu^{3+} contribution to the spectral density (the right-hand tail

of the red curve) extends over a wavelength region where $V(\lambda_0)$ is smaller, indicating that the eye is less sensitive, so that the radiant flux in that region is largely squandered. Since the narrowband phosphor does not generate light at these longer wavelengths to begin with, its provides superior performance.

Phosphor Blends Incorporating KSF. Manganese-doped potassium fluorosilicate (KSF), the narrowband red phosphor examined in detail in Sec. 10.3, is a prime choice for the red component in many blended phosphors because it possesses almost all of the desirable features for phosphors itemized in the introduction to Sec. 10.3 and has a number of special merits in addition:

- Strong absorption in the vicinity of 450 nm, where blue InGaN LEDs emit, along with high photoluminescence quantum yield.
- Saturated, narrowband photoluminescence centered at 631 nm, which is near the He–Ne 633-nm red laser line, where the sensitivity of the eye is appreciable.
- Negligible emission for wavelengths beyond 650 nm, where the light energy is squandered because those wavelengths are invisible to the eye.
- Ability to be consolidated as a transparent ceramic phosphor (as with YAG:Ce^{3+}).

It has long been common practice to blend KSF with the classic yellow phosphor YAG:Ce^{3+} to enable the generation of metameric warm-white light with favorable properties, such as a CRI > 90 for the Munsell saturated-red-color sample R9 (Fig. 9.9-1). However, the recent emergence of various proprietary phosphor blends that incorporate KSF offers an updated palette with outstanding properties.

Specific phosphor blends are usually developed for particular purposes; the constituents and relative proportions of the blend are adjusted to tune the CCT and CRI. We provide three examples for illustration: 1) optimization of the PCLED luminous flux and efficacy for illumination applications (Example 10.4-1); 2) enhancement of the color gamut for display applications (Example 10.4-2); and 3) generation of equal-energy white light for measurement applications (Example 10.4-3). Although the spectral densities for these three examples differ dramatically (Figs. 10.4-2–10.4-4), all appear white to the eye.

EXAMPLE 10.4-1. ***Proprietary Phosphor Blends That Incorporate Narrowband KSF.*** Different manufacturers employ different phosphor formulations, preparation techniques, and blending methods, which are largely proprietary. Nevertheless, when a red phosphor is called for in a blend, manganese-doped potassium fluorosilicate (KSF) is often chosen because of its favorable properties. In this example, we compare the spectral densities for Cree and Philips KSF-containing phosphor blends used to generate metameric warm-white light at 3000 K.

When pumped by blue LED light, the 2023 Cree XHP35.2 Pro9™ 3000-K phosphor blend generates the spectral density portrayed as the green curve in Fig. 10.4-2. The peak near 450 nm derives from the blue InGaN LED pump while the quintet of spectral spikes in the vicinity of 623 nm is a hallmark of the KSF spectrum, as explained in Sec. 10.3. (This curve was also displayed in Fig. 10.4-1, again in green, where it was juxtaposed with the spectral density for Cree's 2023 XHP35.2 conventional 3000-K phosphor blend (red curve) that incorporates the broadband red phosphor YAG:Eu^{3+}.)

The spectral density for the 2023 Philips 3000-K phosphor blend, when pumped by blue LED light, is plotted as the purple curve in Fig. 10.4-2. This blend is used in Philips' line of ultra-efficient white LED-filament retrofit lamps, such as that displayed in Fig. 11.4-1(*d*) and discussed in Example 11.4-3. The Philips and Cree spectral densities are quite similar: they are anchored by solitary peaks near 450 nm, associated with the blue InGaN LED, and are essentially congruent in the vicinity of 623 nm, where they mimic the spectrum of GE's TriGain™ formulation of KSF. The curves differ only in some details in the yellow and green spectral regions.

As detailed in Table 10.6-1 and Sec. 10.6, the use of KSF rather than YAG:Eu^{3+} as the red component in the phosphor blend endows PCLEDs with larger values of the luminous flux P_V, wall-plug luminous efficacy η_{WPE}, and wall-plug luminous efficiency η_{WPC}.

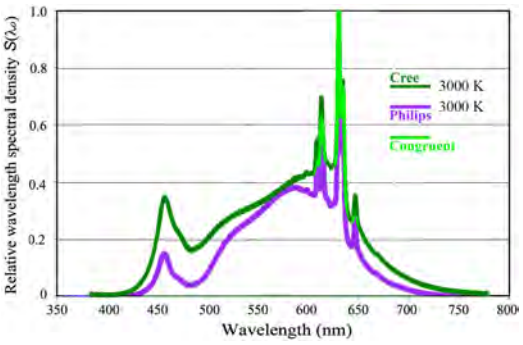


Figure 10.4-2 Spectral densities for 3000 K white light generated by KSF-containing phosphor blends (Cree: green curve; Philips: purple curve; Congruent: lime). These spectra differ substantially from those of conventional phosphors, providing enhanced luminous flux and wall-plug luminous efficacy. The peaks near 450 nm derive from the blue InGaN LED light. (Data adapted from Cree data sheet CLD-DS334-REV1 for XLamp® XHP35.2 Pro9™, 2023; and Philips data sheet MAS LEDBulbND4-60W E27 830 A60 CL G EELA, 2023.)

EXAMPLE 10.4-2. ***Blends Incorporating KSF and Narrowband Green Phosphors.*** Phosphor blends that combine narrowband K_2SiF_6:Mn^{4+} with a narrowband green phosphor, such as β-SiAlON:Eu^{2+} or one of GE's proprietary narrowband green phosphors, serve to increase the color rendering index and color gamut.[†] The increased CRI provides a more accurate, richer, and more intense rendering of the natural colors of illuminated objects, and the wider color gamut is useful for fashioning improved LED backlighting for liquid-crystal displays (LCDs).

The advantage in the display domain arises from the smaller photoluminescence bandwidths associated with the narrowband phosphors, which reduce deleterious crosstalk between the red and green subpixels and thereby improve image quality. As an example, the GE green phosphor illustrated in Fig. 10.4-3 (green curve) generates photoluminescence of shorter wavelength, narrower spectral width, and a more symmetric spectral profile, than does β-SiAlON:Eu^{2+} (purple curve). As a result, blending KSF with this green phosphor, rather than with β-SiAlON:Eu^{2+}, yields a color gamut that is enlarged by $> 5\%$, as well as higher screen brightness.

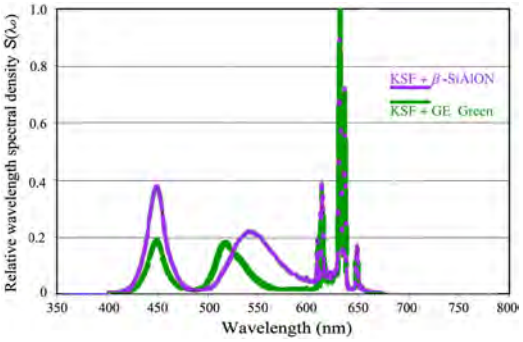


Figure 10.4-3 KSF+β-SiAlON (purple) and KSF+GE green (green) narrowband phosphor blends that generate metameric white light with $T_c \approx 6500$ K. Individual spectral peaks appear at red, green, and blue wavelengths (the latter from the LED). Narrowband phosphors lead to an enlarged color gamut. (Data adapted from S. J. Camardello, et al., Development of New Green Phosphors for Liquid Crystal Display Backlights, *Society for Information Display (SID) Digest*, vol. 52, paper 62-11, pp. 917–919, 2021.)

EXAMPLE 10.4-3. ***Blends That Generate Approximately Equal-Energy White Light.*** The properties of equal-energy white light, with its uniform wavelength-based spectral density (as opposed to metameric white light), are well-understood (Secs. 9.3 and 9.6; Examples 9.6-3 and 9.9-1). Such light has a correlated color temperature $T_c \approx 5460$ K, as established in Example 9.8-2(a). A source whose spectral density more-or-less approximates that of equal-energy white light from the NUV to the NIR can be fabricated by making use of a single NUV GaN LED die whose encapsulant is infused with a blend of phosphors of different colors, and making use of a UV-blocking filter to

[†] L. Wang, X. Wang, T. Kohsei, K. i. Yoshimura, M. Izumi, N. Hirosaki, and R.-J. Xie, Highly Efficient Narrow-Band Green and Red Phosphors Enabling Wider Color-Gamut LED Backlight for More Brilliant Displays, *Optics Express*, vol. 23, pp. 28707–28717, 2015; S. J. Camardello, M. D. Butts, R. Cassidy, J. E. Murphy, G. Parthasarathy, A. A. Setlur, O. P. Siclovan, J. Welch, and A. Yakimov, Development of New Green Phosphors for Liquid Crystal Display Backlights, *Society for Information Display (SID) Digest*, vol. 52, no. 1, book 2, paper 62-11, pp. 917–919, 2021.

attenuate the residual spectral spike from the GaN pump at $\overline{\lambda} \approx \lambda_g \approx 366$ nm. The spectral density from one such device is illustrated in Fig. 10.4-4. Light sources with equal-energy spectra find use in colorimetry, materials characterization, and spectroscopy.

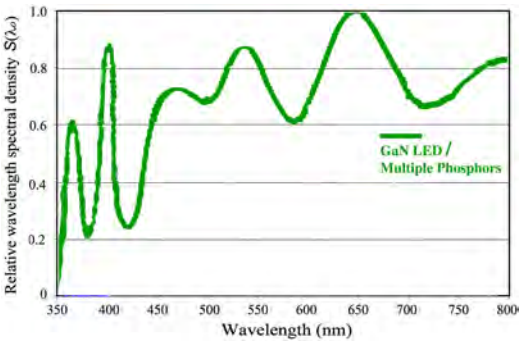


Figure 10.4-4 Spectrum for a PCLED that generates a very rough approximation to equal-energy white light. The device comprises a single NUV GaN LED die under an encapsulant infused with a blend of multiple phosphors, and the light is passed through a Schott GG395 UV-blocking filter. Spectral peaks of the constituent phosphors are evident at several wavelengths; the peak at 366 nm is the residual GaN LED pump light. (Data adapted from Lumedel specification sheet for NewDEL™ Model X3312 Fiber-Coupled Broadband LED Source, 2023.)

Quantum-Dot Photoluminescence. Just as inorganic phosphors generate photoluminescence, so too do chalcogenide and perovskite quantum dots (Fig. 5.8-1). [Quantum dots that emit photoluminescence (Sec. 5.8) are to be distinguished from quantum-dot light-emitting diodes (QLEDs) that emit electroluminescence (Secs. 7.5–7.7).] Although the use of quantum dots to generate photoluminescence has several advantages relative to the use of narrowband inorganic phosphors, in the current state of our technology the disadvantages of using quantum dots outweigh the advantages:

Advantages of Quantum Dots Over Narrowband Phosphors.

- Ideal absorption in the blue and a photoluminescence quantum yield $\eta_{\text{PLQY}} \approx 1$.
- Finely controllable photoluminescence wavelength, tunable by quantum-dot size.
- Range of quantum-dot sizes determines concomitant emission-wavelength range.

Advantages of Narrowband Phosphors Over Quantum Dots.

- Narrower spectra, wider color gamuts, and brighter operation.
- Smaller self-absorption.
- Superior chemical, thermal, high-flux, and humidity stability.
- Longer lifespan and greater reliability; no degradation under on-chip conditions.
- Less complex manufacturing and lower fabrication costs.
- Compatibility with existing LED designs.
- Superior environmental friendliness.

Examples of Quantum-Dot/Phosphor Hybrids.

- A narrowband KSF red phosphor and green quantum dots can be combined in a photoluminescent film that exhibits behavior similar to that of a film of KSF and narrowband inorganic green phosphor. Here, the tunability of quantum dots is joined with the stability of inorganic phosphors.
- A KSF phosphor can be coated on an InGaN blue LED die to create a magenta source that is then used in conjunction with a green-emitting perovskite quantum-dot film. Green perovskite quantum dots have bandwidths $\Delta\lambda_{\text{FWHM}} < 25$ nm, which are narrower than those of green chalcogenide (CdSe) quantum dots.

Despite the wide range of existing photoluminescent materials, the search for, and development of, phosphors with superior properties continues apace.

10.5 DISCRETE COOL-WHITE PCLEDS

As discussed in the introduction to this Chapter, a blue-emitting InGaN LED chip can be used in conjunction with a yellow phosphor to generate metameric white light in a discrete **phosphor-conversion LED (PCLED)** (also called a **phosphor-conversion package**). The mixture of blue and yellow can give rise to white because yellow is itself a combination of red and green (see Sec. 9.3 and particularly Example 9.3-1). The phosphor may be directly coated on the LED die or can take the form of a sheet that overlays the chip. Alternatively, the phosphor can be dispersed within the transparent material that encapsulates the die or it can be remotely located at some distance from the chip — each configuration has its own advantages and uses. A fraction of the blue LED photons that impinge on the phosphor generate yellow photons via photoluminescence. As a result of its salutary properties, cerium-doped yttrium aluminum garnet (Sec.10.3 and Table 10.3-1) is often the phosphor of choice for fabricating discrete cool-white PCLEDs, as well as cool-white PCLED filaments (Sec. 10.7) and cool-white COB PCLEDs (Sec. 10.8).

Evolution of the Discrete White PCLED

The evolution of the discrete white PCLED over the twenty-year period from 2004 to 2024 is illustrated in Fig. 10.5-1. Early devices, such as that portrayed in Fig. 10.5-1(a), made use of InGaN LED chips with a peak wavelength $\lambda_p \approx 465$ nm and $\Delta\lambda_{FWHM} \approx 35$ nm. A fraction of the blue LED photons that impinged on the YAG:Ce^{3+} phosphor generated yellow photoluminescence with a broad spectral bandwidth ($\Delta\lambda_{FWHM} \approx 200$ nm). The result was metameric white light with a wall-plug luminous efficacy $\eta_{WPE} \approx 20$ lm/W. Devices in dual in-line (DIP) packages such as this are designed to be soldered onto a printed-circuit board through holes in the board. DIP LEDs are therefore a type of **through-hole LED**.

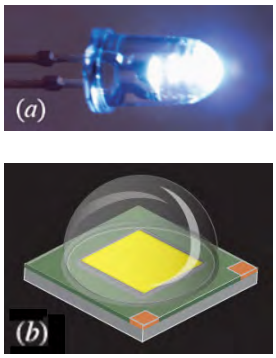


Figure 10.5-1 Evolution of the discrete white phosphor-conversion LED. (a) White-light emission from an early device (ca. 2004) consisting of an InGaN LED die and a yellow phosphor in a 5-mm-diameter, dual in-line package (DIP). This device generated metameric cool-white light with a wall-plug luminous efficacy $\eta_{WPE} \approx 20$ lm/W. (b) Contemporary device (ca. 2024) comprising a single InGaN LED die overlaid with a thin yellow phosphor sheet and a 3-mm-diameter hemispherical lens in a surface-mounted device (SMD) package. Devices such as these provide metameric cool-white light that in practice can attain $P_V > 1000$ lm and $\eta_{WPE} > 200$ lm/W, a factor of 10 larger than that of early devices such as the one portrayed in (a).

A contemporary discrete white PCLED, such as that portrayed in Fig. 10.5-1(b), operates on the same principle, but is packaged as a **surface-mounted device (SMD)**, with electrical contacts lateral to the housing; this offers improved heat-sinking and efficiency along with reduced size. The single LED die in this illustration is supported by a ceramic base and is overlaid with a thin yellow phosphor sheet. The entire device is encapsulated in a hemispherical silicone lens. The viewing angle and operating temperature are similar to those for the single-color MQWLEDs characterized in Table 7.4-1.

The InGaN LED chips used in modern devices generally have shorter peak wavelength ($\lambda_p \approx 450$–460 nm) and narrower spectral width ($\Delta\lambda_{FWHM} \approx 20$ nm) than those employed in first-generation devices. The yellow photoluminescence typically has an average wavelength $\overline{\lambda} \approx 570$ nm and a spectral band that stretches from ≈ 510

to 630 nm, corresponding to $\Delta\lambda_{\text{FWHM}} \approx 120$ nm. This is narrower than that of first-generation devices but is still broadband. Discrete cool-white PCLEDs such as these are widely used in outdoor, roadway, spot, and high-bay lighting applications.

The border between blue and violet is usually defined at $\lambda_0 = 445$ nm (Fig. 2.4-1), but wavelengths in this range are usually referred to as* blue *or* royal blue *in the LED literature, and we adhere to this convention.

Behavior and Characteristics of a Discrete Cool-White PCLED

We now proceed to examine the behavior of InGaN-based discrete cool-white PCLEDs by focusing on the operating parameters for a representative device with the specifications detailed in Table 10.5-1. We do so via a sequence of five examples:

- Example 10.5-1: Spectrum and chromaticity diagram.
- Example 10.5-2: Wavelength conversion efficiency.
- Example 10.5-3: Chromaticity coordinates.
- Example 10.5-4: Correlated color temperature.
- Example 10.5-5: Wall-plug luminous efficacy.

Table 10.5-1 Specifications for a representative discrete (single-die), cool-white, 3×3 mm^2 InGaN PCLED, packaged as a ceramic surface-mounted device (SMD) with a 3.45 mm $\times$ 3.45 mm footprint. Data are presented for operation at a typical current (top row) and at the maximum operating current (bottom row). Successive columns display: current i, forward voltage V, electrical power consumption P_{EL}, luminous flux P_{V}, wall-plug luminous efficacy (WPE), wall-plug luminous efficiency (WPC), chromaticity coordinates (x, y), correlated color temperature T_c, and color rendering index (CRI). The data displayed were collected at an operating temperature of 85 °C and at a viewing (50%-power) angle $2\theta_{1/2} \approx 125°$. (Data adapted from Cree data sheet CLD-DS149-REV6 for XLamp® XP-L2, `https://downloads.cree-led.com/files/ds/x/XLamp-XPL2.pdf`, 2023.)

COOL-WHITE[a]	i^b(A)	V^b(V)	$P_{\text{EL}}^{b,c}$(W)	P_{V}^{c}(lm)	$\eta_{\text{WPE}}^{c,d}(\frac{\text{lm}}{\text{W}})$	η_{WPC}^{d}	x^e	y^e	T_c^f(K)	CRI[g]
TYPICAL	1.05	2.79	2.93	490	167	0.24	0.34	0.36	5000	80
MAXIMUM	3.00	3.04	9.12	1150	126	0.18	0.34	0.35	5000	80

[a] Table entry values are rounded.
[b] The electrical drive power is related to the device current and voltage via $P_{\text{EL}} = iV$, as specified in (7.1-13).
[c] The wall-plug luminous efficacy η_{WPE}, luminous flux P_{V}, and electrical drive power P_{EL} are related by (8.9-4).
[d] The wall-plug luminous efficiency and efficacy are related by $\eta_{\text{WPC}} = \eta_{\text{WPE}}/683$, in accordance with (8.9-9).
[e] The chromaticity coordinates x and y, which are perceptual measures of color, are defined in Sec. 9.6.
[f] The correlated color temperature T_c, a measure of the color of a source of light, is defined in Sec. 9.8.
[g] The color rendering index is a measure of the faithfulness with which the color of an object is rendered (Sec. 9.9).

EXAMPLE 10.5-1. ***Spectrum and Chromaticity Diagram for a Cool-White PCLED.*** The spectrum and chromaticity diagram associated with the metameric cool-white light emitted by a single-die, discrete device such as the one characterized in Table 10.5-1 are displayed in Figs. 10.5-2(a) and (b), respectively. The spectral density presented in Fig. 10.5-2(a) comprises two peaks: a narrow peak ($\Delta\lambda_{\text{FWHM}} \approx 20$ nm) associated with the blue LED light centered at $\lambda_1 = 445$ nm and a broad peak ($\Delta\lambda_{\text{FWHM}} \approx 120$ nm) associated with the yellow photoluminescence with average wavelength $\lambda_2 = 570$ nm. This spectrum is distinctly different from that for a white incandescent source: the spectral density traced in Fig. 9.7-1(a) for a blackbody radiator is a smooth curve that stretches from ultraviolet to infrared wavelengths, and has a single peak.

The rationale underlying the generation of metameric cool-white light by this spectral combination is provided by the chromaticity diagram displayed in Fig. 10.5-2(b). As discussed in connection with

Figs. 9.6-2 and 9.7-2, the wavelengths (specified in nm) at the outer curved boundary of these diagrams represent the collection of fully saturated spectral colors. Also, as explained in Example 9.6-1, all colors that lie on a straight line connecting any two points in the diagram can be generated by mixing the colors at the endpoints of that line. Since a straight line can be drawn connecting the blue (445 nm) and yellow (570 nm) endpoints, and since that line transects the Planckian locus in the vicinity of 5000–7000 K, it follows that metameric cool-white light can be generated by combining blue LED light and yellow photoluminescence. The proportion of blue and yellow light required to do so will be considered in Example 10.5-3.

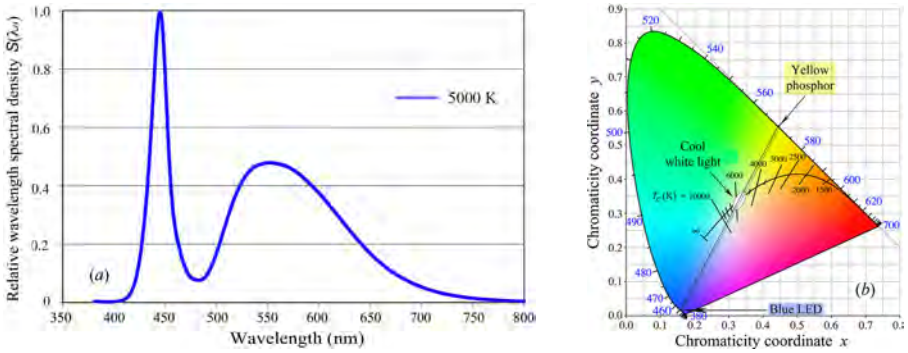


Figure 10.5-2 (a) Spectrum emitted by a discrete cool-white phosphor-conversion LED comprising an InGaN die overlaid with a thin yellow phosphor layer, such as the one specified in Table 10.5-1. The average wavelengths emitted by the die and the phosphor are, respectively, $\lambda_1 = 445$ nm (blue) and $\lambda_2 = 570$ nm (yellow). (b) The blue and yellow points on the outer boundary of the chromaticity diagram can be connected by a straight line that transects the Planckian locus in the vicinity of 5000–7000 K, corresponding to metameric cool-white light.

EXAMPLE 10.5-2. ***Wavelength Conversion Efficiency for a Cool-White PCLED.*** The operation of the PCLED described in Example 10.5-1 is based upon a process in which a fraction of the photons emitted by a blue LED die are converted into yellow photoluminescence photons in a phosphor. This process involves several steps:

Phosphor-Absorption Fraction. The fraction of the blue photons emitted by the die that are absorbed by the yellow phosphor is denoted f, while the unabsorbed fraction that exits the device is $(1 - f)$. The value of f depends on a number of variables, including the nature of the phosphor, its density, and its location relative to the LED die, all of which make it tricky to estimate.

Photoluminescence Quantum Yield (PLQY). Not every blue photon absorbed by the phosphor succeeds in eliciting a yellow photoluminescence photon; some blue photons are lost via nonradiative transitions following absorption. The fraction of absorbed photons that effectively generate yellow photons is specified by the PLQY (Sec. 10.2), which, for the particular YAG:Ce^{3+} yellow phosphor used in the suite of examples considered in this section is $\eta_{\text{PLQY}} \approx 0.80$. More typically, the photoluminescence quantum yield for YAG:Ce^{3+} is $\eta_{\text{PLQY}} \approx 0.90$, as reported in Table 10.3-1.

Complementary Photoluminescence Quantum Defect. Also, as discussed in Sec. 10.2, every blue photon that is successfully converted into a yellow photon loses a portion of its energy in the course of frequency downconversion. The fraction of the energy $h\nu_1$ of a blue photon effective in generating a lower-energy yellow photon of average energy $h\nu_2$ is quantified by the complementary quantum defect $\overline{\eta}_{\text{PLQD}} = h\nu_2/h\nu_1 = \lambda_1/\lambda_2$ defined in (10.2-3). Hence, for a blue photon of average wavelength $\lambda_1 = 445$ nm that gives rise to a yellow photoluminescence photon of average wavelength $\lambda_2 = 570$ nm, the complementary quantum defect is $\overline{\eta}_{\text{PLQD}} \approx \lambda_1/\lambda_2 = {}^{445}/_{570} \approx 0.78$.

The considerations outlined above lead us to conclude that, in the presence of a yellow phosphor, the initial blue LED light of radiant flux P_0 is converted into a combination of both blue and yellow radiant flux, with optical powers given by $(1 - f)P_0$ and $f\eta_{\text{PLQY}}\overline{\eta}_{\text{PLQD}}P_0$, respectively. This outcome can be cast in the form of an idealized spectral density of the emitted light written as

$$S_\lambda(\lambda_0) \approx P_0 \left[(1 - f)\,\delta(\lambda - \lambda_1) + f\eta_{\text{PLQY}}\overline{\eta}_{\text{PLQD}}\,\delta(\lambda - \lambda_2)\right], \tag{10.5-1}$$

where it is assumed that the optical power associated with each color is concentrated in a delta function at its respective average wavelength.

In practice, the yellow component has a substantial linewidth, as is understood from the discussion provided in Sec. 10.3. The power spectral density displayed in Fig. 10.5-2(a) demonstrates that the yellow peak is of lower height, but of greater width, than the blue peak. Numerical estimation of the areas under these two peaks, which represent their respective external optical powers, reveals that the area under the yellow peak is a factor $\mathcal{A} \approx 2.5$ greater than that under the blue peak, which indicates that the majority of the optical power generated by the blue LED is transferred to yellow light via photoluminescence. Having established the emitted yellow-to-blue optical-power ratio $\mathcal{A}$, it is straightforward to compute the phosphor-absorption fraction f specified in (10.5-1):

$$\mathcal{A} \approx \frac{f\eta_{\text{PLQY}}\overline{\eta}_{\text{PLQD}}P_0}{(1-f)P_0} \quad \text{which yields} \quad f \approx \frac{\mathcal{A}}{\mathcal{A}+\eta_{\text{PLQY}}\overline{\eta}_{\text{PLQD}}}. \tag{10.5-2}$$

Inserting the numerical estimates obtained above into (10.5-2), namely $\mathcal{A} \approx 2.5$, $\eta_{\text{PLQY}} \approx 0.80$, and $\overline{\eta}_{\text{PLQD}} \approx 0.78$, yields $f \approx 0.80$, whereupon (10.5-1) can be written as

$$S_\lambda(\lambda_0) \approx 0.20\,P_0\,\delta(\lambda-\lambda_1) + 0.50\,P_0\,\delta(\lambda-\lambda_2). \tag{10.5-3}$$

EXAMPLE 10.5-3. ***Chromaticity Coordinates for a Cool-White PCLED.*** The determination of the chromaticity coordinates for a mixture of light of two colors whose individual chromaticity coordinates are known was described in Example 9.6-1. This technique can be used to estimate the chromaticity coordinates for the white PCLED under consideration. This device contains a blue LED die and a yellow phosphor, which are complementary colors (Example 9.3-1), thereby allowing metameric white light to be generated when the two components are mixed in suitable proportions. The idealized spectral densities set forth in (10.5-1) and (10.5-3) mimic the form of the spectral density displayed in (9.6-1), with the equivalences $P_1 \equiv (1-f)P_0 \approx 0.20P_0$ and $P_2 \equiv f\eta_{\text{PLQY}}\overline{\eta}_{\text{PLQD}}P_0 \approx 0.50P_0$. Employing the idealized spectral density set forth in (10.5-3) accommodates the requirement set forth in Example 9.6-1 that the spectral widths of the blue and yellow light be small in comparison with the bandwidths of the color-matching functions. (The use of this idealization will be justified at the end of this example.)

As established in Example 9.6-1 and specified in (9.6-4), the chromaticity coordinates of the combined light (x, y), are linear combinations of the individual chromaticity coordinates for the blue and yellow components, (x_1, y_1) and (x_2, y_2), respectively, suitably weighted by the functions K_1 and K_2. These weight functions, which are provided in (9.6-3), are in turn determined by the relative optical powers and values of the color matching functions $\bar{x}(\lambda_{1,2})$, $\bar{y}(\lambda_{1,2})$, and $\bar{z}(\lambda_{1,2})$. The parameters required for computing (x, y) are presented in Table 10.5-2.

Table 10.5-2 Parameters employed in the course of estimating the chromaticity coordinates (x, y) for the light emitted by a discrete cool-white PCLED that makes use of a blue die (source 1) in conjunction with a yellow phosphor (source 2). Successive columns of the table display: the average wavelengths of the blue (λ_1) and yellow (λ_2) sources (in nm); the individual chromaticity coordinates for the two sources, $(x_{1,2}, y_{1,2})$, under the assumption that they are spectrally pure; and the CIE 1931 color matching functions $\bar{x}\bar{y}\bar{z}$, evaluated at the average wavelengths of the two sources. The values of the photopic luminous efficiency function at the average wavelengths of the two sources, $V(\lambda_{1,2})$, will be used in estimating the wall-plug luminous efficacy in Example 10.5-5. All of the values provided here are publicly available by consulting tables or online conversion calculators for: 1) converting a specified wavelength to chromaticity coordinates, 2) determining the values of the CIE 1931 color matching functions at a specified wavelength, and 3) determining the value of the photopic luminous efficiency function at a specified wavelength.

SOURCE	$\lambda_{1,2}$	$x_{1,2}$	$y_{1,2}$	$\bar{x}(\lambda_{1,2})$	$\bar{y}(\lambda_{1,2})$	$\bar{z}(\lambda_{1,2})$	$V(\lambda_{1,2})$
BLUE (1)	445	0.1611	0.0138	0.3481	0.0298	1.7826	0.0574
YELLOW (2)	570	0.4441	0.5547	0.7621	0.9520	0.0021	0.9733

Using (9.6-3), in conjunction with (10.5-1), and the values of $\bar{x}$, $\bar{y}$, and $\bar{z}$ set forth in Table 10.5-2, leads to $K_1 = 2.1605P_1 = 0.4315P_0$ and $K_2 = 1.7162P_2 = 0.8570P_0$. Inserting these weights into

(9.6-4), together with the individual chromaticity coordinates for the blue and yellow light provided in Table 10.5-2, yields the chromaticity coordinates for the combined light: $x = 0.3493$ and $y = 0.3735$. These values are close to those of their experimental counterparts, which are provided in Table 10.5-1: $x = 0.34$ and $y = 0.36$. The theoretical values depend strongly on $\mathcal{A}$ (and therefore on f), which reflects the important role played by the relative contributions of the values of the yellow and blue optical powers; yet they depend only weakly on η_{PLQY} and $\overline{\eta}_{\mathrm{PLQD}}$. In accordance with (9.6-4), the chromaticity coordinates for the combined light (x, y) must fall along the straight line that connects the coordinates of the individual sources; examination of Fig. 10.5-2(b) confirms that indeed they do.

A more accurate determination of the chromaticity coordinates can be carried out by replacing the idealized spectral density presented in (10.5-3) with the experimental one, then using (9.5-1) to calculate the XYZ tristimulus values, and finally using (9.5-6a) and (9.5-6b) to compute the chromaticity coordinates. This additional effort may not be necessary, however, since the boundary of the chromaticity diagram in the yellow spectral region, portrayed in Fig. 10.5-2(b), is essentially a straight line. Hence, as explained in Sec. 9.6, the assembly of yellow spectral components that comprise the broadband photoluminescence spectrum all lie along this same straight boundary, and therefore so too do the chromaticity coordinates of mixtures of all pairs of these spectral components. This indicates that the simplified approach of representing the yellow photoluminescence power as a delta function localized at its average wavelength, as posited in (10.5-3), will not lead us far astray.

EXAMPLE 10.5-4. ***Correlated Color Temperature for a Cool-White PCLED.*** As discussed in Example 9.8-2, a straightforward procedure exists for calculating the correlated color temperature (CCT) of a source of white light from its xy chromaticity coordinates. For the discrete cool-white PCLED considered in the foregoing examples, inserting the coordinates $(x, y) = (0.3493, 0.3735)$ established in Example 10.5-3 into (9.8-3b) yields McCamy's intermediate parameter $\zeta = 0.0923$, which, when entered into (9.8-3a), returns $T_c \approx 4920$ K. This value is in good agreement with the experimental value reported in Table 10.5-1, which is $T_c \approx 5000$ K.

This value is not far from that for equal-energy (spectrally uniform) white light considered in Example 9.8-2(a), which is $T_c \approx 5460$ K. It is worth reemphasizing that the quality of the white light generated by a cool-white phosphor-conversion LED, such as the one at hand, and that generated by a spectrally uniform source of white light, such as that considered in Example 9.8-2(a), are nearly indistinguishable perceptually, despite the dramatic differences in their spectral densities.

EXAMPLE 10.5-5. ***Wall-Plug Luminous Efficacy for a Cool-White PCLED.*** The wall-plug luminous efficacy η_{WPE} for the cool-white phosphor-conversion LED considered in this suite of examples is determined by making use of (8.9-7), which specifies that $\eta_{\mathrm{WPE}} \approx 683\,\eta_{\mathrm{PCE}} V(\lambda_0)$ for a monochromatic source of light, where η_{PCE} is the LED power-conversion efficiency and $V(\lambda_0)$ is the photopic luminous efficiency function at the wavelength λ_0. The value of η_{WPE} may be approximated by assuming that the individual blue and yellow contributions, from the LED and from the phosphor, respectively, are monochromatic, and by making use of the wavelength conversion efficiency parameters established in Example 10.5-2. The requisite generalization of (8.9-7) then takes the form

$$\eta_{\mathrm{WPE}} \approx 683\,\eta_{\mathrm{PCE}} \left[(1 - f)\,V(\lambda_1) + f\eta_{\mathrm{PLQY}}\overline{\eta}_{\mathrm{PLQD}} V(\lambda_2)\right]. \tag{10.5-4}$$

We now make use of the parameter values set forth in Example 10.5-2, namely, $(1 - f) \approx 0.20$ and $f\eta_{\mathrm{PLQY}}\overline{\eta}_{\mathrm{PLQD}} \approx 0.50$, and assume further that the power-conversion efficiency for the blue LED assumes the plausible value $\eta_{\mathrm{PCE}} \approx 1/2$, as inferred from (7.1-15). Inserting these values, together with those for the photopic luminous efficiency functions drawn from Table 10.5-2 [$V(\lambda_1) = 0.0574$ and $V(\lambda_2) = 0.9733$] into (10.5-4) provides $\eta_{\mathrm{WPE}} \approx 1/2 \cdot 683\,[0.20 \cdot 0.0574 + 0.50 \cdot 0.9733] \approx 170$ lm/W. This value is close to the experimental result reported in Table 10.5-1, which is 167 lm/W. The conversion from wall-plug luminous efficacy to wall-plug luminous efficiency, via (8.9-9), yields $\eta_{\mathrm{WPC}} \approx 170/683 \approx 0.25$, which is close to 0.24, the experimental value provided in Table 10.5-1.

The value for the wall-plug luminous efficacy depends strongly on η_{PLQY} and $\overline{\eta}_{\mathrm{PLQD}}$, and only weakly on f, whereas the reverse is true for the chromaticity coordinates (Example 10.5-3). The underlying reason for this dichotomy is inherent in the form of (10.5-4): η_{PLQY} and $\overline{\eta}_{\mathrm{PLQD}}$ govern the yellow contribution to the luminous efficacy, which is far more potent than the blue contribution since $V(\lambda_2) \gg V(\lambda_1)$.

Consistency of Model and Experimental Parameters for a Cool-White PCLED. The unique, common set of model parameters associated with the discrete cool-white PCLED examined in Examples 10.5-1–10.5-5 are displayed in Table 10.5-3 (top row). The model values for the chromaticity coordinates, correlated color temperature, wall-plug luminous efficacy, and wall-plug luminous efficiency are all in good agreement with the experimental values reported in Table 10.5-1. This indicates that the modeling is internal consistent and confirms that our understanding of the relevant aspects of LED operation (Chapter 7), color vision (Chapter 8), colorimetry (Chapter 9), and phosphor photoluminescence (Secs. 10.2–10.4) are in accord with the operation of these devices.

Table 10.5-3 The top row of the table provides a summary of the model parameters used in, and obtained from, the calculations pertinent to the discrete cool-white PCLED considered in Examples 10.5-1–10.5-5. Successive columns of the table display: the ratio of yellow-to-blue optical power (radiant flux) $\mathcal{A}$; photoluminescence quantum yield $\eta_{\mathtt{PLQY}}$; complementary photoluminescence quantum defect $\overline{\eta}_{\mathtt{PLQD}}$; phosphor-absorption fraction f; chromaticity-coordinate weight functions $K_{1,2}$; chromaticity coordinates (x, y); correlated color temperature T_c (K); estimated power-conversion efficiency $\eta_{\mathtt{PCE}}$; wall-plug luminous efficacy $\eta_{\mathtt{WPE}}$ (lm/W); and wall-plug luminous efficiency $\eta_{\mathtt{WPC}}$. The model parameters are in good accord with the experimental parameters reported in Table 10.5-1. The bottom row of the table displays the model parameters when the ratio of yellow-to-blue optical power is increased to $\mathcal{A} = 10$.

$\mathcal{A}$	$\eta_{\mathtt{PLQY}}$	$\overline{\eta}_{\mathtt{PLQD}}$	f	K_1	K_2	x	y	T_c	$\eta_{\mathtt{PCE}}$	$\eta_{\mathtt{WPE}}$	$\eta_{\mathtt{WPC}}$
2.5	0.80	0.78	0.80	0.4315	0.8570	0.3493	0.3735	4920	0.5	170	0.25
10	0.80	0.78	0.94	0.1269	1.0080	0.4125	0.4942	3972	0.5	197	0.29

Impediments Attendant to Reducing the CCT of a Cool-White PCLED. The CCT of the light emitted from a cool-white PCLED can be reduced by increasing the yellow-to-blue optical-power ratio $\mathcal{A}$; this can be implemented, for example, by increasing the density of yellow phosphor in the device. The bottom row of Table 10.5-3 reveals, for example, that increasing $\mathcal{A}$ from 2.5 to 10 results in a reduction of the CCT by about 1000 K (from 4920 to 3972 K). The calculations were carried out by fixing the blue and yellow wavelengths (so that all entries in Table 10.5-2 remain the same), the power-conversion efficiency $\eta_{\mathtt{PCE}}$, and the phosphor photoluminescence parameters $\eta_{\mathtt{PLQY}}$ and $\overline{\eta}_{\mathtt{PLQD}}$. As expected, the increase in the value of $\mathcal{A}$ is accompanied by an increase in the phosphor-absorption fraction f (from 0.80 to 0.94), as provided by (10.5-2), and by an increase in the wall-plug luminous efficacy $\eta_{\mathtt{WPE}}$ (from 170 to 197 lm/W), as specified in (10.5-4).

Although increasing $\mathcal{A}$ reduces T_c, it also moves the chromaticity coordinates (x, y) away from the Planckian locus, which renders the light increasingly yellowish and degrades the CRI. This is a consequence of the fact that the chromaticity coordinates of the generated light are constrained to lie along the line connecting the yellow and blue terminals on the diagram portrayed in Fig. 10.5-2(b). Indeed, since a value for the CCT can be calculated even when the chromaticity coordinates are somewhat remote from the Planckian locus, some illumination engineers decline to use it as a metric and rely instead on the CRI, even though the CRI has its own limitations (Sec. 9.9).

A practical and widely used method for reducing the CCT while enforcing a high value of the CRI relies on the introduction of a red phosphor, as described in the Sec. 10.6.

10.6 DISCRETE WARM-WHITE PCLEDS

In this section we demonstrate that blended phosphors, such as those introduced in Sec. 10.4, can be used to generate warm-white light with high values of the CRI. In particular, we show that introducing a red phosphor into the blend allows the CCT of the emitted light to be decreased and its CRI to be increased, thereby making the light warmer and improving its color rendering qualities. When a conventional (broadband) red phosphor is used, these benefits are accompanied by a reduction in the luminous flux and wall-plug luminous efficacy but the use of a narrowband red phosphor mitigates this degradation. We compare the performance of discrete warm-white PCLEDs that rely on these two classes of phosphors, and explicitly demonstrate the benefits and limitations of each version. Devices such as these are widely used for general indoor and outdoor lighting.

Characteristics of a Warm-White PCLED (Broadband-Red Phosphor)

We first consider discrete warm-white PCLEDs that make use of conventional (broadband) phosphor technology. The specifications for a device with a CCT of 3000 K and a CRI of 90 are presented in the upper portion of Table 10.6-1; the associated spectrum and chromaticity diagram are displayed in Example 10.6-1. Devices based on phosphor blends that incorporate narrowband-red will be discussed subsequently.

Table 10.6-1 Specifications for representative single-die, warm-white, 3.2 mm × 3.2 mm, high-density InGaN PCLEDs packaged as ceramic surface-mounted devices (SMDs) with a 3.45 mm × 3.45 mm footprint. The upper and lower tables correspond to devices fabricated using phosphor blends that incorporate broadband-red (YAG:Eu^{3+}) and narrowband-red (KSF:Mn^{4+}) phosphors, respectively. Data are presented for operation at several values of the forward current. Successive columns display: forward current i, forward voltage V, electrical power consumption P_{EL}, luminous flux P_{V}, wall-plug luminous efficacy (WPE) $\eta_{\text{WPE}}(\frac{\text{lm}}{\text{W}})$, wall-plug luminous efficiency (WPC) η_{WPC}, chromaticity coordinates (x, y), correlated color temperature T_c, and color rendering index (CRI). The data displayed were collected at an operating temperature of 85 °C and with a viewing (50%-power) angle of $2\theta_{1/2} \approx 130$–$135°$. (Data adapted from Cree data sheets CLD-DS199-REV7 and CLD-DS334-REV1 for XLamp® XHP35.2 and XLamp® XHP35.2 Pro9™, respectively, at `https://downloads.cree-led.com/files/ds/x/XLamp-XHP35.2.pdf`, 2023 and `https://downloads.cree-led.com/files/ds/x/XLamp-XHP35.2-Pro9.pdf`, 2023.)

WARM-WHITE[a]	i^b(A)	V^b(V)	$P_{\text{EL}}^{b,c}$(W)	P_{V}^c(lm)	$\eta_{\text{WPE}}^{c,d}(\frac{\text{lm}}{\text{W}})$	η_{WPC}^d	x^e	y^e	T_c^f(K)	CRI[g]
BROADBAND-RED BLEND	0.35	11.2	3.92	460	117	0.17	0.43	0.40	3000	90
	0.70	11.9	8.33	820	98	0.14	0.43	0.40	3000	90
	1.50	13.1	19.7	1400	71	0.10	0.43	0.39	3000	90
NARROWBAND-RED BLEND	0.35	11.2	3.92	490	125	0.18	0.43	0.40	3000	90
	0.70	11.9	8.33	880	106	0.16	0.43	0.40	3000	90

[a] Table entry values are rounded.
[b] The electrical drive power is related to the device current and voltage via $P_{\text{EL}} = iV$, as specified in (7.1-13).
[c] The wall-plug luminous efficacy η_{WPE}, luminous flux P_{V}, and electrical drive power P_{EL} are related by (8.9-4).
[d] The wall-plug luminous efficiency and efficacy are related by $\eta_{\text{WPC}} = \eta_{\text{WPE}}/683$, in accordance with (8.9-9).
[e] The chromaticity coordinates x and y, which are perceptual measures of color, are defined in Sec. 9.6.
[f] The correlated color temperature T_c, a measure of the color of a source of light, is defined in Sec. 9.8.
[g] The color rendering index is a measure of the faithfulness with which the color of an object is rendered (Sec. 9.9).

EXAMPLE 10.6-1. ***Spectrum and Chromaticity Diagram for a Warm-White PCLED.*** A discrete white PCLED that is pumped by the light from a blue LED die and uses a phosphor blend that incorporates a broadband-red phosphor exhibits the spectrum and chromaticity diagram displayed in Fig. 10.6-1. The specifications for this device are displayed in the upper portion of Table 10.6-1. The spectral density for $T_c = 3000$ K, illustrated as the red curve in Fig. 10.6-1(a), has three principal features: a peak at 445 nm associated with the blue LED light, a point of inflection near 570 nm representing the presence of broadband yellow photoluminescence, and another peak in the vicinity of 615 nm corresponding to the red photoluminescence. This curve is identical to the one displayed in Fig. 10.4-1. The photoluminescence spectra of conventional phosphors have considerable bandwidths, as is understood from the discussion provided in Sec. 10.3.

As discussed in the text surrounding Fig. 9.6-2, and in Example 9.6-2, all colors lying within an arbitrary triangle traced out on the xy chromaticity diagram can be generated by mixing the three colors at the vertices of that triangle in appropriate proportions. Consider the triangle in Fig. 10.6-1(b) formed from the designated blue, yellow, and red vertices (associated with the LED light, the yellow photoluminescence, and the red photoluminescence, respectively). Since the triangle includes the region of the Planckian locus that encompasses 2700–3500 K, metameric warm-white light can be generated by using a blue LED die in conjunction with this yellow–red phosphor blend.

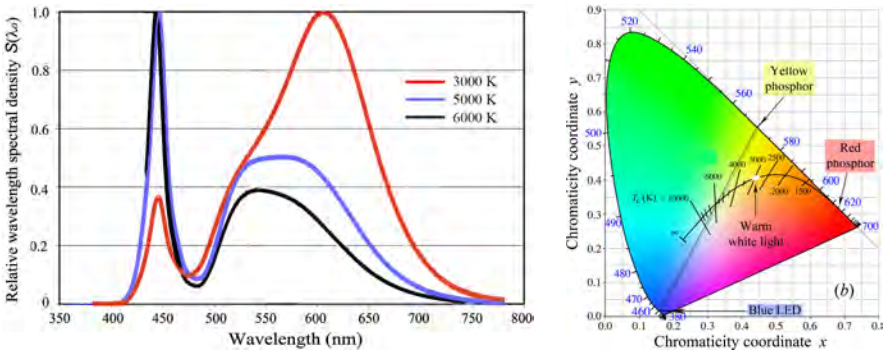


Figure 10.6-1 (a) Spectral density of the 3000-K warm-white emission from a discrete phosphor-conversion LED comprising an InGaN die in a package incorporating a phosphor blend that contains both yellow and red conventional phosphors (red curve). The specifications for the warm-white PCLED that generates this spectrum are displayed in the upper portion of Table 10.6-1. The peak wavelengths of the light emitted by the die and blended phosphor are, respectively, 445 nm (blue), 570 nm (yellow), and 615 nm (red). The 5000-K (blue) and 6000-K (black) curves, which represent cool-white light, are generated by yellow phosphors rather than by blends. (Data adapted from Cree data sheet CLD-DS199-REV7 for XLamp® XHP35.2, 2023.) (b) The blue, yellow, and red locations designated on the outer boundary of the chromaticity diagram form the vertices of a triangle that encompasses the Planckian locus over the range 2700–3500 K, which corresponds to warm white light.

Establishing the CCT. The CCTs and CRIs for PCLEDs that employ phosphor blends may be modified by adjusting the compositions and relative proportions of the various constituents that comprise the blend. The spectral densities represented by the 5000-K (blue) and 6000-K (black) curves in Fig. 10.6-1, both representing metameric cool-white light, result from the characteristics of the particular yellow phosphors used in the Cree XLamp® XHP35.2 family of PCLEDs. In particular, the 5000-K yellow phosphor is similar to, but distinct from, the 5000-K yellow phosphor that gives rise to the spectral density portrayed by the blue curve in Fig. 10.5-2.

Merits and Limitations of Incorporating a Red Phosphor. The principal merit of generating warm white light by incorporating a red phosphor in the blend is the ability to maintain a large value of the CRI, which endows illuminated objects with a more natural appearance by virtue of the increased spectral reach of the incident light. However, this salutary increase in the CRI is usually accompanied by an unwelcome decrease in the luminous flux P_{V}, wall-plug luminous efficacy η_{WPE}, and wall-plug luminous efficiency η_{WPC}, by virtue of both the reduced complementary photoluminescence quantum defect $\overline{\eta}_{\mathrm{PLQD}}$ (Sec. 10.2) and the reduced photopic luminous efficiency function $V(\lambda_0)$ in the red (Sec. 8.5). A hint of these tradeoffs emerges when the entries in the upper portion of Table 10.6-1 are compared with those in Table 10.5-1: the values of P_{V}, η_{WPE}, and η_{WPC} for a warm-white PCLED (CCT = 3000 K, CRI = 90) are noticeably smaller than those for a cool-white PCLED (CCT = 5000 K, CRI = 80). The comparison must be viewed with a note of caution, however, since the devices are from different PCLED families.

Modeling the Behavior of Blended-Phosphor Devices. In modeling the operation of devices that contain a blend of phosphors, the wavelength conversion efficiencies from both blue-to-yellow light and blue-to-red light can be accommodated by introducing a red-phosphor absorption fraction g alongside the yellow-phosphor absorption fraction f introduced in Example 10.5-2. This approach can then be employed to obtain an expression for the wall-plug luminous efficacy that is more general than, but analogous to, that considered in Example 10.5-5. The chromaticity coordinates can be established by applying the technique set forth in Example 9.6-2, which is a generalization of the method used in Examples 9.6-1 and 10.5-3. The correlated color temperature follows from the chromaticity coordinates, as explained in Example 10.5-4.

Characteristics of a Warm-White PCLED (Narrowband-Red Phosphor)

Having considered discrete warm-white PCLEDs that rely on broadband phosphor technology in the previous section, we turn now to comparing them with discrete warm-white PCLEDs that make use of narrowband phosphor technology. As discussed in Sec. 10.6, broadband-red phosphor blends offer high CRI, but simultaneously lead to a diminution of the device parameters P_{V}, η_{WPE}, and η_{WPC}, relative to the values for cool-white PCLEDs. As described in Sec. 10.4, the operation of narrowband red phosphors is expected to be more favorable in this respect, as a result of their smaller photoluminescence bandwidths.

The specifications for discrete warm-white PCLEDs that rely on narrowband- and broadband-red phosphor blends, but are otherwise identical, are presented in the lower and upper portions of Table 10.6-1, respectively. Both devices emit metameric white light with a CCT of 3000 K and a CRI of 90, and the light emitted by both is indistinguishable to the human eye. Furthermore, the current, voltage, electrical power consumption, and xy chromaticity coordinates of the two devices match exactly. Yet, their spectral densities, displayed as the green and red curves in Fig. 10.4-1, respectively, differ significantly, as do several other key measures. The relative performance advantages of these two types of warm-white PCLED are summarized below:

Performance Advantages of Discrete Narrowband-Red PCLEDs.

- Because of its smaller photoluminescence bandwidth, the luminous flux P_{V}, wall-plug luminous efficacy η_{WPE}, and wall-plug luminous efficiency η_{WPC} are 7–8% greater than values for the broadband-red warm-white PCLED, for the reasons explained in Sec. 10.4.
- The color gamut is enhanced relative to that of the broadband-red device (Sec. 9.5).

Performance Advantages of Discrete Broadband-Red PCLEDs.

- The heat retention in the broadband-red warm-white PCLED is lower than that in the narrowband-KSF device, for which the maximum allowed forward current is

limited because of the relatively long (8-ms) decay time of the parity- and spin-forbidden $^2E \rightarrow {}^4A_2$ emission transition.
- The lower heat retention in the broadband-red device supports a maximum luminous flux $P_V = 1400$ lm, whereas P_V is limited to 880 lm for the narrowband-KSF device.

Incorporating a narrowband, rather than broadband, red phosphor in the blend for a discrete warm-white PCLED improves the luminous flux and wall-plug luminous efficacy but it can also limit the maximum attainable luminous flux.

Variations on the Theme of Discrete Phosphor-Conversion Devices

Discrete warm- and cool-white PCLEDs exist in many configurations and can be used in many systems, as illustrated by the following examples:

- *Designer-Phosphor Devices.* Various versions of multiple-phosphor PCLEDs can be implemented by coupling blue, violet, or ultraviolet LEDs with designer collections of phosphors that exhibit photoluminescence at diverse wavelengths.
- *Lamps with Adjustable Color Temperature.* Retrofit lamps that make use of multiple cool-white (6200 K) and multiple warm-white (2200 K) discrete PCLEDs of adjustable intensities can be manually tuned to emit white light with a correlated color temperature that ranges over 2200–6200 K (Example 11.4-4).

10.7 PCLED FILAMENTS

An LED filament is a 1D chain of tens (or hundreds) of discrete, unpackaged, blue LED chips mounted to a transparent glass or sapphire filament, a construction commonly referred to as **chip-on-glass (COG)**. Encapsulating the LED filament in a silicone resin that contains a yellow phosphor such as YAG:Ce^{3+} enables cool-white light to be generated via photoluminescence, as displayed in Fig. 10.7-1(a). A broadband- or narrowband-red phosphor can be incorporated into the phosphor blend to allow warm-white light to be emitted, as described in Secs. 10.4–10.6.

The white light emitted by a PCLED filament is portrayed in Fig. 10.7-1(b), where the drive current has been reduced to a small fraction of its normal operating value to enable the light emitted by the individual LED chips to be resolved; operation at the normal drive current is illustrated in Fig. 10.7-1(c). While the CCT is ordinarily determined by the characteristics of the phosphor, it can alternatively be controlled by replacing some of the blue chips with red ones and relying on additive color mixing (Sec. 11.3).

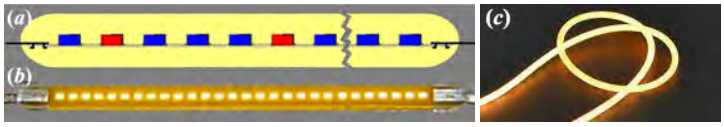


Figure 10.7-1 (a) Sketch of a PCLED filament comprising a chain of blue LEDs mounted to a glass filament and embedded in a phosphor that is designed to emit metameric cool- or warm-white light. Alternatively, a small proportion of red LEDs can be substituted for blue ones to reduce the CCT of the light via additive color mixing. (b) White-light emission from an LED filament comprising 28 individual LEDs; the drive current has been reduced to 5% of its normal operating value so that the light emitted by the individual LED chips can be resolved. (c) Metameric white light at $T_c \approx 3000$ K generated by a PCLED filament of diameter 1.5 mm and length 25 cm.

Phosphor-conversion LED filaments such as the one displayed in Fig. 10.7-1 are used in white LED-filament retrofit lamps, such as that displayed in Fig. 11.4-1(d) and discussed in Example 11.4-3.

10.8 CHIP-ON-BOARD PCLEDS

The exposition provided thus far in this chapter has been principally directed toward exploring the behavior and characteristics of discrete, single-die, phosphor-conversion LEDs that serve as sources of cool- and warm-white light. However, many lighting applications call for sources whose luminous flux is greater than that available from single-die devices. A chip-on-board (COB) PCLED, which is a module consisting of a large number of closely-spaced, individual LED dies mounted on a substrate (board) and enveloped in phosphor, offers substantially increased luminous flux. This 2D device can be viewed as a generalized version of the 1D multiple-die, PCLED filament discussed in Sec. 10.7. COB PCLEDs are widely used in a broad variety of indoor and outdoor venues for track, spot, task, downlight, automotive, industrial, horticultural, and stadium lighting.

Evolution of the White Chip-on-Board PCLED

The simple 2D array portrayed in Fig. 10.8-1(a), which contains four dies, can be thought of as a rudimentary COB PCLED. Each die is similar to the one illustrated in Fig. 10.5-1(b), and has a comparable viewing angle ($2\theta_{1/2} \approx 120°$). From the perspective of optics, arrays such as this typically approximate a point source and rely on optical components to configure the light beam into the desired spatial pattern.

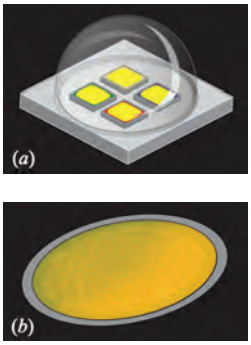


Figure 10.8-1 (a) A 2D PCLED array comprising four InGaN dies overlaid with thin phosphor sheets. This surface-mounted array, which is encapsulated in a 5-mm-diameter hemispherical lens, can be thought of as a rudimentary COB device. Drawing 18 W of electrical power, it emits metameric cool-white light with a luminous flux $P_V \approx$ 2250 lm at a wall-plug luminous efficacy $\eta_{WPE} \approx 125$ lm/W. (b) An illuminated chip-on-board (COB) PCLED containing tens of InGaN LED dies embedded in a 2.2-cm-diameter, 1.5-mm-thick layer of phosphor. A device such as this, reported in Table 10.8-1, draws 235 W of electrical power at its maximum operating current and delivers metameric warm-white light with a luminous flux of $P_V \approx 28700$ lm at a wall-plug luminous efficacy $\eta_{WPE} \approx 122$ lm/W. The spectral density of the emitted light is displayed as the green curve in Fig. 10.8-2.

A **chip-on-board (COB) PCLED**, sometimes called an **LED integrated array**, comprises a 2D array of tens (or hundreds) of densely packed, discrete, unpackaged, blue LED chips (the maximum number of chips is limited principally by thermal-management considerations). The array is encapsulated in a phosphor-containing resin and configured as a single circuit. As depicted in Fig. 10.8-1(b), it is mounted on a printed-circuit board of phenolic or aluminum, or on a substrate such as sapphire or glass. COB PCLEDs offer good thermal performance and provide diffuse sources of lighting of high uniformity. They are suitable for a wide variety of directional and non-directional single-color lighting applications. Viewing angles and operating temperatures are comparable with those for discrete PCLEDs: $2\theta_{1/2} \approx 115°$ and 85 °C, respectively. Chip-on-board devices are commercially available in a broad range of sizes, die densities, operating voltages, operating currents, luminous fluxes, luminous

efficacies, correlated color temperatures, and color rendering indices. Two or more COB PCLEDs connected in series form a **multiple COB (MCOB) device**.

Behavior and Characteristics of White COB PCLEDs

We proceed to discuss the characteristics of the InGaN-based COB PCLEDs exemplified in Table 10.8-1, under typical and maximum-current operating conditions. We consider in turn the behavior of: 1) a 5000-K cool-white COB PCLED based on a broadband yellow phosphor (upper table); 2) a 3000-K warm-white COB PCLED based on a broadband-red phosphor blend (middle table); and 3) a 3000-K warm-white COB PCLED based on a narrowband-red phosphor blend (lower table). Examination of the entries in the table reveals that these three COB devices are electrically identical and have the same values of the CRI.

The luminous flux and luminous efficacy values differ, however, since the devices utilize different phosphors. The COB PCLED with the narrowband-red phosphor blend exhibits higher values of these parameters than those offered by the COB PCLED with the broadband-red phosphor blend, as is the case with discrete PCLEDs, and for the same reasons (Sec. 10.6). However, the values for the narrowband-red phosphor blend also exceed those for the cool-white phosphor, which was not the case for the discrete PCLEDs (Sec. 10.6).

The spectral densities of the metameric white light emitted by the three COB PCLEDs compared in Table 10.8-1 are considered in Example 10.8-1, while the enhancement in luminous flux offered by COB PCLEDs relative to discrete PCLEDs that make use of comparable phosphors is detailed in Example 10.8-2.

EXAMPLE 10.8-1. ***Spectra for COB PCLEDs That Emit Metameric White Light.*** The spectral densities for the metameric white light emitted by the three COB PCLEDs characterized in Table 10.8-1 are displayed in Fig. 10.8-2. The narrow peaks in the vicinity of 450 nm arise from the InGaN pump while the peaks in the yellow and red regions are associated with the different phosphor blends. These three spectral densities closely resemble those displayed earlier for Cree discrete phosphor-conversion LEDs (Secs. 10.5 and 10.6), revealing that the two types of device make use of similar phosphor blends. In particular, the blue curve in Fig. 10.8-2 is similar to the blue curve in Fig. 10.5-2(a) for the Cree XLamp® XP-L2 PCLED; the red curve in Fig. 10.8-2 resembles the red curve in Fig. 10.6-1(a) for the Cree XLamp® XHP35.2 PCLED; and the green curve in Fig. 10.8-2 is closely related to the green curve in Fig. 10.4-2 for the Cree XLamp® XHP35.2 Pro9™ PCLED (both green curves display the unmistakable fingerprint of the spectral-spike quintet in the vicinity of 623 nm that is the hallmark of the manganese-doped potassium fluorosilicate spectrum, as explained in Sec. 10.3).

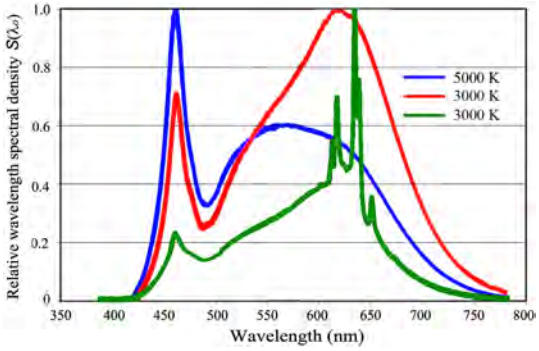


Figure 10.8-2 COB PCLED spectral densities under InGaN LED pumping. ***Blue Curve*:** 5000-K cool-white light from a broadband YAG:Ce^{3+} yellow phosphor. ***Red Curve*:** 3000-K warm-white light from a phosphor blend using the broadband YAG:Eu^{3+} red phosphor. ***Green Curve*:** 3000-K warm-white light from a phosphor blend using the narrowband KSF:Mn^{4+} red phosphor. (Data adapted from Cree data sheets CLD-DS260-REV5 and CLD-DS333-REV2 for XLamp® CMU2287 and CMU2287 Pro9™, respectively, 2023.)

Table 10.8-1 Specifications for three representative white, 2.2-cm-diameter, InGaN chip-on-board (COB) PCLEDs with a 2.8 cm × 2.8 cm footprint. The three devices are identical except for their phosphors. Data are provided for operation at a typical value of the forward current (upper row of each table) and at the maximum permissible operating current (lower row of each table).

UPPER TABLE (COOL-WHITE): Data in the upper table correspond to a device fabricated using a broadband-yellow phosphor (YAG:Ce^{3+}) that generates cool-white light with a CCT of 5000 K. (Data in the upper and middle tables adapted from Cree data sheet CLD-DS260-REV5 for XLamp® CMU2287, `https://downloads.cree-led.com/files/ds/x/XLamp-CMU2287.pdf`, 2023.)

MIDDLE AND LOWER TABLES (WARM-WHITE): Data in the middle and lower tables correspond to devices that generate warm-white light with a CCT of 3000 K using phosphor blends that incorporate broadband-red (YAG:Eu^{3+}) and narrowband-red (KSF:Mn^{4+}) phosphors, respectively. (Data in the lower table adapted from Cree data sheet CLD-DS333-REV2 for XLamp® CMU2287 Pro9™, `https://downloads.cree-led.com/files/ds/x/XLamp-CMU2287-Pro9.pdf`, 2023.)

UNITS: Successive columns display: forward current i, forward voltage V, electrical power consumption P_{EL}, luminous flux P_{V}, wall-plug luminous efficacy (WPE) $\eta_{\mathtt{WPE}}(\frac{\mathrm{lm}}{\mathrm{W}})$, wall-plug luminous efficiency (WPC) $\eta_{\mathtt{WPC}}$, chromaticity coordinates (x, y), correlated color temperature T_c, and color rendering index (CRI). The data were collected at an operating temperature of 85 °C and with a viewing (50%-power) angle of $2\theta_{1/2} \approx 115°$.

COB[a]	i^b(A)	V^b(V)	$P_{\mathrm{EL}}^{b,c}$(W)	P_{V}^{c}(lm)	$\eta_{\mathtt{WPE}}^{c,d}(\frac{\mathrm{lm}}{\mathrm{W}})$	$\eta_{\mathtt{WPC}}^{d}$	x^e	y^e	T_c^f(K)	CRI[g]
COOL-WHITE	1.62	50.5	81.8	12100	148	0.22	0.34	0.36	5000	90
	4.20	56.0	235	26000	110	0.16	0.34	0.36	5000	90
WARM-WHITE (BB-RED BLEND)	1.62	50.5	81.8	11600	142	0.21	0.43	0.40	3000	90
	4.20	56.0	235	25000	106	0.15	0.43	0.40	3000	90
WARM-WHITE (NB-RED BLEND)	1.62	50.5	81.8	13350	163	0.24	0.43	0.40	3000	90
	4.20	56.0	235	28700	122	0.18	0.43	0.40	3000	90

[a] Table entry values are rounded.
[b] The electrical drive power is related to the device current and voltage via $P_{\mathrm{EL}} = iV$, as specified in (7.1-13).
[c] The wall-plug luminous efficacy $\eta_{\mathtt{WPE}}$, luminous flux P_{V}, and electrical drive power P_{EL} are related by (8.9-4).
[d] The wall-plug luminous efficiency and efficacy are related by $\eta_{\mathtt{WPC}} = \eta_{\mathtt{WPE}}/683$, in accordance with (8.9-9).
[e] The chromaticity coordinates x and y, which are perceptual measures of color, are defined in Sec. 9.6.
[f] The correlated color temperature T_c, a measure of the color of a source of light, is defined in Sec. 9.8.
[g] The color rendering index is a measure of the faithfulness with which the color of an object is rendered (Sec. 9.9).

EXAMPLE 10.8-2. ***Luminous-Flux Comparison for COB and Discrete PCLEDs.*** The *raison d'être* of the chip-on-board device is to provide a luminous flux greater than that available from single-die devices; this is achieved by densely packing tens (or hundreds) of individual dies into a single integrated module. Table 10.8-2 reiterates the maximum-current operating characteristics of the three COB PCLEDs examined in Table 10.8-1 and compares them with those of the three discrete PCLEDs that make use of similar phosphor blends, as described in Example 10.8-1. The metric used for the comparison is the ratio of the COB PCLED luminous flux to the discrete PCLED luminous flux, $P_{\mathrm{V}}(\mathtt{COB})/P_{\mathrm{V}}(\mathtt{DISCRETE})$, which is provided in the rightmost column of Table 10.8-2. For the particular examples considered, the values for the COB PCLED luminous flux are tens of times greater than those for the discrete PCLEDs.

Table 10.8-2 Comparison of parameter values for three white, 2.2-cm-diameter, InGaN COB PCLEDs that employ three distinct phosphors with three white, 3.2 mm × 3.2 mm, InGaN discrete PCLEDs that employ similar phosphors. The upper table represents devices that generate cool-white light whereas the middle and lower tables represent devices that generate warm-white light. The three COB devices are identical except for their phosphors. The data provided represent operation at the maximum permissible operating current for each device.

***UPPER TABLE (COOL-WHITE)*:** Data in the upper table correspond to devices fabricated using a broadband-yellow phosphor that generates cool-white light with a CCT of 5000 K. (Data adapted from Cree data sheets CLD-DS260-REV5 and CLD-DS149-REV6 for XLamp® CMU2287 and XP-L2, respectively, at `https://downloads.cree-led.com/files/ds/x/XLamp-CMU2287.pdf`, 2023 and `https://downloads.cree-led.com/files/ds/x/XLamp-XPL2.pdf`, 2023.)

***MIDDLE TABLE (WARM-WHITE)*:** Data for devices that generate warm-white light with a CCT of 3000 K using phosphor blends that incorporate a broadband-red phosphor. (Data adapted from Cree data sheets CLD-DS260-REV5 and CLD-DS199-REV7 for XLamp® CMU2287 and XHP35.2, respectively, at `https://downloads.cree-led.com/files/ds/x/XLamp-CMU2287.pdf`, 2023 and `https://downloads.cree-led.com/files/ds/x/XLamp-XHP35.2.pdf`, 2023.)

***LOWER TABLE (WARM-WHITE)*:** Data in the lower table correspond to devices that generate warm-white light with a CCT of 3000 K using phosphor blends that incorporate the narrowband-red phosphor manganese-doped potassium fluorosilicate. (Data adapted from Cree data sheets CLD-DS333-REV2 and CLD-DS334-REV1 for XLamp® CMU2287 Pro9™ and XHP35.2 Pro9™, respectively, at `https://downloads.cree-led.com/files/ds/x/XLamp-CMU2287-Pro9.pdf`, 2023 and `https://downloads.cree-led.com/files/ds/x/XLamp-XHP35.2-Pro9.pdf`, 2023.)

***UNITS*:** Successive columns display: forward current i, forward voltage V, electrical power consumption $P_{\mathtt{EL}}$, luminous flux $P_{\mathtt{V}}$, wall-plug luminous efficacy (WPE) $\eta_{\mathtt{WPE}}(\frac{\mathrm{lm}}{\mathrm{W}})$, wall-plug luminous efficiency (WPC) $\eta_{\mathtt{WPC}}$, correlated color temperature T_c, color rendering index (CRI), and ratio of COB PCLED luminous flux to discrete PCLED luminous flux $P_{\mathtt{V}}(\mathtt{COB})/P_{\mathtt{V}}(\mathtt{DISCRETE})$.

DEVICE (TABLE NO.)[a]	i(A)	V(V)	$P_{\mathtt{EL}}$(W)	$P_{\mathtt{V}}$(lm)	$\eta_{\mathtt{WPE}}(\frac{\mathrm{lm}}{\mathrm{W}})$	$\eta_{\mathtt{WPC}}$	T_c(K)	CRI	$\frac{P_{\mathtt{V}}(\mathtt{COB})}{P_{\mathtt{V}}(\mathtt{DISCRETE})}$
CW COB (10.8-1)[b]	4.20	56.0	235	26000	110	0.16	5000	90	22.6
CW DISCRETE (10.5-1)[b]	3.00	3.04	9.12	1150	126	0.18	5000	80	
WW BB COB (10.8-1)[b]	4.20	56.0	235	25000	106	0.15	3000	90	17.9
WW BB DISCRETE (10.6-1)[b]	1.50	13.1	19.7	1400	71	0.10	3000	90	
WW NB COB (10.8-1)[b]	4.20	56.0	235	28700	122	0.18	3000	90	32.6
WW NB DISCRETE (10.6-1)[b]	0.70	11.9	8.33	880	106	0.16	3000	90	

[a] TABLE specifies the table number to which these data are posted. Table entry values are rounded.

[b] Abbreviations: COB = CHIP-ON-BOARD PCLED, DISCRETE = DISCRETE PCLED, CW = COOL-WHITE, WW = WARM-WHITE, BB = BROADBAND-RED PHOSPHOR BLEND, NB = NARROWBAND-RED PHOSPHOR BLEND.

BIBLIOGRAPHY

Luminescence, Photoluminescence, Luminophores, and Phosphors

R.-S. Liu and X.-J. Wang, eds., *Phosphor Handbook: Fundamentals of Luminescence*, CRC Press/Taylor & Francis, 3rd ed. 2022.

R.-S. Liu and X.-J. Wang, eds., *Phosphor Handbook: Experimental Methods for Phosphor Evaluation and Characterization*, CRC Press/Taylor & Francis, 3rd ed. 2022.

V. Dubey, N. Dubey, M. Michalska Domańska, M. Jayasimhadri, and S. J. Dhoble, eds., *Rare-Earth-Activated Phosphors: Chemistry and Applications*, Woodhead/Elsevier, 2022.

T. S. Teets, *Photoluminescence*, American Chemical Society, 2021.

A. C. Berends, M. A. van de Haar, and M. R. Krames, YAG:Ce^{3+} Phosphor: From Micron-Sized Workhorse for General Lighting to a Bright Future on the Nanoscale, *Chemical Reviews*, vol. 120, pp. 13461–13479, 2020.

C. C. Lin, W.-T. Chen, and R. S. Liu, Phosphors for White LEDs, in R. Karlicek, C.-C. Sun, G. Zissis,

and R. Ma, eds., *Handbook of Advanced Lighting Technology*, Springer, 2017.

D. R. Vij, ed., *Luminescence of Solids*, Springer, 2012.

R.-J. Xie, Y. Q. Li, N. Hirosaki, and H. Yamamoto, *Nitride Phosphors and Solid-State Lighting*, CRC Press/Taylor & Francis, 2011.

P. Hänninen and H. Härmä, eds., *Lanthanide Luminescence: Photophysical, Analytical and Biological Aspects*, Springer, 2011.

S. Ye, F. Xiao, Y. X. Pan, Y. Y. Ma, and Q. Y. Zhang, Phosphors in Phosphor-Converted White Light-Emitting Diodes: Recent Advances in Materials, Techniques and Properties, *Materials Science and Engineering: Reports*, vol. 71, pp. 1–34, 2010.

X. Piao, K.-i. Machida, T. Horikawa, H. Hanzawa, Y. Shimomura, and N. Kijima, Preparation of $CaAlSiN_3$:Eu^{2+} Phosphors by the Self-Propagating High-Temperature Synthesis and Their Luminescent Properties, *Chemistry of Materials*, vol. 19, pp. 4592–4599, 2007.

M. J. Weber, ed., *Selected Papers on Phosphors, Light Emitting Diodes, and Scintillators: Applications of Photoluminescence, Cathodoluminescence, Electroluminescence, and Radioluminescence*, SPIE Optical Engineering Press (Milestone Series Volume 151), 1998.

M. J. Weber, ed., *Selected Papers on Photoluminescence of Inorganic Solids*, SPIE Optical Engineering Press (Milestone Series Volume 150), 1998.

Narrowband Red Phosphors

W. E. Cohen, F. Du, W. W. Beers, and A. M. Srivastava, Review — The K_2SiF_6:Mn^{4+} (PFS/KSF) Phosphor, *ECS Journal of Solid State Science and Technology*, vol. 12, 076004, 2023.

F. Du, C. D. Nelson, and S. A. Krosschell, Coated Manganese Doped Phosphors, *U.S. Patent 11,060,023 B2*, Patented July 13, 2021, Filed August 9, 2018.

K. Alberi, J. Murphy, and A. Setlur, Materials for Solid-State Lighting Applications, in K. Alberi, M. Buongiorno Nardelli, A. Zakutayev, L. Mitas, S. Curtarolo, A. Jain, M. Fornari, N. Marzari, I. Takeuchi, M. L. Green, M. Kanatzidis, M. F. Toney, S. Butenko, B. Meredig, S. Lany, U. Kattner, A. Davydov, E. S. Toberer, V. Stevanovic, A. Walsh, N.-G. Park, A. Aspuru-Guzik, D. P. Tabor, J. Nelson, J. Murphy, A. Setlur, J. Gregoire, H. Li, R. Xiao, A. Ludwig, L. W. Martin, A. M. Rappe, S.-H. Wei, and J. Perkins, The 2019 Materials by Design Roadmap, *Journal of Physics D: Applied Physics*, vol. 52, 013001, Sec. 13, pp. 31–32, 2019.

J. E. Murphy, F. Garcia-Santamaria, A. A. Setlur, and S. Sista, PFS, K_2SiF_6:Mn^{4+}: The Red-Line Emitting LED Phosphor Behind GE's TriGain™ Technology Platform, *Society for Information Display (SID) Digest*, vol. 46, no. 1, book 2, paper 62.4, pp. 927–930, 2015.

E. V. Radkov, A. A. Setlur, A. M. Srivastava, and L. S. Grigorov, Red Line Emitting Phosphor for Use in LED Applications, *U.S. Patent 7,648,649 B2*, Patented January 19, 2010, Filed February 13, 2007.

A. G. Paulusz, The Predictive use of the Configurational Co-Ordinate Model for Luminescent Centres, *Journal of Luminescence*, vol. 17, pp. 375–384, 1978.

A. G. Paulusz, Efficient Mn(IV) Emission in Fluorine Coordination, *Journal of the Electrochemical Society*, vol. 120, pp. 942–947, 1973.

K. Th. Wilke, K. Albers, and R. Mannheim, Die Lumineszenz der komplexen Fluoride, *Zeitschrift für Physikalische Chemie*, vol. 2130, pp. 191–224, 1960.

A. L. Smith, New Manganese-Activated Fluoride Phosphors, *Journal of the Electrochemical Society*, vol. 101, pp. 189–194, 1954.

Narrowband Green Phosphors

I. Hendy, J. E. Murphy, and A. A. Setlur, Latest Advances in Narrow-Band Phosphors and Their Role in Color Management, *Information Display*, vol. 39, no. 3, pp. 37–42, May/June 2023.

S. J. Camardello, A. A. Setlur, J. E. Murphy, and M. D. Butts, Uranium-Based Phosphors and Compositions for Displays and Lighting Applications, *U.S. Patent Application Publication 2023/0287265 A1*, Published September 14, 2023, Filed October 14, 2022.

D. Baumann, S. Peschke, and P. Schmid, Narrow-Band Green Luminophore, *U.S. Patent Application Publication 2023/0123606 A1*, Published April 20, 2023, Filed March 30, 2021.

S. J. Camardello, M. D. Butts, R. Cassidy, J. E. Murphy, G. Parthasarathy, A. A. Setlur, O. P. Siclovan, J. Welch, and A. Yakimov, Development of New Green Phosphors for Liquid Crystal Display Backlights, *Society for Information Display (SID) Digest*, vol. 52, no. 1, book 2, paper 62-11, pp. 917–919, 2021.

H. Zhu, Y.-Q. Li, K. Gong, and S. L Dunscombe, Coated Narrow Band Green Phosphor, *U.S. Patent 11,015,116 B2*, Patented May 25, 2021, Filed March 10, 2020.

S. Li, L. Wang, D. Tang, Y. Cho, X. Liu, X. Zhou, L. Lu, L. Zhang, T. Takeda, N. Hirosaki, and R. J. Xie, Achieving High Quantum Efficiency Narrow-Band β-SiAlON:Eu^{2+} Phosphors for High-Brightness LCD Backlights by Reducing the Eu^{3+} Luminescence Killer, *Chemistry of Materials*, vol. 30, pp. 494–505, 2018.

L. Wang, X. Wang, T. Kohsei, K. i. Yoshimura, M. Izumi, N. Hirosaki, and R.-J. Xie, Highly Efficient Narrow-Band Green and Red Phosphors Enabling Wider Color-Gamut LED Backlight for More Brilliant Displays, *Optics Express*, vol. 23, pp. 28707–28717, 2015.

N. Hirosaki, R.-J. Xie, K. Kimoto, T. Sekiguchi, Y. Yamamoto, T. Suehiro, and M. Mitomo, Characterization and Properties of Green-Emitting β-SiAlON:Eu^{2+} Powder Phosphors for White Light-Emitting Diodes, *Applied Physics Letters*, vol. 86, 211905, 2005.

Quantum-Dot and Hybrid Phosphors

N. T. Kalyani, S. J. Dhoble, M. Michalska-Domańska, B. Vengadaesvaran, H. Nagabhushana, and A. K. Arof, eds., *Quantum Dots: Emerging Materials for Versatile Applications*, Woodhead/Elsevier, 2023.

K. Upadhyay, S. Thomas, and R. K. Tamrakar, eds., *Hybrid Phosphor Materials: Synthesis, Characterization, and Applications*, Springer, 2022.

R.-S. Liu and X.-J. Wang, eds., *Phosphor Handbook: Novel Phosphors, Synthesis, and Applications*, CRC Press/Taylor & Francis, 3rd ed. 2022.

P. O. Anikeeva, J. E. Halpert, M. G. Bawendi, and V. Bulović, Electroluminescence from a Mixed Red–Green–Blue Colloidal Quantum Dot Monolayer, Nano Letters, vol. 7, pp. 2196–2200, 2007.

Historical Accounts and Seminal Publications

See also the bibliographies in Chapters 5–7.

I. Akasaki, Fascinated Journeys into Blue Light (Nobel Lecture in Physics, 2014), *Reviews of Modern Physics*, vol. 87, pp. 1119-1131, 2015; H. Amano, Growth of GaN on Sapphire via Low-Temperature Deposited Buffer Layer and Realization of p-type GaN by Mg Doping Followed by Low-Energy Electron Beam Irradiation (Nobel Lecture in Physics, 2014), *Reviews of Modern Physics*, vol. 87, pp. 1133-1138, 2015; S. Nakamura, Background Story of the Invention of Efficient Blue InGaN Light Emitting Diodes (Nobel Lecture in Physics, 2014), *Reviews of Modern Physics*, vol. 87, pp. 1139-1151, 2015.

Class for Physics of the Royal Swedish Academy of Sciences, Scientific Background on the Nobel Prize in Physics 2014: Efficient Blue Light-Emitting Diodes Leading to Bright and Energy-Saving White Light Sources, *Kungliga Vetenskapsakademien*, pp. 1–9, 2014.

S. Nakamura and M. R. Krames, History of Gallium-Nitride-Based Light-Emitting Diodes for Illumination, *Proceedings of the IEEE*, vol. 101, pp. 2211–2220, 2013.

I. Akasaki, Key Inventions in the History of Nitride-Based Blue LED and LD, *Journal of Crystal Growth*, vol. 300, pp. 2–10, 2007; Erratum: vol. 310, p. 2683, 2008; Erratum: vol. 312, pp. 351–357, 2010.

CHAPTER

11

LED LIGHTING

11.1 MERITS OF LED LIGHTING 340

11.2 SINGLE-COLOR LEDS 343

11.3 ADDITIVE COLOR-MIXING LEDS 344

11.4 RETROFIT LED LAMPS 348

11.5 HYBRID LEDS 353

11.6 LED LUMINAIRES 356

11.7 OLED LIGHT PANELS 358

11.8 SMART AND CONNECTED LED LIGHTING 361

11.9 LED PERFORMANCE METRICS 363

In 1962, **Nick Holonyak, Jr. (1928–2022)**, working with **Saverio (Samuel) Bevacqua** at the General Electric Research Laboratory in Syracuse, NY, developed a bright GaAsP LED and laser diode that emitted light in the red.

M. George Craford (born 1938), a mentee of Nick Holonyak, co-invented the yellow GaAsP:N LED in 1971. He led a team that in 1990 developed AlInGaP LEDs for generating high-brightness yellow, orange, and red light.

LED Lighting: Devices and Colorimetry. Malvin Carl Teich. Google Books. Published 2024.

LED lighting, where the observer views the light scattered from surrounding objects, is an essential technology in an enormous number of arenas, including exhibition and entertainment lighting; landscape lighting; aerospace and military lighting; automotive lighting; and residential, architectural, and street lighting. Also called **solid-state lighting**, it is nearly universally used for illumination applications. LED lighting offers excellent color rendering quality and can be designed to provide dynamic lighting that renders any color or any combination thereof, including white. It can also offer personalizing lighting throughout the day, simultaneously integrating the art of illumination with systems that govern various sensory environments. **Connected lighting systems** comprise luminaires outfitted with sensors that network with these systems and with each other to dynamically modify the resident space. The capabilities of LED lighting are increasingly vital for agricultural, horticultural, and germicidal applications.

This chapter can be considered to be a continuation of Chapter 10, which considers in detail the generation of metameric white light by phosphor-conversion light-emitting diodes (PCLEDs). These efficient lighting components are widely used in many facets of LED lighting. This chapter also relies extensively on Chapter 7, which is devoted to the operation and behavior of LEDs, and on Chapters 8 and 9, which narrate the principles of human color vision and colorimetry, respectively.

From a historical perspective, the incandescent filament lamp, a predecessor to the LED, served as the workhorse of artificial lighting from 1879, shortly after its invention by Thomas Edison and his British rival Joseph Swan, until the early 2000s. The principal limitation of incandescent sources is that they emit graybody radiation so that the conversion of electrical power to optical power strongly favors the infrared; only about 5% of the optical power is emitted in the visible region while some 95% fails to provide light since it is emitted in the infrared as heat. The fluorescent lamp, which entered the marketplace in the late 1930s, was somewhat more efficient and replaced the incandescent lamp in some venues but its light was widely deemed to be inferior in quality. The use of fluorescent light accelerated with the advent of the compact fluorescent lamp (CFL), which became widely available about 1995, but incandescent lighting maintained its primacy by virtue of its ideal color rendering, simple construction, low price, and familiarity to the public.

In the early 2000s, however, incandescent lighting finally yielded its primacy to LED lighting, which is highly efficient and offers luminous-flux levels and wall-plug luminous efficacies that are substantially superior to those available with incandescent and fluorescent lighting. LED sources also offer manifold advantages relative to their traditional counterparts, as discussed above.

LED lighting is based on light-emitting diodes. By way of a historical introduction to the development of LEDs for lighting, the first high-brightness LED in the visible region was fabricated by Nick Holonyak (p. 338) and Saverio Bevacqua using the compound semiconductor $\mathrm{GaAs}_{1-x}\mathrm{P}_x$. As reported in the manuscript they submitted to *Applied Physics Letters* on 17 October 1962 that was published on 1 December 1962,[†] their p–n junction device functioned as a red LED at room temperature and as a red laser diode that emitted light at 710 nm when its temperature was lowered sufficiently. It was subsequently discovered by George Craford and his colleagues that $\mathrm{GaAs}_{1-x}\mathrm{P}_x$ could be induced to emit light at wavelengths shorter than red, albeit inefficiently, by doping with nitrogen.[‡]

[†] N. Holonyak, Jr. and S. F. Bevacqua, Coherent (Visible) Light Emission from $\mathrm{GaAs}_{1-x}\mathrm{P}_x$ Junctions, *Applied Physics Letters*, vol. 1, pp. 82–83, 1 December 1962 (submitted 17 October 1962).

[‡] M. G. Craford, R. W. Shaw, A. H. Herzog, and W. O. Groves, Radiative Recombination Mechanisms in GaAsP Diodes With and Without Nitrogen Doping, *Journal of Applied Physics*, vol. 43, pp. 4075–4083, 1972.

The salutary features of LED lighting, in the context of traditional lighting technologies, are presented in Sec. 11.1. Single-color LEDs and some of their uses, particularly in architectural lighting, are discussed in Sec. 11.2. Methods for generating white light, other than those involving PCLEDs, include the use of additive color-mixing devices (Sec. 11.3) and hybrid devices (Sec. 11.5). These approaches make use of AlInGaP, the material of choice for generating light in the red, orange, yellow-orange (amber), and yellow; and InGaN, the material of choice for generating light in the green, blue, and violet (Sec. 7.3). Retrofit lamps and variations thereof are considered in Sec. 11.4 and LED luminaires are studied in Sec. 11.6. Smart lighting and connected lighting, as well as human-centric lighting, IoT lighting, LiFi, and the myriad implementations of these variants, are explored in Sec. 11.8. The properties of white OLED light panels are introduced in Sec. 11.7. Finally, a performance comparison of the figures of merit for various sources of illumination is presented in Sec. 11.9. It will become clear in this chapter that LEDs serve as the underlying sources for all manner of lighting devices.

As detailed in Chapter 7 and summarized in Sec. 10.1, new types of LEDs introduced in recent decades, including quantum-dot, organic, and perovskite devices, are also under development.

11.1 MERITS OF LED LIGHTING

Traditional Technologies

The advent of LED lighting has rendered largely obsolete many of the traditional lighting technologies that have been employed since the late nineteenth century, when electricity became widely available for practical use. As detailed below, the operating principles underlying traditional technologies rely on either gas-discharge phenomena or on incandescence. The light produced in a gas discharge is characterized as spontaneous emission (Sec. 4.4) whereas incandescence involves an interplay among spontaneous emission, stimulated emission, and absorption (Secs. 4.7). Incandescence results from the transitions of free and valence electrons in hot solid materials (Sec. 4.8).

Incandescent Lamps. The quintessential example of an incandescent lamp is the old-fashioned glass light bulb containing a thin tungsten filament that is ohmically heated by passing an electric current through it (Secs. 4.8 and 9.7). Tungsten is the material of choice because it has the lowest vapor pressure and highest melting point of any metal. Incandescent lighting was widely used throughout the twentieth century because of its convenience, appealing spectral properties, and optimal color rendering. However, its color is limited to shades of reddish-white and it is highly inefficient, with only about 5% of the consumed energy converted to visible light. The lifespan is limited to approximately 1500 h, which is roughly the same as that for the mantles used in *incandescent gas* lamps, which were employed for street lighting before *incandescent electric* lamps became available.

Halogen Incandescent Lamps. Halogen lamps are incandescent sources whose bulbs contain a small amount of a halogen gas such as bromine. The halogen and tungsten atoms react chemically so that the evaporated tungsten is redeposited on the filament when the halogen cools. This approach allows the efficiency to be increased from 5% to about 7% and extends the lifespan from 1500 h to roughly 4000 h.

Fluorescent Lamps. Fluorescent lamps are low-pressure gas-discharge devices that operates by passing an electric current through mercury vapor. The ensuing excitation of the Hg atoms results in the emission of ultraviolet photons that strike a phosphor coating deposited on the interior of the glass envelope. The color of the resulting photoluminescence depends on the nature of the phosphor (Sec. 10.2). Fluorescent lamps

are more efficient than incandescent lamps, but their color rendering is substantially inferior. They also contain mercury, which is environmentally hazardous.

Low-Pressure Gas-Discharge Lamps. Low-pressure gas-discharge lamps operate by passing an electric current through a gas such as neon. The atoms are excited (or ionized) by collisions and the ensuing decay to the ground state gives rise to spontaneous emission (fluorescence) with a color that depends on the atomic species (Sec. 4.4). Buffer gases are often added to the mix to facilitate excitation. Many monoatomic species (and their ions) exhibit fluorescence at innumerable wavelengths. Neon signs, which were introduced in the early 1900s, serve as an example. Although seldom used for illumination, their eye-catching red-orange glow was ubiquitous in advertising until the 1980s when LED signage became available. A blue glow is obtained by using argon instead of neon; white is produced by xenon and colors ranging from green to purple are generated by mixtures of krypton and xenon. Noble gases are employed because of their stability and lack of reactivity. Whatever the choice of gas, however, the term "*neon* sign" is universally used. The introduction of a properly designed optical resonator into a neon sign can provide optical feedback and lead to stimulated emission (Sec. 4.5), the process underlying laser action. Indeed, the *neon laser*, which is usually referred to as the *helium-neon laser* since helium serves as a buffer, was the first gas laser to be operated (in 1960).

Low-Pressure Sodium Lamps. Low-pressure sodium (LPS) gas-discharge lamps, introduced in the 1930s, emit monochromatic yellow light on their closely spaced Na D-line doublet. They are often used for outdoor illumination because of their long lifespan, ease of filtering (which minimizes light pollution for astronomical viewing), and large wall-plug luminous efficiency $\eta_{\mathtt{WPC}}$. The latter is a consequence of the proximity of the D-line doublet wavelength ($\lambda_{\mathrm{NaD}} = 589$ nm) to the yellowish-green peak wavelength of the photopic luminous efficiency function ($\lambda_0 = 555$ nm). The value of the WPC is readily determined via (8.9-10), which provides $\eta_{\mathtt{WPC}} = \eta_{\mathtt{PCE}}\, V(589\text{ nm}) \approx 0.822\,\eta_{\mathtt{PCE}}$ lm/W. Since empirical measurements show that $\eta_{\mathtt{PCE}} \approx 0.27$, this leads to to $\eta_{\mathtt{WPC}} \approx 0.22$. A serious difficulty with LPS lamps is that the emitted light is totally devoid of color-rendering capability because of its monochromaticity (CRI = 0); the colors of illuminated objects are therefore difficult to discern and appear unnatural.

High-Pressure Sodium Lamps. High-pressure sodium (HPS) lamps operate on the same principle as low-pressure sodium lamps and are generally used for the same purposes. However, the increased pressure results in a broader spectrum (Sec. 4.6) and a greater lifespan. The broadened spectrum in turn leads to a color rendering index CRI = 15.

High-Pressure Mercury-Vapor Lamps. The operation of these gas-discharge lamps is similar to that of high-pressure sodium lamps except that the generation of a plasma requires the mercury to evaporate and a number of Hg lines at discrete visible wavelengths are excited. The ensuing light appears bluish-white with $T_c \approx 4000$ K and a color rendering index CRI = 50. In some devices, phosphors are coated on the interior of the quartz envelope to introduce other colors via photoluminescence, much as with fluorescent lamps. Most of the high-pressure mercury-vapor lamps used for lighting in Europe were replaced by HPS lamps in the 1960s, since the latter are more efficient.

High-Pressure Xenon Lamps. Just as sodium and mercury vapor are used to generate light in high-pressure lamps, so too is the noble gas xenon. Since the spectrum of Xe is much broader than that of Na or Hg vapor, however, these lamps generate light with a far higher color rendering index, i.e., CRI = 95.

Ceramic Metal-Halide Lamps. The luminous efficacy and color rendering index of high-pressure discharge lamps such as HPS can be substantially enhanced by incorporating metal halide compounds in the gas mixture, which greatly expands the number of visible spectral lines. Such compounds include various rare-earth halides

and the halides of Sc, Na, Tl, and In. A salutary feature of these lamps is that the color temperature may be adjusted by suitably choosing the added compounds. Metal-halide lamps outperform high-pressure Hg and Xe lamps with respect to luminous efficacy, and are far superior to HPS lamps with respect to color rendition. Ceramic metal-halide lamps incorporate ceramic arctubes, which allow higher operating temperature and thus increased luminous efficacy and color rendering.

Carbon-Arc Lamps. Carbon-arc lamps were invented by Sir Humphry Davy in the early 1800s and gained wide use in the late 1800s as the first practical, electrically excited sources for street lighting. They consist of a pair of carbon electrodes held at a high potential difference and separated by a small gap. The electric field at the gap ionizes the air between the electrodes and creates an arc (plasma) that links them via a conductive path. The high plasma temperature in turn leads to incandescence from the carbon electrodes and to the evaporation of some of the carbon atoms. These atoms are then excited by collisional mechanisms and emit light on de-excitation. By virtue of their large luminous flux, carbon-arc lamps were also used, in conjunction with mirrors and lenses, as searchlights for military and maritime applications. In particular, they were coupled with Fresnel lenses (Sec. 1.5) to generate the light beacons that emanated from many lighthouses in the period from 1850 until the early 1900s.

Salutary Features of LED Lighting

At its core, LED lighting relies on spontaneous recombination radiation generated at the junction region between two different materials (Chapter 7) and on the behavior of photoluminescent phosphors (Chapter 10). It offers a great many salutary features in comparison with the traditional incandescent and gas-discharge lighting technologies considered above. Indeed, it is the most efficient and versatile lighting technology ever developed, as is evidenced by the entries provided in Table 11.9-1 that compare representative illumination parameters for all manner of light sources. It is not an exaggeration to proclaim that LEDs have revolutionized lighting worldwide.

High Efficiency. Large values of the wall-plug luminous efficacy (WPE), and the accompanying wall-plug luminous efficiency (WPC), reveal that LEDs use far less electrical power than incandescent and gas-discharge sources to generate a given luminous flux. Indeed, solar panels (in conjunction with batteries for energy storage) can be used to power LEDs and can therefore bring light to remote and isolated geographical regions.

Because the spectral density of LED light is readily confined to the visible region, LED sources offer luminous efficacies far greater than those attainable with incandescent sources.

High-Quality Color Rendering. LED lighting offers excellent color rendering, which indicates that colored objects appear natural under LED illumination.

Convenience. LEDs are compact, monolithic devices that are resistant to damage from shock, vibration, external impacts, and weather. With appropriate heat-sinking, they operate from low to high temperatures. They offer instantaneous switch-on and continuous dimmability without flicker. They operate noiselessly and emit little ultraviolet radiation or infrared heat that could damage the objects they illuminate. They can be fabricated using mature manufacturing technology and do not contain toxic materials, thereby facilitating disposal.

Long Operational Life, Slow Failure, and Low Cost. LEDs are endowed with lifespans that can exceed 100000 hours, far longer than the 1500 h for a typical incandescent lamp, 4000 h for a halogen incandescent lamp, and 10000 h for a compact fluorescent lamp. When LEDs fail, they do so in a gradual rather than in a sudden manner. These features result in greatly reduced long-term replacement and maintenance costs.

Broad Choice of Directionality, Hue, Saturation, and Luminance. Multi-LED arrays and optics offer directional or omnidirectional emission. LED lighting provides luminance and chromaticity ranges that span the gamut of human vision, including a continuum of whites.

Dynamic-, Smart-, and Connected-Lighting Capabilities. The colors, temporal irradiance patterns, and spatial distributions of light produced by LEDs can be dynamically programmed. Furthermore, the electronic drivers can communicate wirelessly with each other and with collections of sensors to provide smart networks, e.g., LED traffic signals controlled by sensors embedded in the street pavement.

11.2 SINGLE-COLOR LEDS

Single-color LEDs are usually MQWLEDs that are fabricated from semiconductor materials with specific bandgap wavelengths and operate via electroluminescence. **ELLEDs** such as these are generally designated by either their peak wavelength (PWL) (Secs. 7.2–7.4) or dominant wavelength (DWL) (Sec. 9.6). MQWLEDs may instead be fashioned into **PCLEDs** that make use of photoluminescence (Sec. 10.2) from broadband or narrowband phosphor blends (Secs. 10.3 and 10.4), in which case they are designated by their colors.

Single-color LEDs have applications in the display, signage, automotive, and sensing domains, as well as consumer electronics, entertainment, and horticulture, along with clinical medicine. Examples include:

- Traffic signals (Fig. 7.4-1), automobile taillights, and emergency-vehicle alerts.
- Information displays and advertising.
- Fluorescence- and absorption-based sensing.
- Backlighting and status indication.
- Event and stage lighting.
- Architectural lighting (Secs. 11.2-1 and 11.2-2).
- Plant growth and aquarium maintenance.
- Body fluid monitoring and photodynamic therapy.

Devices such as these are commercially available in many colors, at low, medium, and high optical powers, and with various footprints, as exemplified by the listing provided below (adapted from Cree color LED portfolio data sheet FS05R21 dated April 2023):

PCLED photoluminescence colors.

- Blue
- Cyan
- Mint
- Lime
- Yellow
- Amber
- Red–Orange
- Red
- Magenta
- Purple

ELLED electroluminescence wavelengths.

- Violet: 400–420 nm PWL
- Royal Blue: 450–465 nm PWL
- Blue: 465–480 nm DWL
- Cyan: 490–510 nm DWL
- Green: 520–535 nm DWL
- Amber: 585–595 nm DWL
- Red–Orange: 610–620 nm DWL
- Red: 620–630 nm DWL
- Photo Red: 650–670 nm PWL
- Far Red: 720–740 nm PWL

Architectural Lighting

Lighting expositions in the public square offer a fine venue for qualitatively illustrating the effectiveness of single-color LEDs that offer a broad range of colors (hue, saturation,

and luminance). By way of example, we demonstrate how LED lighting highlights the grandeur of two iconic structures: the Eiffel Tower in Paris and the Empire State Building in New York City.

Eiffel Tower in Paris. An engaging example that illustrates the effectiveness of LED lighting technology is provided by the illuminated Eiffel Tower in Paris, images of which are displayed in Fig. 11.2-1.

Figure 11.2-1 The Eiffel Tower of Paris, locally known as *la dame de fer*, illuminated by LED lighting. This 330-m high tower, erected as the entrance arch for the 1889 World's Fair in Paris, may well be the most recognizable structure in the world. Over the years, the iconic shape of the Eiffel Tower has successively been highlighted by gas lamps, incandescent lamps, fluorescent lights, high-pressure sodium-vapor lamps, and now by a programmable, dynamic LED lighting system that is synchronizable with music and sound.

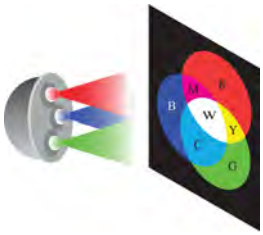
Empire State Building in New York City. Another dramatic example that illustrates some of the capabilities of LED lighting technology is provided by the system that illuminates the Empire State Building, as illustrated in Fig. 11.2-2.

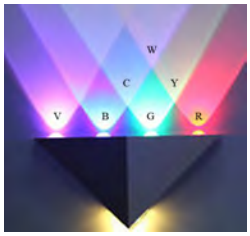
Figure 11.2-2 The Empire State Building is a 103-story, Art-Deco skyscraper in New York City that was completed in 1931. The name of the building derives from the nickname for New York State: *The Empire State*. Its exterior was first illuminated by metal-halide lamps and floodlights in 1964, which were replaced by a computer-controlled LED lighting system in 2012. This system provides programmable and dynamic lighting, offering a vast array of colors and other display options that can be synchronized with music or sound.

Drive Circuitry. The electronic drivers for the LEDs used in architectural lighting systems typically employ drive circuitry similar to that schematized in Figs. 7.4-2 and 7.4-3. A collection of LEDs of a particular color are usually connected in series and driven by a pulse-width-modulated (PWM) current provided by a drive transistor. The radiant flux generated is determined by the average current passing through the LEDs, which in turn is governed by the duty cycle of the PWM current. Banks of LEDs of different colors (usually red, green, and blue) are used to generate light of arbitrary hue, saturation, and luminance, including white. An addressable microprocessor system is usually used to control the relative light levels generated by the different-color LEDs, which enables the overall color and intensity of the light to vary with time and position in an arbitrary prescribed manner. Collections of such lighting units are readily concatenated into lighting networks.

11.3 ADDITIVE COLOR-MIXING LEDS

The *second method* for generating metameric white light using LEDs (of the three codified in Sec. 10.1) relies on multiple single-color dies. Light from red, green, and blue sources can be superposed to yield a broad range of colors, including metameric white, as may be understood from the principles of additive color mixing disclosed in Chapter 9. Initially introduced in Sec. 9.1, this process is portrayed in Fig. 11.3-1. Variable hue lamps that operate via color mixing are discussed in Sec. 11.5.





R + G	→	Y (Yellow)
G + B	→	C (Cyan)
B + R	→	M (Magenta)
R + G + B	→	W (White)

Figure 11.3-1 Additive color mixing. *Left*: A device that generates light of tunable color can be constructed from LEDs that emit light in the Red (R), Green (G), and Blue (B), as schematically illustrated. *Center*: Light produced by a wall sconce constructed from LEDs that emit, from left to right, Violet (V), Blue, Green, and Red. (Adapted from *Wandlampe* by Mattes, 4 July 2018, via Wikimedia Commons.) *Right*: The color-mixing rules specified in the table delineate the colors exhibited in the images at left and center. The yellow (Y), cyan (C), and white (W) generated by the wall sconce exhibit limited brightness because of reduced luminance, as exemplified in Examples 9.3-1–9.3-3.

Color-Mixing LEDs (CMLEDs)

A **color-mixing light-emitting diode (CMLED)**, also called a **color-mixing package**, comprises an array of several individually addressable dies in close proximity that are incorporated into a single LED package. Modern CMLEDs contain red, green, blue, and white (RGBW) dies, as sketched in Fig. 11.3-2. The RGB emitters are depicted by the colors of the light they generate. Their individual radiant power levels are electrically adjustable and their properties resemble those of the discrete, small-area, III–V MQWLEDs specified in the upper portion of Table 7.4-1.

The white phosphor-conversion element, whose radiant power is also electrically adjustable, serves to enhance the luminous flux and wall-plug luminous efficacy of the CMLED. This element is similar to the discrete white PCLED displayed in Fig. 10.5-1(*b*). Portrayed as yellow in Fig. 11.3-2 (the color of the phosphor), its specifications are similar to those of the cool- or warm-white PCLEDs reported in Tables 10.5-1 and 10.6-1, respectively.

CMLEDs are typically supplied in pockets on a tape that is wound on a reel to facilitate automated assembly. They are seldom used as sources of metameric white light; PCLEDs serve this purpose. Rather, CMLEDs are used in color-changing, stage, architectural, and entertainment applications. RGBW COB (chip-on-board) LEDs can be fabricated and are sometimes used for commercial and industrial applications, including stage, architectural, and landscape lighting.

Characteristics of an Additive Color-Mixing LED. The specifications for a representative CMLED, the Cree XLamp® XM-L Color Gen 2 High Density LED, are presented in Table 11.3-1. The data reported are for each of the four dies operating independently, and their peak wavelengths λ_p are designed to fall within the ranges

Figure 11.3-2 Sketch of a CMLED, a multicolor color-mixing LED comprising individually addressable red, green, and blue dies, along with a white phosphor-conversion emitter of selectable CCT. The dies are housed within a ceramic surface-mounted device (SMD) capped by a 5-mm-diameter hemispherical lens. CMLEDs such as this can be electrically tuned to emit essentially any color within the gamut of human vision, including metameric white light.

indicated. In this particular example the phosphor-conversion element emits cool-white light at 6000 K, but CMLEDs are also available in which the constituent PCLED emits neutral- or warm-white light. The entries labeled ARRAY represent overall values when all RGBW elements are fully energized.

The upper and lower portions of Table 11.3-1 represent operation at typical and maximum permitted current levels, respectively. Comparing them makes it clear that the luminous flux P_{V} for each die increases dramatically as the current level increases, but this comes at the expense of a substantial diminution of the wall-plug luminous efficacy η_{WPE} as a consequence of efficiency droop (Sec. 7.4).

The associated spectral densities and chromaticity diagram are displayed and discussed in Example 11.3-1.

Table 11.3-1 Specifications for a high-density RGBW multicolor color-mixing LED (CMLED) with individually addressable elements packaged as a 5 mm × 5 mm surface-mounted device (SMD). Data are presented for operation at a typical current (upper table) and at the maximum operating current (lower table), when each die is operated independently. Successive columns display the following parameters: range of peak wavelengths λ_p, current i, forward voltage V, electrical power consumption P_{EL}, luminous flux P_{V}, wall-plug luminous efficacy (WPE), wall-plug luminous efficiency (WPC), chromaticity coordinates (x, y), and correlated color temperature T_c. The data displayed were collected at an operating temperature of 25 °C and at a viewing (50%-power) angle $2\theta_{1/2} \approx 120°$. (Data adapted from Cree Data Sheet CLD-DS273-REV8 for XLamp® XM-L Color Gen 2 LEDs, `https://downloads.cree-led.com/files/ds/x/XLamp-XMLDCL.pdf`, 2023.)

	RGBW CMLED[a]	λ_p(nm)	i^b(A)	V^b(V)	$P_{\mathrm{EL}}^{b,c}$(W)	P_{V}^c(lm)	$\eta_{\mathrm{WPE}}^{c,d}(\frac{\mathrm{lm}}{\mathrm{W}})$	η_{WPC}^d	x^e	y^e	T_c^f(K)
TYPICAL	RED	620–630	0.35	2.1	0.735	80	109	0.16	–	–	–
	GREEN	520–535	0.35	2.6	0.910	155	170	0.25	–	–	–
	BLUE	450–465	0.35	2.9	1.015	23	22	0.03	–	–	–
	WHITE	–	0.35	2.9	1.015	155	153	0.22	0.32	0.34	6000
	ARRAY	–	–	–	3.675	413	112	0.16	–	–	–
MAXIMUM	RED	620–630	1.75	2.8	4.90	340	69	0.10	–	–	–
	GREEN	520–535	1.75	3.2	5.60	411	73	0.11	–	–	–
	BLUE	450–465	1.75	3.5	6.13	81	13	0.02	–	–	–
	WHITE	–	1.75	3.6	6.30	543	86	0.13	0.32	0.34	6000
	ARRAY	–	–	–	23.0	1375	60	0.09	–	–	–

[a] Table entry values are rounded.
[b] The electrical drive power is related to the device current and voltage via $P_{\mathrm{EL}} = iV$, as specified in (7.1-13).
[c] The wall-plug luminous efficacy η_{WPE}, luminous flux P_{V}, and electrical drive power P_{EL} are related by (8.9-4).
[d] The wall-plug luminous efficiency and efficacy are related by $\eta_{\mathrm{WPC}} = \eta_{\mathrm{WPE}}/683$, in accordance with (8.9-9).
[e] The chromaticity coordinates x and y, which are perceptual measures of color, are defined in Sec. 9.6.
[f] The correlated color temperature T_c, a measure of the color of a source of light, is defined in Sec. 9.8.

EXAMPLE 11.3-1. ***Spectra and Chromaticity Diagram for a Color-Mixing LED.*** The spectra and chromaticity diagram for a CMLED such as that illustrated in Fig. 11.3-2 are displayed in Fig. 11.3-3. The red (R), green (G), and blue (B) spectral densities portrayed in Fig. 11.3-3(a) exhibit peaks at $\lambda_p = 625$, 530, and 455 nm, respectively, which fall within the design ranges specified in Table 11.3-1. The peaks of the black curve, which give rise to metameric cool-white (W) light at $T_c \approx 6000$ K, are associated with a blue InGaN LED die ($\lambda_p \approx 445$ nm, $\Delta\lambda_{FWHM} \approx 20$ nm) and a thin yellow phosphor overcoating ($\overline{\lambda} \approx 570$ nm, $\Delta\lambda_{FWHM} \approx 120$ nm, see Sec. 10.3). This curve is similar to the black curve presented in Fig. 10.6-1 for a cool-white PCLED.

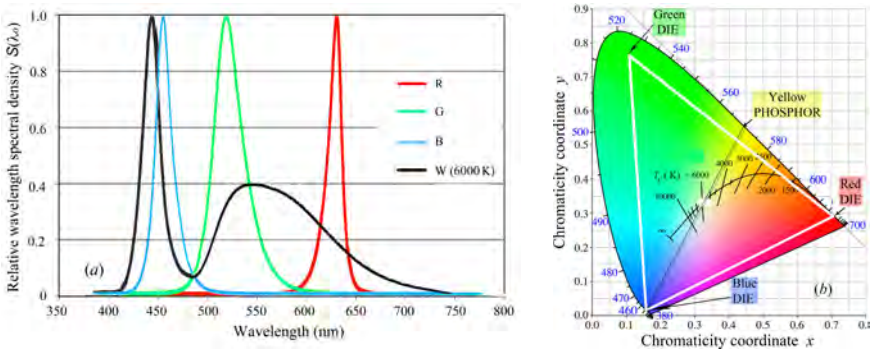


Figure 11.3-3 (a) Spectral densities for the light emitted by individually addressable R, G, B, and W dies in a CMLED such as that pictured in Fig. 11.3-2 and specified in Table 11.3-1. When operated independently at a temperature of 25 °C and at a current of 350 mA per die, the peak wavelengths of the RGB LED curves lie in the wavelength regions $\lambda_p = 620$–630, 520–535, and 450–465 nm, respectively. When operated at 25 °C and 350 mA, the phosphor-conversion (W) element emits metameric cool-white light with a CCT $\approx$ 6000 K (black curve). (b) The designated wavelengths of the red, green, and blue dies form the vertices of a triangle (white) on the chromaticity diagram that contains nearly the entirety of the Planckian locus and a substantial portion of the human color gamut. The yellow phosphor and its 445-nm blue die yield metameric cool-white light, much as illustrated in Fig. 10.5-2(b).

The chromaticity coordinates associated with these four sources are displayed in Fig. 11.3-3(b). As explained in Sec. 9.6, the coordinates for the light generated by the red die lie on the outer boundary of the diagram, the coordinates for the light generated by the blue die lie near the boundary, and the coordinates for the light generated by the green die lie interior to the diagram. These colors form the vertices of the white triangle traced out on the chromaticity diagram. It follows from the discussion surrounding Figs. 9.6-2 and 9.6-4, and Example 9.6-2, that all colors within the triangle can be generated from mixtures of the three colors at the vertices. The phosphor-conversion element generates metameric white light ($T_c \approx 6000$ K) with chromaticity coordinates that lie within the white oval near the center of the diagram, which is located at the intersection of the Planckian locus and the line joining the coordinates of the yellow phosphor and the 445-nm blue die.

Modeling the Behavior of Color-Mixing LEDs. The behavior of additive color-mixing LEDs can be modeled by augmenting the RGB coordinates on the chromaticity diagram of Fig. 11.3-3(b) with the blue and yellow coordinates associated with the phosphor-conversion element. This process serves to convert the white triangle into an irregular pentagon with a slightly enlarged color gamut. In principle, the chromaticity coordinates for light consisting of an arbitrary mixture of these five colors could then be established by generalizing the method set forth in Example 9.6-2. The correlated color temperature follows from the chromaticity coordinates, as explained in Example 10.5-4, and the wall-plug luminous efficacy could be estimated by making use of the approach provided in Example 10.5-5.

Merits and Limitations of White-Light Generation Using CMLEDs. We briefly compare the relative advantages of CMLEDs and PCLEDs for generating metameric white light. PCLEDs are the dominant technology for the reasons indicated below.

Advantages of CMLEDs Over PCLEDs.

- The individual $\mathfrak{RGB}$ primaries are independently controlled, enabling dynamic, adaptive CCT tuning and enhanced CRI; colors other than white can be accessed.
- The future possibility of using WQLED, WOLED, and QPeWLED elements with improved efficiencies (Secs. 7.5–7.7).
- CMLEDs sidestep the reduction of luminous flux and WPE arising from the imperfect photoluminescence quantum yield and complementary photoluminescence quantum defect in PCLEDs (Sec. 10.2).
- The ultimate CMLED theoretical WPE is roughly estimated to be 325 lm/W, which exceeds the estimated PCLED value of 255 lm/W, since CMLEDs are not subject to quantum yield or complementary quantum defect losses.

Advantages of PCLEDs Over CMLEDs.

- Enhanced CRI using narrowband-red and narrowband-green phosphor blends.
- Absence of color instability from the different temperature dependencies, current dependencies, and degradation rates of the constituent chips in a CMLED.
- Simpler drive electronics.
- No multiple constituent-chip beams whose emission patterns require blending.
- Avoids the relative inefficiency of green and amber MQWLEDs ("green gap").
- Simpler design; only one chip type so precise control of chip location not required.
- More straightforward and mature manufacturing process.
- Lower materials and production costs.

11.4 RETROFIT LED LAMPS

A retrofit lamp is a lighting device that replaces an existing traditional light source, such as incandescent (or fluorescent) lamp, with an energy-efficient LED alternative without having to significantly modify the existing fixture. Almost all modern white retrofit lamps are based on phosphor-conversion light-emitting diodes (PCLEDs) such as those described in Chapter 10.

Lamps, Modules, and Engines. Multiple discrete LEDs, LED filaments, and chip-on-board LEDs are often incorporated into glass or plastic housings called **bulbs** to create **LED lamps**, devices that generate artificial light. An **LED module** (or **LED light engine**) consists of an LED lamp, along with its driver, associated optics, and primary thermal management system, all assembled in a compact package. A white LED lamp that is designed to be a drop-in replacement for a tungsten incandescent lamp is called an **LED retrofit lamp**.

Bases. Retrofit lamps for home use typically operate at line voltage. Some are dimmable while others are not. Most employ the metal screw base developed by Edison in the late 1800s. The most commonly used Edison bases are designated E26 and E12 in the U.S. (120 V), and E27 and E11 in the E.U. (240 V). The number following "E" represents the diameter of the base in mm, measured across the peaks of its thread. E26 and E27 are referred to as medium (or standard) Edison screw bases (MES) whereas E12 and E11 are called candelabra bases. E26 and E27 are sufficiently close in size that they are generally physically interchangeable. Other Edison screw bases are also

used. While right-handed Edison screw bases are the norm, left-handed versions find use for special purposes, such as avoiding theft in public spaces such as the subway.

Bulbs (Envelopes). Used since the early twentieth century, classic pear-shaped A-series bulbs are designed to optimize the distribution of the emitted light. The most commonly encountered versions are designated A19 in the U.S., where the number 19 represents the major bulb diameter in units of eighths of an inch; and A60 in the E.U. and elsewhere, where the number 60 represents the major bulb diameter in mm. Since $^{19}/_{8} = 2.375$ in ≈ 60 mm, the diameters of the U.S. A19 and the E.U. A60 bulbs are nearly the same. A-series bulbs with various other diameters are also available.

Operating Characteristics. The operating characteristics of an LED lamp, generally indicated on its packaging, include its electrical drive power P_{EL} (W or kWh/1000 h), luminous flux P_{V} (lm), wall-plug luminous efficacy η_{WPE} (lm/W), viewing angle $2\theta_{1/2}$ (°), correlated color temperature (CCT) T_c (K), color rendering index (CRI), lifespan (h), and whether or not the device is dimmable and is suitable for outdoor use. Lifespans typically range from 15000 to 50000 hours. The book locations where these parameters are defined is provided in Table 11.4-1.

The electrical power consumed by an incandescent lamp of equivalent luminous flux is also sometimes stated. Moreover, the parameter R_a is sometimes substituted for the CRI; R_a is computed as the average CRI over Munsell color samples R1–R8, whereas the CRI is the average over the full set R1–R15 (Sec. 9.9 and Fig. 9.9-1). The value of the wall-plug luminous efficacy η_{WPE} (lm/W) for an LED source determines its E.U. energy class, as specified in Table 11.4-2.

Table 11.4-1 LED lamp operating parameters and definitions.

PARAMETER	P_{EL} (W)	P_{V} (lm)	η_{WPE} $(\frac{\text{lm}}{\text{W}})$	$2\theta_{1/2}$ (°)	T_c (K)	CRI
DEFINITION	Eq. (7.1-13)	Eq. (8.8-1)	Eq. (8.9-4)	Fig. 7.2-2	Sec. 9.8	Sec. 9.9

European Union Lighting Energy Classes. In an effort designed to help consumers make informed choices when purchasing LED lights and lamps, the European Union introduced a new energy labeling system for light sources in 2021. As indicated in Table 11.4-2, a lighting device is assigned to one of seven energy classes that stretch from "A" to "G," depending on the value of its wall-plug luminous efficacy η_{WPE} (lm/W). A-class and G-class devices have the highest and lowest values of WPE and are therefore the most and least energy-efficient, respectively.

Table 11.4-2 Lighting energy classes established by the European Union in 2021.

CLASS	A	B	C	D	E	F	G
η_{WPE} $(\frac{\text{lm}}{\text{W}})$	$\geqslant 210$	185–209	160–184	135–159	110–134	85–109	$\leqslant 84$

Evolution of the White LED Retrofit Lamp

The evolution of the retrofit lamp over the time period from 2005 to 2024 is illustrated in Fig. 11.4-1. Early devices, such as the spot lamp pictured in Fig. 11.4-1(a), consisted of collections of dual in-line (DIP) PCLEDs. The Philips L-Prize lamp depicted in Fig. 11.4-1(b) was a landmark innovation in the world of lighting. More modern devices, such as the one portrayed in Fig. 11.4-1(c), rely on assemblies of surface-mounted devices (SMDs). Devices with the highest wall-plug luminous efficacies, such as that displayed in Fig. 11.4-1(d), make use of LED filaments. These lamps are described in further detail below.

Figure 11.4-1 Evolution of the white LED retrofit lamp from 2004 to 2024. LED lamps, including those shown here, generally contain multiple LEDs. (a) Cool-white DIP LED retrofit spot lamp (ca. 2005). (b) Philips warm-white L-prize LED retrofit lamp (ca. 2011). (c) Contemporary warm-white SMD LED retrofit lamp (ca. 2020). (d) Contemporary neutral-white, ultra-efficient, LED-filament retrofit lamp (ca. 2024).

Early DIP LED Retrofit Lamps. The early (ca. 2005) LED retrofit spot lamp portrayed in Fig. 11.4-1(a) comprised 38 LEDs facing in a common direction, each in a traditional dual in-line package (DIP) [Fig. 10.5-1(a)]. This directional lamp consumed 1.4 W of electrical power and generated cool-white light with a wall-plug luminous efficacy $\eta_{\mathrm{WPE}} \approx 50$ lm/W. It was not dimmable.

Philips L-Prize LED Retrofit Lamp. A substantial advance in retrofit lamps was attained with the development of the Philips L-Prize lamp in 2011 [Fig. 11.4-1(b)]. Its introduction marked a significant milestone in the evolution of home lighting technology. The L-Prize lamp, which operates as a hybrid device and emits warm-white light, is discussed in Sec. 11.5.

White SMD Retrofit Lamps. Warm-white LED retrofit lamps, such as the one displayed in Fig. 11.4-1(c) (ca. 2020), are designed to be morphologically similar to their incandescent counterparts. Omnidirectional and directional versions of these lamps are examined in Examples 11.4-1 and 11.4-2, respectively.

White LED-Filament Lamps. White LED-filament lamps, such as the MASTER UltraEfficient LED bulb portrayed in Fig. 11.4-1(d) (ca. 2024), are designed to closely resemble their incandescent counterparts while providing excellent efficiency. As reported in Sec. 10.7, and displayed in Fig. 10.7-1, the light from these devices is generated by many small, unpackaged, blue LED chips mounted on transparent filaments and embedded in phosphor. With suitable juxtaposition of the filaments, the chip-on-glass (COG) architecture offers a uniform radiation pattern over a large solid angle.

The use of a multiple chips that operate at low current, rather than a single chip operating at high current, mitigates the deleterious effects of efficiency droop and increases the lifespan of the device. Filling the bulb with a gas such as He provides

thermal management and eliminates the necessity for heat-sinking. Because the individual InGaN MQWLED chips operate at low currents and low optical powers, under some circumstances they might be able to be replaced with white quantum-dot LEDs (WQLEDs), white organic LEDs (WOLEDs), or white perovskite LEDs (PeWLEDs); the properties of these three classes of devices are compared in Table 7.6-1.

While the LED-filament lamp was not well-accepted when it was first introduced in 2008, these devices offer A-class performance and are now ubiquitous. The details relating to the operation of one such LED lamp are offered in Example 11.4-3.

Retrofit-Lamp Variants. Retrofit lamps configured in the following forms, some with build-in optics, are also available:

- Globes
- Candles
- Highbays
- Lowbays
- Spots
- Corncobs
- Reflectors
- Post-Tops
- UFOs
- Decoratives
- Designers
- Specialties

LED retrofits are also available for fluorescent lamps; however, the sale of fluorescents is now widely prohibited by law because they contain mercury, which is environmentally hazardous.

EXAMPLE 11.4-1. ***Contemporary Omnidirectional White SMD Retrofit Lamp.*** The interior of a lamp such as that displayed in Fig. 11.4-1(c) can be partitioned into four chambers by reflective metallic dividers, each serving as a substrate for two discrete surface-mounted LEDs such as those displayed in Fig. 10.5-1(b), so that the lamp contains eight SMD LEDs. This lamp is both omnidirectional and dimmable, and has a plastic shroud that is vented at both the top and bottom, enabling it to be cooled by convection. It consumes $P_{\text{EL}} = 10$ W of electrical power and generates metameric warm-white light with a luminous flux $P_{\text{V}} = 815$ lm, corresponding to just over 100 lm per LED and an overall wall-plug luminous efficacy $\eta_{\text{WPE}} = 82$ lm/W. Since the equivalent incandescent lamp consumes 60 W of electrical power, this lamp uses $\approx {}^{1}/_{6}$ of the energy of an equivalent incandescent that generates the same amount of light. The emitted light has a CCT of $T_c = 2700$ K and a CRI of 90. The lamp has a lifespan of 25000 h. LED retrofit lamps such as these are available with a broad range of bulb shapes, lamp bases, luminous fluxes, CCTs, and CRIs. The luminous flux can be enhanced by making use of chip-on-board (COB) devices (Sec. 10.8) in place of SMDs.

EXAMPLE 11.4-2. ***Directional White SMD Retrofit Lamp.*** A warm-white LED retrofit lamp with a different design from that considered in Example 11.4-1 is depicted in the cutaway diagram presented in Fig. 11.4-2. This lamp consumes an electrical power $P_{\text{EL}} \approx 10$ W and produces the same level and quality of light as an incandescent bulb that consumes an electrical power of 100 W. It is dimmable and produces metameric warm-white light with a luminous flux $P_{\text{V}} = 1500$ lm, so that the overall wall-plug luminous efficacy is $\eta_{\text{WPE}} = 150$ lm/W and each of its 10 discrete LEDs produces 150 lm. The emitted light exhibits a correlated color temperature $T_c \approx 2700$ K, a color rendering index CRI = 90, and the lamp has a lifespan of 25000 hours. LED lamps can be dimmed either by reducing the applied voltage or by using a pulse-width modulated current driver, as described in Sec. 7.4.

EXAMPLE 11.4-3. ***Ultra-Efficient White LED-Filament Retrofit Lamp.*** The A-class LED-filament retrofit lamp displayed in Fig. 11.4-1(d) (ca. 2024) closely resembles its incandescent counterpart and has roughly the same weight. It makes use of a Philips phosphor blend whose spectral density is displayed in Fig. 10.4-2 (green curve). This device consumes 4.0 W of electrical power and generates metameric warm-white light with a luminous flux $P_{\text{V}} = 840$ lm, near-ideal wall-plug luminous efficacy $\eta_{\text{WPE}} = 210$ lm/W, correlated color temperature $T_c = 3000$ K, and color rendering index CRI = 85. This lamp has the classic A60 shape, a clear-glass bulb, and a lifespan of 50000 hours. A-class Philips white retrofit E27 lamps such as this are available with many different characteristics, but all have $\eta_{\text{WPE}} = 210$ lm/W (energy class A), 50000 h lifespan, CRI = 85, and all are devoid of Hg and other hazardous substances (`https://www.lighting.philips.com/home`):

Figure 11.4-2 Cutaway view of a directional white LED retrofit lamp (ca. 2020). The bulb contains an array of ten SMDs along with a heat sink, a diffusing globe, and an E26 Edison screw base. The circuitry incorporated within the lamp serves as a **built-in driver**. A collection of LEDs connected in series can be driven by a DC current obtained by rectifying line-voltage AC with diodes and capacitors. The LEDs can also be directly driven by an AC current, emitting light every other half cycle. Alternatively, they can be wired as two antiparallel strands of series-connected LEDs, resulting in half of them emitting light every half cycle.

- $P_{EL} = 2.3$, 4.0, 5.2, or 7.3 W.
- $P_V = 485$, 840, 1095, or 1535 lm.
- $T_c = 2700$, 3000, or 4000 K.
- Dimmable and nondimmable versions.
- Various bulb shapes and bases.
- Clear or frosted bulb finishes.

Retrofit LED Lamps of Adjustable Color Temperature

Lamps incorporating PCLEDs that employ different phosphors can generate metameric white light with adjustable CCT and luminance. The *Philips Hue White Ambiance Smart Bulb*, introduced in 2017, incorporates two types of PCLEDs, warm-white devices with a CCT of 2200 K and cool-white devices with a CCT of 6200 K. As illustrated in Example 11.4-4, this lamp delivers white light with 50000 gradations between warm-and-cool white light. The CCT is controllable via a mechanical slider or wirelessly via Bluetooth or Zigbee (the latter protocol is commonly used in smart home applications).

The theoretical underpinnings of why CCT tunability can be implemented in a device such as this were established in Example 9.6-1 and quantified in (9.6-4). The chromaticity coordinates of the superposed light, which are suitably weighted linear combinations of the individual chromaticity coordinates for the constituent warm- and cool-white PCLEDs, fall along the straight line connecting them. Figure 9.7-2 reveals that the Planckian locus between 2200 K and 6200 K can indeed be approximated by a straight line.

EXAMPLE 11.4-4. ***Philips Hue Adjustable Color Temperature Retrofit Lamp.*** The *Philips Hue White Ambiance Smart Bulb* contains 16 warm-white (2200 K) and 16 cool-white (6200 K) discrete PCLEDs. It emits white light with a correlated color temperature that is adjustable over the range 2200–6500 K, which accommodates 50000 shades of white. Figure 11.4-3 illustrates how changing the relative proportions of the light emitted from the two classes of LEDs modifies the color temperature. Neutral-white light at 4200 K, for example, is generated by fully energizing both the 2200-K and 6200-K PCLEDs.

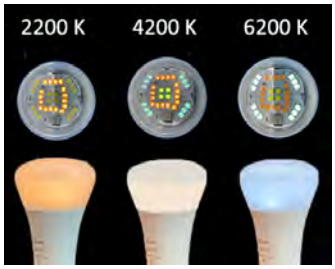


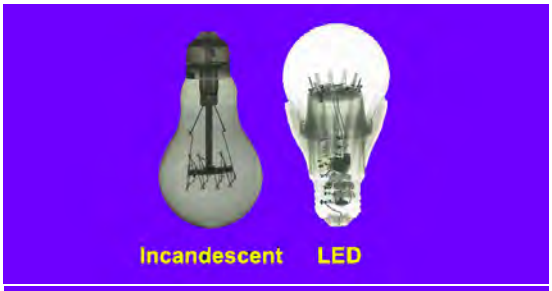
Figure 11.4-3 Images of the *Philips Hue White Ambiance Smart Bulb* operating at three values of the correlated color temperature: 2200 K, 4200 K, and 6200 K. The middle row of the figure portrays the activated LEDs when the diffusing bulb is removed; the bottom row displays the light emitted from the diffusing bulb itself. Warm-, neutral-, and cool-white light appear yellowish, whitish, and bluish, respectively. The salient physical and operating characteristics of this lamp are summarized below.

Summary of the salient physical and operating characteristics of the *Philips White Ambiance* bulb:

- Form factor: A19. Screw base: E26. Height: 112 mm. Diameter: 62 mm. Weight: 78 g.
- Input voltage: 110–130 V. Operating temperature: -20 °C to $+45$ °C. Lifespan: 25000 h.
- $P_{\mathrm{EL}} = 10.5$ W; Equivalent incandescent wattage: 75 W; Energy savings: $^{64.5}/_{75} = 86\%$.
- $P_{\mathrm{V}} = 1055$ lm @4000 K; $P_{\mathrm{V}} = 806$ lm @2700 K. Dimmable. CRI $= 80$.
- $\eta_{\mathrm{WPE}} = 100$ lm/W @4000 K; $\eta_{\mathrm{WPE}} = 77$ lm/W @2700 K.
- $\eta_{\mathrm{WPC}} = 0.15$ @4000 K; $\eta_{\mathrm{WPC}} = 0.11$ @2700 K.
- Continuously tunable CCT from 2200 to 6500 K (50000 shades of warm-to-cool white).
- Bluetooth and Zigbee communications protocols.

Tungsten Incandescents Supplanted by PCLED Retrofits

The superiority of the phosphor-conversion retrofit lamp relative to the traditional tungsten-filament incandescent lamp is artistically highlighted in the depiction below:



The royal-blue background and yellow font used in this illustration represent, respectively, the colors of the LED light and the phosphor used to generate metameric cool-white light for LED lighting. The development of the white phosphor-conversion LED at the turn of the 21st century upended traditional incandescent and fluorescent lighting. The X-ray image of a tungsten-filament incandescent lamp appears overturned at left, while that of a modern LED retrofit lamp stands upright at right.

11.5 HYBRID LEDS

The *third method* for generating metameric white light (of the three codified in Sec. 10.1) is a hybrid of the first two: It joins the phosphor-conversion approach discussed in Secs. 10.5 and 10.6 with the additive color-mixing approach elaborated in Sec. 11.3. Hybrid LEDs comprise two or more dies of different colors (e.g., blue and red) together with one or more phosphors. They can be implemented in a wide variety of configurations and are suitable for the generation of light with a broad palette of colors.

In this section, we consider two hybrid devices. The first is the *Philips L-Prize LED Retrofit Lamp* developed in 2011 (and retired in 2013), which is often thought of as the "poster child" and progenitor of the hybrid approach. The second is the *Philips Hue White and Color Ambiance Smart Bulb*, which generates light whose color can be tuned over the full gamut of human vision and remains in service as of 2024. This lamp was introduced in 2012, reissued in 2015 with an enhanced color palette, and enabled with Bluetooth in 2019.

Although hybrid devices were initially developed for the efficient generation of white light, they were no match for PCLEDs in doing so. Hybrids instead found their place as sources of light of variable hue. Since their introduction, Philips' hybrid LEDs have offered innovative lighting solutions that have coupled highly efficient LED lighting with digital control technology.

Philips L-Prize LED Retrofit Lamp

Images of the interior and exterior of the Philips L-prize white retrofit lamp are displayed in Figs. 11.5-1(a) and 11.4-1(b), respectively. Its development in 2011 garnered for Philips Lighting North America (now Signify) the US$10-million *L Prize* established by the U.S. Department of Energy in 2008. More formally called the *Bright Tomorrow Lighting Prize*, this award was designed to "spur lighting manufacturers to develop high-quality, high-efficiency LED lighting products to replace the common incandescent light bulb."

Specifically, the prize sought to foster the development of a lamp with multiple LEDs that 1) consumed modest electrical power ($P_{\text{EL}} = 10$ W), 2) generated warm-white light ($T_c = 2700$ K) with a luminous flux equivalent to that of a traditional 60-W incandescent lamp ($P_{\text{V}} = 800$ lm), and 3) had a long lifespan (25000 h). The reduction of electrical-power consumption from 60 W to 10 W corresponded to a power savings of ${}^{50}\!/_{60} = 83\%$. As described in Example 11.5-1, Philips met these goals by developing a hybrid configuration that relied on blue LEDs illuminating remote yellowish-green phosphor panels, along with red LEDs. The L-Prize Lamp has a notable place in the history of LED lighting: it paved the way for the widespread adoption of LED technology and set new standards in the lighting industry that fostered substantial energy savings.

EXAMPLE 11.5-1. ***Operation and Spectrum of the Philips L-Prize Retrofit Lamp.*** The 2011 Philips L-Prize Retrofit Lamp has three identical sections, one of which is pictured in Fig. 11.5-1(a). Six individual surface-mounted LEDs are visible, three blue and three red. All were operated below their maximum power specifications to minimize efficiency droop. Each section is capped by a plastic cover that contains the remotely located yellow-green phosphor. An exterior view of the intact lamp is portrayed in Fig. 11.4-1(b).

The spectral density of the light emitted by this lamp, displayed in Fig. 11.5-1(b), comprises three peaks: a narrow peak near 450 nm associated with the blue LED light, a broad peak in the vicinity of 550 nm representing the yellow-green photoluminescence from the phosphor, and a narrow peak near 625 nm ($\Delta\lambda_{\text{FWHM}} \approx 20$ nm) associated with the red AlInGaP LED light. The shape of the spectrum is reminiscent of that of Philips' modern KSF-containing narrowband-red phosphor blend displayed in Fig. 10.4-2 (purple curve), which generates warm-white light at $T_c = 3000$ K when illuminated by blue LED light.

The color gamut supported by the L-Prize Lamp is similar to that of the gamut triangle presented in Fig. 10.6-1(b) for a warm-white PCLED, but here the triangle's vertices are defined by the wavelengths of the blue LED light, the yellow-green photoluminescence, and the red LED light (rather than the red phosphor). As always, any color within the triangle can be generated by mixing the three colors at its vertices in appropriate proportions. Since the gamut triangle includes the region of the Planckian locus encompassing 2700–3500 K, the Philips L-prize hybrid device was able to generate metameric warm-white light.

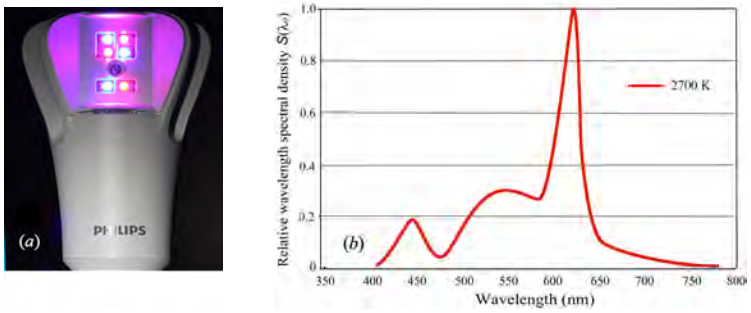


Figure 11.5-1 (a) Interior view of a compartment of the Philips L-prize Retrofit Lamp. Each of the three compartments contains three blue and three red LEDs, and is enclosed by a plastic cap containing yellow-green phosphor; the lamp as a whole contains 18 LEDs. An exterior view of the intact lamp is displayed in Fig. 11.4-1(b). Drawing 10 W of electrical power, this lamp generates metameric warm-white light ($T_c = 2700$ K) with a luminous flux $P_V = 940$ lm, which corresponds to 52 lm per LED and an overall wall-plug luminous efficacy $\eta_{WPE} = {}^{940}/_{10} = 94$ lm/W. (b) The spectral density of the light emitted by the L-prize lamp. Three peaks are evident, with approximate wavelengths 450 nm (blue), 550 nm (yellow-green), and 625 nm (red). These peaks represent, respectively: 1) blue light from the InGaN LEDs, 2) yellow-green photoluminescence from the phosphor when excited by the blue LED light, and 3) red light from the AlInGaP LEDs.

After entering the marketplace in 2012, the L-Prize Lamp was discontinued in 2013 in favor of other models that could be produced at lower cost. The performance of the warm-white PCLEDs of today, which incorporate narrowband-red phosphor blends, is superior to that of the more complex, warm-white hybrid devices of yesterday. For example, the wall-plug luminous efficacy of the Philips L-Prize lamp was $\eta_{WPE} = 94$ lm/W, as illustrated in Example 11.5-1, while that of a modern LED-filament lamp containing multiple warm-white PCLEDs is $\eta_{WPE} = 210$ lm/W, as demonstrated in Example 11.4-3.

Philips Hue White and Color Ambiance LED Retrofit Lamp

Lamps that make use of additive color mixing can deliver essentially any color or shade of white. The *Philips Hue White and Color Ambiance Smart Bulb*, introduced in 2012, is the progenitor of such devices and has a devoted audience. This lamp operates either in hue/saturation or color-temperature mode, offering light of variable hue with sixteen million possible colors, or metameric white light with a CCT that can be continuously tuned over the range 2000–6500 K, respectively. Like the 2017 Philips Hue Adjustable Color Temperature Retrofit Lamp analyzed in Example 11.4-4, it has wireless connectivity and can be controlled via a mobile-phone app. This device is deconstructed in Example 11.5-2.

EXAMPLE 11.5-2. ***Philips Hue Adjustable-Color Retrofit Lamp Deconstructed.*** The *Philips Hue White and Color Ambiance Smart Bulb* introduced in 2012 contains an LED board, a power board, and a logic board:

- The aluminum printed-circuit LED board, displayed in Fig. 11.5-2(a), supports the LEDs and implements thermal management. It contains 16 discrete LEDs: two red (R1, R2); one blue (B); one green (G); four surrounding warm-white PCLEDs (W1–W4); and eight cool-white PCLEDs in the outer ring (C1–C8).
- The power board, pictured in Fig. 11.5-2(b), rectifies the incoming AC voltage, and regulates, surge-protects, and distributes the resulting low-voltage DC.
- The logic board, depicted in Fig. 11.5-2(c), contains the LED drivers, implements power management, wireless communications/security, and firmware storage/updating.

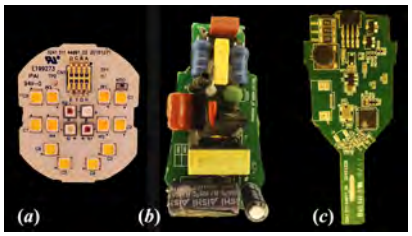


Figure 11.5-2 Components of the Philips Hue White and Color Ambiance Bulb. (*a*) The LED board supports four centrally located colored LEDs, surrounded by four warm-white PCLEDs and, in the outer ring, by eight cool-white PCLEDs. (*b*) The power board distributes electrical power to the logic and LED boards. (*c*) The logic board serves as the central processing unit. The salient physical and operating characteristics of the lamp are summarized below.

Summary of salient physical and operating characteristics of *Philips White and Color Ambiance* bulb:

- Form factor: A19. Screw base: E26. Height: 112 mm. Diameter: 62 mm. Weight: 70 g.
- Input voltage: 110–130 V. Operating temperature: -20 °C to $+45$ °C. Lifespan: 25000 h.
- $P_{\text{EL}} = 10.5$ W; Equivalent incandescent wattage: 75 W; Energy savings: $^{64.5}/_{75} = 86\%$.
- $P_{\text{V}} = 1055$ lm @4000 K; $P_{\text{V}} = 806$ lm @2700 K. Dimmable. CRI $= 80$.
- $\eta_{\text{WPE}} = 100$ lm/W @4000 K; $\eta_{\text{WPE}} = 77$ lm/W @2700 K.
- $\eta_{\text{WPC}} = 0.15$ @4000 K; $\eta_{\text{WPC}} = 0.11$ @2700 K.
- Dual inner and outer diffusers. Bluetooth and Zigbee communications protocols.
- Sixteen million colors and white CCT that is continuously tunable from 2000 to 6500 K.

11.6 LED LUMINAIRES

As discussed in the previous chapter, and in earlier sections of this chapter, **LED lamps** are devices that generally contain multiple LEDs and generate light. **LED luminaires**, on the other hand, are self-contained lighting units with housings that incorporate one or more LED lamps or LED modules, along with the ancillary components necessary for their operation. Called a **light fixture** or **light fitting** in common parlance, a luminaire often contains an integrated power supply and the means for controlling it. The housing serves to hold the lamps in position and to secure the optical components that spatially shape the emitted light and guide it to the exterior of the housing.

The design of a luminaire establishes its esthetics and plays a central role in determining the distribution and directions of the light emanating from it. The wall-sconce illustrated in the center panel of Fig. 11.3-1, for example, is a luminaire. The lamps within a luminaire can operate on the basis of phosphor-conversion or color-mixing processes, or both. LED-filament and chip-on-board lamps are widely used in luminaires because of the high values of luminous flux and luminous efficacy they offer.

Types of Luminaires. An extensive variety of LED luminaires of different types and specifications are available. Luminaire designs vary widely and assume many configurations, including:

- Downlights
- Wallwashers
- Floodlights
- Troffers
- Panels
- Cove lights
- Projectors
- Tracklights
- Chandeliers
- Pendants
- Sconces
- Lanterns
- Streetlights
- Roadlights
- Table lamps
- Floor lamps
- Path Lights
- Bollards
- Columns
- Tubes
- Bars
- Strips
- Ropes
- Strings

A number of representative luminaires are pictured in Fig. 11.6-1.

Figure 11.6-1(a) displays an indoor recessed ceiling luminaire comprising 16 white PCLEDs, each capped with a molded-plastic optical element similar to that depicted in Fig. 1.4-3. As exemplars of catadioptric optics, these devices make use of both total internal reflection and refraction to guide the light from the LED chip where it is generated to the exterior of the housing. Figure 11.6-1(b) displays a portion of an illuminated outdoor garden using a flexible smart lightstrip that emits diffused light with a color can be chosen at will. Figure 11.6-1(c) depicts an outdoor luminaire designed for use in parks and other public spaces; a luminaire similar to this is discussed in Example 11.6-1. Finally, Fig. 11.6-1(d) displays a luminaire styled as a table lamp that incorporates a smart bulb, such as that highlighted in Example 11.5-2. Customized 3D luminaires can also be printed.

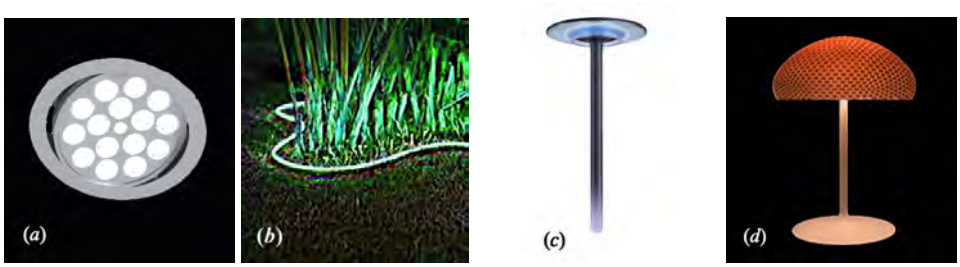


Figure 11.6-1 Luminaires. (a) A recessed ceiling LED luminaire comprising a collection of metameric-white PCLEDs. (b) A smart LED light strip illuminating a portion of an outdoor garden. (c) An outdoor post-top LED luminaire used for illumination in a city park. (d) A luminaire styled as a table lamp that incorporates a retrofit lamp of adjustable color.

As a further indication of some of the kinds of LED luminaires that are available, a compilation of the lighting equipment required to outfit a small theater (along with the estimated costs of the various components) is provided in Example 11.6-2.

EXAMPLE 11.6-1. ***Operation of an Outdoor Post-Top LED Luminaire.*** Post-top LED luminaires are outdoor lighting fixtures designed to be mounted atop a post or pole. Commonly used for illuminating streets, parks, walkways, and parking lots, post-top luminaires are known for their functional, decorative, and aesthetic attributes. A typical post-top luminaire might draw $P_{\text{EL}} = 50$ W of electrical power and generate metameric neutral-white light with: 1) a luminous flux $P_{\text{V}} = 5000$ lm, 2) a luminaire luminous efficacy $\eta_{\text{LUM}} = 100$ lm/W, 3) a correlated color temperature $T_c \approx 4000$ K, 4) a color rendering index CRI ≈ 70, and 5) a lifespan ≈ 100000 h. Luminaires such as these may be customized to draw between 15 and 115 W of electrical power, generate metameric white light with a CCT in the range 3000–4000 K and a CRI between 70 and 80; and exhibit a luminous flux and luminous efficacy in the range $P_{\text{V}} = 1200$–9800 lm and $\eta_{\text{LUM}} = 90$–105 lm/W, respectively.

EXAMPLE 11.6-2. ***LED Lighting and Ancillary Equipment for a Small Theater.*** A multipurpose theater that holds 275 people and serves a town of roughly 40000 persons requires the following lighting equipment (approximate costs are also indicated):

- LED Wash Lights for Stage Lighting (6–8 units): to create even illumination and color washes. Approximately US$1000 per unit.
- LED Fresnel Spotlights for Stage Lighting (4–6 units): to create focused light for key and back-lighting. Approximately US$4000 per unit.
- LED Ellipsoidal Spotlights for Stage Lighting (2–4 units): to create sharp beams for special effects. Approximately US$3000 per unit.
- LED Follow Spotlights for Onstage Performer Tracking (1–2 units): for precision, dynamic lighting, especially in musical and dance performances. Approximately US$15000 per unit.

- LED Cyclorama Lights for Stage Lighting: to provide even illumination across backdrops and cycloramas. Approximately US$3000.
- Lighting Console: a control system capable of handling complex cues and programming. Approximately US$12000.
- Cabling, Rigging, and Trusses for Hanging Lights and Scenery: DMX cables, power cables, connectors, safety cables, clamps, and trusses. Approximately US$20000.
- LED Dimmer and Relay Rack Backstage: for control of stage lighting power levels. Approximately US$9000.
- LED House Lights (25 units). Approximately US$8000.
- LED Exit Signs (20 units). Approximately US$1000.

Not included in the tabulation provided above are architectural lighting for the lobby, aisles, and other non-stage areas; emergency lighting; hazers and fog machines for creating special lighting effects; software; and sound systems. Design and installation costs are also additional.

Luminaire Wall-Plug Luminous Efficacy (LUM) and Luminous Efficiency (LUC). Echoing the definition of the wall-plug luminous efficacy for lamps provided in (8.9-4), the **luminaire wall-plug luminous efficacy** $\eta_{\mathtt{LUM}}$ is defined as

$$\eta_{\mathtt{LUM}} = \frac{P_{\mathtt{V}}}{P_{\mathtt{EL}}}\,, \tag{11.6-1}$$

Luminaire Wall-Plug Luminous Efficacy (lm/W or LPW)

where $P_{\mathtt{V}}$ is now the output luminous flux of the luminaire and $P_{\mathtt{EL}}$ is the electrical drive power provided to it. The units of this quantity are again lm/W, but they are sometimes written as luminaire-lm/W or luminaire-lm/circuit-W.

The efficacy of a luminaire can be 30% or more below that of its constituent lamp(s), as a consequence of driver, thermal, and optical/fixture losses:

$$\eta_{\mathtt{LUM}} \lesssim 0.7\,\eta_{\mathtt{WPE}}. \tag{11.6-2}$$

One factor that contributes to this loss is the legacy form factor of most LED bulbs, which is suboptimal for dissipating heat as well as for generating light that can be efficiently redirected into a desired spatial configuration. Nevertheless, optical/fixture losses can often be reduced below 10% for well-designed luminaires.

Similarly mimicking the definition of wall-plug luminous efficiency set forth in (8.9-9), the **luminaire wall-plug luminous efficiency** $\eta_{\mathtt{LUC}}$ is defined by normalizing the luminaire wall-plug luminous efficacy to its maximum possible value of 683, i.e.,

$$\eta_{\mathtt{LUC}} = \eta_{\mathtt{LUM}}/683\,. \tag{11.6-3}$$

This dimensionless quantity has a value that lies between zero and unity.

11.7 OLED LIGHT PANELS

OLED light panels are large-area light sources fashioned from organic light-emitting diodes (OLEDs), which are efficient generators of electroluminescence in the blue, green, and red. **White organic light-emitting diodes** (**WOLEDs**), fabricated in a manner such as that prescribed in Fig. 7.6-1, generate white light via additive color mixing, as depicted in Fig. 11.3-1. They have nearly unity internal quantum efficiency and

provide excellent color rendition. A white OLED light panel comprises a single, broad-area, serially stacked OLED, such as that displayed in Fig. 7.6-1(*b*).

OLED light panels offer large-area diffuse lighting that is devoid of glare, without the necessity for supplementary optics, so that they can be located in close proximity to objects and human users. This is in contrast to most discrete LEDs, which emit as point sources. OLED panels also emit little heat so they can be used to illuminate delicate objects and sensitive materials, such as artwork and foodstuffs. When fabricated on transparent plastic substrates, OLED light panels are lightweight, thin, and flexible. They can therefore be fashioned into bendable or rollable sheets, and can be molded into unique shapes that serve as 3D diffuse sources of metameric white light. Indeed, although typically designed to emit white light, OLED panels can be configured to emit light of any color, and offer dynamic color tuning. The use of luminous surfaces in artificial lighting offers new horizons for reshaping the manner in which humans interact with light.

White OLED light panels can be fabricated on substrates that are rigid or flexible. A flexible white OLED light patch is sketched in Fig. 11.7-1(*a*). The operating luminance of panels such as these can be adjusted by modifying the duty cycle of a pulse-width-modulation (PWM) driver or by altering the applied DC current. As with other LED lighting sources, increasing the luminance results in an increase in panel brightness, but this comes at the expense of an increase in surface temperature, as well as a decrease in luminous efficacy and lifespan. The specifications and operating properties of flexible and rigid OLED light panels are considered in Example 11.7-1, while the spatial distribution and spectral density of the emitted light are described in Examples 11.7-2 and 11.7-3, respectively.

Figure 11.7-1 (*a*) White organic light-emitting diode (OLED) light panels can take the form of thin, lightweight, and bendable panels when fabricated on transparent plastic substrates. They can also take the form of rigid panels when fabricated on thin sheets of glass. Such panels offer large-area homogeneous illumination with little glare, and find use in LED lighting. (*b*) Conceptualization of a flexible transparent OLED (TOLED) light panel. (*c*) A captivating OLED luminaire constructed from flexible white OLED light panels. (Projet de fin d'étude réalisé en 2010 à l'ESDMAA, Institut du Design à Yzeure (France) par Elodie Saugues, Designer Produits Française à Stotzheim (France). Cette Lampe propose un nouvel usage de la lumière grâce à la technologie OLED. Une lumière qu'on manipule, qu'on emporte avec soi, grâce à ces trois feuilles autonomes et nomades que l'on peut enrouler ou laisser à plat. Les feuilles peuvent être regrouper grâce à un socle qui permet de les recharger par induction et de créer une lampe autoporté avec ces trois chandelles. Design by Elodie Saugues, reproduced with permission.)

EXAMPLE 11.7-1. ***Flexible and Rigid OLED Light Panels.*** A number of the operating properties of flexible and rigid OLED light panels are specified and their performance is compared with that of MQWLEDs.

- *Flexible OLED light panel.* OLED light panels are available in a broad range of sizes, shapes, luminances, and color temperatures. The 1.3-W flexible light panel pictured in Fig. 11.7-1(*a*), manufactured by LG Chem (Model P6BA30), is fabricated on a transparent polyimide substrate.

It has a thickness of $^1/_4$ mm and emits light over a region with dimensions 18.8 cm $\times$ 4.1 cm, corresponding to an area $A = 0.0077\,\mathrm{m}^2$. This particular light panel generates metameric warm-white light with a luminous flux $P_\mathrm{V} = 75$ lm, luminance $L_\mathrm{V} = 3000$ cd/m^2, and a luminous efficacy $\eta_\mathrm{WPE} = 58$ lm/W. The light has a correlated color temperature $T_c = 3000$ K and a color rendering index CRI = 90, and the panel has a lifespan of 40000 hours.

- *Rigid OLED light panel.* A larger 13.5-W OLED light panel by LG Chem, with dimensions 32 cm $\times$ 32 cm $\times$ 1 mm, is fabricated on glass and is therefore not flexible. This source provides warm-white light with a luminous flux $P_\mathrm{V} = 800$ lm, luminance $L_\mathrm{V} = 3000$ cd/m^2, and a wall-plug luminous efficacy $\eta_\mathrm{WPE} = 59$ lm/W. It too has a correlated color temperature $T_c = 3000$ K, a color rendering index CRI = 90, and a lifespan of 40000 hours. OLED films are routinely fabricated on glass panels as large as 3×3 m^2, although panel size cannot be made arbitrarily large because of the necessity of maintaining constant luminance across the panel.
- *OLED light-panel performance.* In the current state of their development, it is therefore clear that OLED performance measures such as wall-plug luminous efficacy η_WPE fall well behind those of MQWLEDs. Their cost also remains relatively high.

EXAMPLE 11.7-2. ***Spatial Distribution of Light Emitted by an OLED Light Panel.*** The spatial light-emission pattern of an OLED panel is quite close to that of a Lambertian radiator. In accordance with Table 8.8-1 and (8.8-3), the luminous intensity and luminance of a radiator with a viewing angle $2\theta_{1/2} = 120°$, are $I_\mathrm{V} = P_\mathrm{V}/2\pi(1 - \cos\theta_{1/2}) = P_\mathrm{V}/\pi$ and $L_\mathrm{V} = P_\mathrm{V}/\pi A$, respectively. Minor deviations of the OLED spatial-emission radiation pattern from Lambertian behavior are accommodated by incorporating a factor called the **Lambertian ratio** R_L into the expression for the luminance, rendering it as $L_\mathrm{V} = P_\mathrm{V}/\pi R_\mathrm{L} A$. For the light panel depicted in Fig. 11.7-1(a), the accompanying data provides $R_\mathrm{L} = 1.12$, so that $L_\mathrm{V} \approx 2768$ cd/m^2, which is indeed close to the value of 3000 cd/m^2 specified by the manufacturer. This luminance value is within the optimal range for visual comfort, which is generally considered to be 2500–3000 cd/m^2.

EXAMPLE 11.7-3. ***Spectral Density of Light Emitted by an OLED Light Panel.*** The spectrum of the light emitted by a white OLED panel such as that displayed in Fig. 11.7-1(a) is depicted in Fig. 11.7-2. The spectral density comprises three peaks, at $\lambda_\mathrm{p} = 450$, 550, and 610 nm, associated with the blue-, green-, and red-emitting organic layers contained within the device, respectively. The OLED spectrum bears a resemblance to that presented in Fig. 11.3-3(a) for a color-mixing LED, indicating that the triangle in the OLED chromaticity diagram delineating the accessible range of colors would be similar to that displayed in Fig. 11.3-3(b). Hence, most colors visible to humans, including the entire range of whites, can be generated by OLED panels.

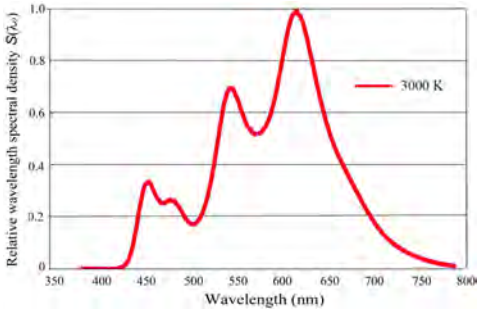


Figure 11.7-2 Spectrum associated with the light emitted by a white OLED light panel such as that portrayed in Fig. 11.7-1(a). The three spectral peaks at $\lambda_\mathrm{p} = 450$, 550, and 610 nm represent OLED emission in the blue, green, and red, respectively. This spectral density is not unlike that of the light emitted by the color-mixing LED displayed in Fig. 11.3-3(a).

EXAMPLE 11.7-4. ***Light Cloud Panorama Exhibition at the Merck Innovation Center.*** The Light Cloud exhibition at the Merck Innovation Center featured in Fig. 11.7-3 comprises four curved and concentrically placed bent-steel supports that host OLED panels whose locations are slightly shifted with respect to each other. The 576 light panels are controlled by 42 sensors and 12 audio channels that are distributed throughout the exhibition and respond to the movements and sounds of visitors. The experience, which is enriched by the mirrored ceiling and walls of the exhibition hall, is a mosaic of movement, light, and sound.

Figure 11.7-3 *Light Cloud Panorama* at the Merck Innovation Center in Darmstadt, Germany. Inaugurated in 2019, this sculpture comprises 576 double-sided OLED light panels by OLEDWorks. The exhibition was developed and realized by Tamschick Media+Space GmbH (Berlin) in collaboration with the Studio for Media Architectures *iart* (Basel), and it garnered the *German Design Award* in 2019. The installation responds to visitors' movements by interactively generating light and sound. An audiovisual presentation is available at `https://iart.ch/en/work/light-cloud`. (©Merck KGaA, Darmstadt Germany, reproduced with permission.)

Transparent OLED (also called **TOLED**) light panels are also available. Transparent when not in use, these thin structures find use in LED lighting; they can be embedded in the windows of a residence, for example, or in the sunroof of an automobile. In the display domain, they serve as novel dynamic information and advertising displays. While TOLEDs are typically fabricated as rigid structures, a flexible TOLED light strip is conceptualized in Fig. 11.7-1(*b*).

An OLED luminaire constructed of flexible OLED panels is displayed in Fig. 11.7-1(*c*).

11.8 SMART AND CONNECTED LED LIGHTING

The terms smart lighting and connected lighting are often used interchangeably, but there are subtle differences. **Smart lighting** refers to bulbs, fixtures, and systems that can be controlled electronically, typically through an app or smart home ecosystem. This includes basic features like dimming and turning lights on/off remotely, but can also extend to scheduling, color changing, and integration with other smart devices. **Connected lighting**, on the other hand, not only allows for control but also incorporates two-way communication between the lighting system and other devices or networks. This enables more advanced features like occupancy sensing, daylight harvesting, and data collection for optimizing energy usage or creating personalized lighting experiences. Connected lighting often plays a role in the broader **Internet of Things (IoT)** ecosystem, sharing data and interacting with other connected devices in the home or city. In essence, smart lighting focuses on control and convenience, while connected lighting adds the element of communication and data exchange, which opens the door to more sophisticated automation and integration within a broader smart environment.

Both smart and connected lighting can play a role in human-centric, plant-centric, and ecologically conscious lighting practices. **Human-centric lighting** prioritizes the well-being and comfort of people, often mimicking natural daylight patterns to support circadian rhythms and improve mood and productivity. Smart and connected lighting offer precise control over correlated color temperature and intensity, enabling users to create dynamic lighting schemes that adapt to different activities and times of day. **Plant-centric lighting** caters to the specific needs of plants, providing the right spectrum and intensity of light for optimal growth and health. Smart and connected systems

can be programmed to deliver customized light schedules based on the type of plant, ensuring they receive the necessary light for photosynthesis and overall well-being. Ecologically conscious lighting aims to minimize environmental impact and energy consumption. Smart and connected lighting offer features like occupancy sensing, daylight harvesting, and automated dimming directed toward minimizing energy use. Additionally, connected systems can collect data on usage patterns, allowing for further optimization and resource conservation.

LED Light-Control Processes

LED lighting is compatible with, and can implement, light-control processes such as the following:

- Wireless control.
- Dynamic control.
- Synchronization with external inputs, including music, voice, and kinetic motion.
- Synchronization with diurnal temporal, spectral, and spatial rhythyms.
- Lifi internet connectivity.
- Disinfection of air, water, surfaces, and objects with UVC (Fig. 2.4-1) light.

11.9 LED PERFORMANCE METRICS

Performance metrics for illumination have been established for all manner of light sources. The upper and lower portions of Table 11.9-1 report parameters for traditional and LED sources of white light, respectively (Sec. 11.1). Representative values for the luminous flux, wall-plug luminous efficacy (WPE), wall-plug luminous efficiency (WPC), correlated color temperature (CCT), and color rendering index (CRI) are provided. The table entries reveal that LED sources, whether PCLEDs, COBs, CMLEDs, retrofits, or luminaires, exhibit larger values of the WPE, WPC, and CRI than traditional sources, thereby confirming their superiority for white illumination. This advantage has been artistically highlighted for retrofit lamps in the figure presented on page 353.

From a practical perspective, phosphor-conversion packages are expected to attain values of the WPE as high as $\approx$ 255 lm/W in the near term, while color-mixing packages may ultimately reach their estimated theoretical limit of $\approx$ 325 lm/W.[†] The discrepancy between the WPE values for the two classes of LEDs arises principally from the limitations imposed on phosphor-conversion devices by the photoluminescence quantum yield $\eta_{\mathtt{PLQY}}$ $(\leqslant 1)$ and the complementary photoluminescence quantum defect $\overline{\eta}_{\mathtt{PLQD}}$ $(\leqslant 1)$, as discussed in Sec. 10.2. Since color-mixing devices do not make use of photoluminescence, they are not subject to these limitations.

Finally, we point out that the wall-plug luminous efficacies for both phosphor-conversion and color-mixing devices are proportional to the underlying chip power-conversion efficiency $\eta_{\mathtt{PCE}}$. As discussed in Sec. 7.1, in the current state of development of LED technology, the power-conversion efficiency is $\eta_{\mathtt{PCE}} \approx {}^{3}\!/_{4}$ ${}^{1}\!/_{2}$ and ${}^{1}\!/_{4}$ for blue, red, and green devices, respectively, which leaves room for improvement.

[†] P. M. Pattison, M. Hansen, and J. Y. Tsao, Lighting Efficacy: Status and Directions, *Comptes Rendus Physique*, vol. 19, pp. 134–145, 2018; M. Pattison, M. Hansen, N. Bardsley, G. D. Thomson, K. Gordon, A. Wilkerson, K. Lee, V. Nubbe, and S. Donnelly, *2022 Solid-State Lighting R&D Opportunities*, U.S. Department of Energy, DOI:10.2172/1862626, Technical Report EE-2542, February 2022.

Table 11.9-1 Representative illumination parameters for various (mostly white) traditional and LED sources. Successive columns display: electrical drive power P_{EL} (W), luminous flux P_{V} (lm), wall-plug luminous efficacy η_{WPE} (lm/W), wall-plug luminous efficiency η_{WPC}, correlated color temperature CCT (K), and color rendering index (CRI).

TRADITIONAL SOURCES[a]	P_{EL}^{b}	P_{V}^{b}	$\eta_{\mathrm{WPE}}^{b,c}$	η_{WPC}^{c}	CCT	CRI
Ideal blackbody radiation source at 6640 K[d]	–	–	96	0.14	6640	100
Red laser-pointer beam ($P_0 = 2$ mW at 650 nm)[e]	0.04	0.15	3.7	0.01	–	–
Carbon-arc lamp[f]	1000	15000	15	0.02	4500	75
Tungsten incandescent lamp (unfiltered)[f]	100	1500	15	0.02	2700	100
Tungsten halogen quartz incandescent lamp[f]	100	1700	17	0.02	3000	100
High-pressure mercury lamp[f]	100	5000	50	0.07	4000	50
High-pressure xenon lamp[f]	100	6000	60	0.09	6200	95
Compact fluorescent lamp[f]	25	1750	70	0.10	3000	80
Linear fluorescent lamp (1-inch diameter – T8)[f]	25	2250	90	0.13	3000	85
Ceramic metal-halide lamp[f]	250	22500	90	0.13	4200	90
High-pressure sodium lamp[f]	100	10000	100	0.15	2100	15
Low-pressure sodium lamp[f]	100	15000	150	0.22	1800	0
LED SOURCES[a]	P_{EL}^{b}	P_{V}^{b}	$\eta_{\mathrm{WPE}}^{b,c}$	η_{WPC}^{c}	CCT	CRI
Cool-white single-die discrete PCLED (Cree)[g]	2.9	490	167	0.24	5000	80
Cool-white single-die discrete PCLED (Cree)[g]	9.1	1150	126	0.18	5000	80
Warm-white single-die discrete PCLED (Cree)[h]	3.9	460	117	0.17	3000	90
Warm-white single-die discrete PCLED (Cree)[h]	19.7	1400	71	0.10	3000	90
Warm-white chip-on-board (COB) device (Cree)[i]	81.8	13350	163	0.24	3000	90
Warm-white chip-on-board (COB) device (Cree)[i]	235	28700	122	0.18	3000	90
Cool-white additive color-mixing CMLED (Cree)[j]	3.7	413	112	0.16	6000	90
Cool-white additive color-mixing CMLED (Cree)[j]	23.0	1375	60	0.09	6000	90
Warm-white omnidirectional retrofit LED lamp (Cree)[k]	10	815	82	0.12	2700	90
Warm-white directional retrofit LED lamp[l]	10	1500	150	0.22	2700	90
Warm-white LED-filament retrofit lamp (Philips)[m]	4.0	840	210	0.31	3000	85
Neutral-white LED-filament retrofit lamp (Philips)[m]	7.3	1535	210	0.31	4000	85
White adjustable-CCT LED lamp (Philips)[n]	10.5	1055	100	0.15	4000	80
Warm-white L-prize hybrid retrofit LED lamp (Philips)[o]	10	940	94	0.14	2700	92
Neutral-white outdoor LED luminaire (Philips)[p,q]	50	5000	100	0.15	4000	70
Warm-white flexible OLED light panel (LG Chem)[r]	1.3	75	58	0.08	3000	90
Warm-white rigid OLED light panel (LG Chem)[r]	13.5	800	59	0.09	3000	90

[a] Table entry values are rounded.
[b] The wall-plug luminous efficacy η_{WPE}, luminous flux P_{V}, and electrical drive power P_{EL} are related by (8.9-4).
[c] The wall-plug luminous efficiency and efficacy are related by $\eta_{\mathrm{WPC}} = \eta_{\mathrm{WPE}}/683$, in accordance with (8.9-9).
[d] The electrical drive power P_{EL} does not exist for light from a blackbody source, so in place of η_{WPE} and η_{WPC}, we report the luminous efficacy of radiation η_{LER} and the luminous efficiency of radiation $\eta_{\mathrm{LER}}/\eta_{\mathrm{LER}}^{\mathrm{MAX}} = \eta_{\mathrm{LER}}/683$, respectively. The luminous efficacy and efficiency of blackbody radiation are maximized at $T_c \approx 6640$ K.
[e] This source is fully characterized in Example 8.9-2. The CCT and CRI are generally used only for white light.
[f] The operating principles and limitations of these traditional technologies are summarized in Sec. 11.1.
[g] Table 10.5-1 and Examples 10.5-1–10.5-5. [h] Table 10.6-1 and Example 10.6-1. [i] Table 10.8-1 and Example 10.8-1. [j] Table 11.3-1 and Example 11.3-1. [k] Example 11.4-1 and Fig. 11.4-1(*c*). [l] Example 11.4-2 and Fig. 11.4-2. [m] Example 11.4-3 and Fig. 11.4-1(*d*). [n] Example 11.4-4 and Fig. 11.4-3. [o] Example 11.5-1 and Fig. 11.4-1(*b*). [p] Example 11.6-1 and Fig. 11.6-1(*c*). [q] The luminaire wall-plug luminous efficacy η_{LUM} provided in (11.6-1), and the luminaire wall-plug luminous efficiency $\eta_{\mathrm{LUC}} = \eta_{\mathrm{LUM}}/683$ set forth in (11.6-3), are reported in place of η_{WPE} and $\eta_{\mathrm{WPC}} = \eta_{\mathrm{WPE}}/683$, respectively. [r] Examples 11.7-1–11.7-3.

BIBLIOGRAPHY

LED Lighting

See also the bibliographies in Chapters 5, 6, 7, and 10.

M. Karlen and C. Spangler, *Lighting Design Basics*, Wiley, 4th ed. 2024.

E. F. Schubert, *Light-Emitting Diodes*, Google Books, 4th ed. 2023.

L. Annaniah, M. S. M. Saheed, and R. Jose, *LED Packaging Technologies*, Wiley–VCH, 2023.

S.-W. R. Lee, J. C. C. Lo, M. Tao, and H. Ye, *From LED to Solid State Lighting: Principles, Materials, Packaging, Characterization, and Applications*, Wiley–VCH, 2021.

C. Weisbuch, E. Spitz, and A. David, eds., LEDs: The New Revolution in Lighting/Les LED: la nouvelle révolution de l'éclairage (Special Issue), *Comptes Rendus Physique*, vol. 19, no. 3, pp. 85–182, 2018.

P. M. Pattison, M. Hansen, and J. Y. Tsao, Lighting Efficacy: Status and Directions, *Comptes Rendus Physique*, vol. 19, pp. 134–145, 2018.

R. Karlicek, C.-C. Sun, G. Zissis, and R. Ma, eds., *Handbook of Advanced Lighting Technology*, Springer Nature, 2017.

M. F. Gendre, G. Lister, Y. Liu, H. Schöpp, S. Franke, and S. A. Mucklejohn, Part VIII: Conventional Light Sources, in R. Karlicek, C.-C. Sun, G. Zissis, and R. Ma, eds., *Handbook of Advanced Lighting Technology*, Springer Nature, pp. 1012–1171, 2017.

R. Lenk and C. Lenk, *Practical Lighting Design With LEDs*, Wiley–VCH, 2017.

National Academies of Sciences, Engineering, and Medicine, Assessment of Solid-State Lighting: Phase Two, The National Academies Press, DOI:10.17226/24619, 2017.

G. Lozano, S. R. K. Rodriguez, M. A. Verschuuren, and J. Gómez Rivas, Metallic Nanostructures for Efficient LED Lighting, *Light: Science & Applications*, **5**, e16080; doi:10.1038/lsa.2016.80, 2016.

M. H. Crawford, J. J. Wierer, A. J. Fischer, G. T. Wang, D. D. Koleske, G. S. Subramania, M. E. Coltrin, J. Y. Tsao, and R. F. Karlicek, Jr., Solid-State Lighting: Toward Smart and Ultra-Efficient Materials, Devices, Lamps and Systems, in D. L. Andrews, ed., *Handbook of Photonics*, Volume 3, *Photonics Technology and Instrumentation*, Wiley-Science Wise, 2015.

V. K. Khanna, *Fundamentals of Solid-State Lighting: LEDs, OLEDs, and Their Applications in Illumination and Displays*, CRC Press/Taylor & Francis, 2014.

M. N. Khan, *Understanding LED Illumination*, CRC Press/Taylor & Francis, 2014.

J. J. Wierer, Jr., J. Y. Tsao, and D. S. Sizov, Comparison Between Blue Lasers and Light-Emitting Diodes for Future Solid-State Lighting, *Laser & Photonics Reviews*, vol. 7, pp. 963–993, 2013.

T. W. Murphy, Maximum Spectral Luminous Efficacy of White Light, *Journal of Applied Physics*, vol. 111, 104909, 2012.

S. Liu and X. Luo, *LED Packaging for Lighting Applications: Design, Manufacturing, and Testing*, Wiley–Chemical Industry Press, 2011.

U.S. Department of Energy Technical Reports on LED Lighting

M. Pattison, M. Hansen, N. Bardsley, G. D. Thomson, K. Gordon, A. Wilkerson, K. Lee, V. Nubbe, and S. Donnelly, *2022 Solid-State Lighting R&D Opportunities*, U.S. Department of Energy, DOI:10.2172/1862626, Technical Report EE-2542, February 2022.

M. Pattison, M. Hansen, N. Bardsley, C. Elliott, K. Lee, L. Pattison, and J. Tsao, *2019 Lighting R&D Opportunities*, U.S. Department of Energy, DOI:10.13140/RG.2.2.30048.64001, Technical Report EE-2008, January 2020.

M. Pattison, N. Bardsley, C. Elliot, M. Hansen, K. Lee, L. Pattison, J. Tsao, and M. Yamada, *2018 Solid-State Lighting R&D Opportunities*, J. Brodrick, ed., U.S. Department of Energy, DOI:10.13140/RG.2.2.12827.52009, Technical Report EE-1907, January 2019.

M. Pattison, N. Bardsley, M. Hansen, L. Pattison, S. Schober, K. Stober, J. Tsao, and M. Yamada, *Solid-State Lighting 2017 Suggested Research Topics Supplement: Technology and Market Context*, J. Brodrick, ed., U.S. Department of Energy, DOI:10.13140/RG.2.2.13316.63364, Technical Report, September 2017.

N. Bardsley, M. Hansen, L. Pattison, M. Pattison, K. Stober, V. Taylor, J. Tsao, and M. Yamada, *2016 Solid-State Lighting R&D Plan*, J. Brodrick, ed., U.S. Department of Energy, DOI:10.13140/RG.2.1.2800.7929, Technical Report EE-1418, June 2016.

N. Bardsley, S. Bland, M. Hansen, L. Pattison, M. Pattison, K. Stober, and M. Yamada, *2015 Solid-State Lighting R&D Plan*, Bardsley Consulting, eds., U.S. Department of Energy, DOI:10.13140/RG.2.1.4897.9442, Technical Report EE-1228, May 2015.

Commercial LED Lighting Catalogs and Websites

Cree LED, *LED Components: Product & Application Guide*, Cree LED, Durham, NC 27709, `https://downloads.cree-led.com/files/fs/Product-Guide.pdf`, May 2023 (FS36R9).

Signify (Philips) LED Lighting, *Professional LED Lighting Catalog*, Signify N.V., Eindhoven, Netherlands, `https://www.assets.signify.com/is/content/Signify/Assets/signify/global/20230202-global-trade-catalog-december-2022.pdf`, December 2022.

Philips: `https://www.lighting.philips.com/home`

Lumileds: `https://lumileds.com/`

Osram Sylvania Inc.: `https://www.osram.us/cb/`

Nichia Corporation: `https://led-ld.nichia.co.jp/en/product/index.html`

Seoul Semiconductor: `https://www.seoulsemicon.com/kr`

Retrofit Lamps and Luminaires

G. Ozenen, *Architectural Interior Lighting*, Springer, 2024.

J. Livingston, *Designing With Light: The Art, Science, and Practice of Architectural Lighting Design*, Wiley, 2nd ed. 2022.

W. van Bommel, *Interior Lighting: Fundamentals, Technology and Application*, Springer, 2019.

W. van Bommel, *Road Lighting: Fundamentals, Technology and Application*, Springer, 2016.

G. Gordon, *Interior Lighting for Designers*, Wiley, 5th ed. 2015.

K. Hakata and T. Matsuoka, Light-Emitting Device and Lighting Apparatus Incorporating Same, *U.S. Patent 8,400,051*, Patented March 19, 2013, Filed January 13, 2009.

OLED and Flexible Lighting Panels

R. Mertens, *The OLED Handbook: A Guide to OLED Technology, Industry & Market*, OLED-Info, 2023.

T. W. Lee, ed., *Graphene for Flexible Lighting and Displays*, Woodhead/Elsevier, 2019.

T. W. Lee, ed., *Graphene for Flexible Lighting and Displays*, Woodhead/Elsevier, 2019.

M. Koden, *OLED Displays and Lighting*, Wiley, 2017.

Smart, Connected, Human-Centric, and Horticultural Lighting

P. Bertoldi, ed., Proceeding of the 11th International Conference on Energy Efficiency in Domestic Appliances and Lighting (EEDAL'22), Publications Office of the European Union, Luxembourg, JRC132721, ISBN 978-92-76-99908-9, DOI:10.2760/356891, 2023.

T. Q. Khanh, P. Bodrogi, and T. Q. Vinh, *Human Centric Integrative Lighting*, Wiley–VCH, 2023.

U. Singh, A. Abraham, A. Kaklauskas, and T.-P. Hong, eds., *Smart Sensor Networks: Analytics, Sharing and Control*, Springer, 2022.

F. Markowitz, *Outdoor Lighting for Pedestrians*, Routledge/Taylor & Francis, 2022.

C. Samarakoon, H. W. Choi, S. Lee, X.-B. Fan, D.-W. Shin, S. Y. Bang, J.-W. Jo, L. Ni, J. Yang, Y. Kim, S.-M. Jung, L. G. Occhipinti, G. A. J. Amaratunga, and J. M. Kim, Optoelectronic System and Device Integration for Quantum-Dot Light-Emitting Diode White Lighting with Computational Design Framework, *Nature Communications*, vol. 13, 4189, 2022.

S. Winchip, *Fundamentals of Lighting*, Bloomsbury, 4th ed. 2022.

N. Santhi and M. Spitschan, eds., *Circadian and Visual Neuroscience*, Elsevier, 4th ed. 2022.

L. Marcelis, E. Goto, B. Grodzinski, S. Torre, J. Wargent, and B. Bugbee, *Crop Physiology Under LED Lighting*, Frontiers in Plant Science, 2022.

T. Kozai, G. Niu, and J. Masabni, eds., *Plant Factory: Basics, Applications and Advances*, Academic/Elsevier, 2022.

T. Kozai, G. Niu, and J. Takagaki, eds., *Plant Factory: An Indoor Vertical Farming System for Efficient Quality Food Production*, Academic/Elsevier, 2nd ed. 2022.

A. Wolska, D. Sawicki, and M. Tafil-Klawe, *Visual and Non-Visual Effects of Light: Working Environment and Well-Being*, CRC Press/Taylor & Francis, 2021.

M. Rossi, *Circadian Lighting Design in the LED Era*, Springer, 2019.

M. Anpo, H. Fukuda, and T. Wada, eds., *Plant Factory Using Artificial Light: Adapting to Evironmental Disruption and Clues to Agricultural Innovation*, Elsevier, 2019.

T. Kozai, *Smart Plant Factory: The Next Generation Indoor Vertical Farms*, Springer, 2018.

Y. Zou, *Light-Emitting Diode Lighting Quality Effects on Morphology, Growth, Flowering, and Carotenoid Content of Potted Plant, Cut Flower, and Medicinal Flower Production*, Ph.D. Dissertation, University of California, Davis, 2018.

S. Dutta Gupta, ed., Light Emitting Diodes for Agriculture: Smart Lighting, Springer, 2017.

T. Kozai, K. Fujiwara, and E. S. Runkle, eds., *LED Lighting for Urban Agriculture*, Springer, 2016.

P. R. Boyce, *Human Factors in Lighting*, CRC Press/Taylor & Francis, 3rd ed. 2014.

Laser-Diode Lighting

F. Rahman, Diode Laser-Excited Phosphor-Converted Light Sources: A Review, *Optical Engineering*, vol. 61, 060901, 2022.

C. Wu, Z. Liu, Z. Yu, X. Peng, Z. Liu, X. Liu, X. Yao, and Y. Zhang, Phosphor-Converted Laser-Diode-Based White Lighting Module with High Luminous Flux and Color Rendering Index, *Optics Express*, vol. 28, pp. 19085–19096, 2020.

K. A. Denault, M. Cantore, S. Nakamura, S. P. DenBaars, and R. Seshadri, Efficient and Stable Laser-Driven White Lighting, *AIP Advances*, vol. 3, no. 7, 072107, 2013.

APPENDIX

A

FOURIER TRANSFORM

This appendix provides a review of the Fourier transform along with some of its salient features.

A.1 DEFINITION, PROPERTIES, AND EXAMPLES

The harmonic function $F\exp(j2\pi\nu t)$, with frequency ν and complex amplitude F, is an important mathematical element in science and engineering. The variable t represents time and the frequency ν has units of Hz. The real part of this function, $|F|\cos(2\pi\nu t + \arg\{F\})$, is a cosine function with amplitude $|F|$ and phase $\arg\{F\}$. The harmonic function can be viewed as a building block from which other functions may be constructed by superposition.

Specifically, the Fourier theorem specifies that an arbitrary complex-valued function $f(t)$, satisfying a relatively unrestrictive set of conditions, may be decomposed into an integral comprising a superposition of harmonic functions of different frequencies and complex amplitudes,

$$f(t) = \int_{-\infty}^{\infty} F(\nu)\exp(j2\pi\nu t)\,d\nu. \tag{A.1-1}$$

Inverse Fourier Transform

The component of frequency ν has a complex amplitude $F(\nu)$ given by

$$F(\nu) = \int_{-\infty}^{\infty} f(t)\exp(-j2\pi\nu t)\,dt. \tag{A.1-2}$$

Fourier Transform

The quantity $F(\nu)$ is called the **Fourier transform** of $f(t)$, and $f(t)$ is termed the **inverse Fourier transform** of $F(\nu)$. The functions $f(t)$ and $F(\nu)$ form a **Fourier-transform pair**; knowledge of one enables the other to be unambiguously determined.

We have adopted the convention that the harmonic temporal function $F\exp(j2\pi\nu t)$ is designated by a frequency ν that is positive. This choice is arbitrary and some authors adopt the opposite convention, defining the Fourier transform and its inverse in (A.1-2) and (A.1-1) with positive and negative signs in their exponents, respectively.

In statistical communication theory, the functions $f(t)$ and $F(\nu)$ represent the time-domain and frequency-domain representations of a signal, respectively. The absolute-squares of these quantities, $|f(t)|^2$ and $|F(\nu)|^2$, are referred to as the **signal power** and **energy spectral density**, respectively. If $|F(\nu)|^2$ extends over a wide range of frequencies, the signal is said to be of broad bandwidth.

LED Lighting: Devices and Colorimetry. Malvin Carl Teich.
Google Books. Published 2024.

Properties of the Fourier Transform

A number of important properties of the Fourier transform are provided below. These properties can be derived by direct application of the definitions in (A.1-1) and (A.1-2).

- *Linearity.* The Fourier transform of the sum of two functions is the sum of their Fourier transforms, indicating that the superposition principle of linear systems is obeyed.
- *Scaling.* If $f(t)$ has a Fourier transform $F(\nu)$, and τ is a real scaling factor, then $f(t/\tau)$ has a Fourier transform $|\tau|F(\tau\nu)$. Hence, if $f(t)$ is scaled by a factor τ, its Fourier transform is scaled by a factor $1/\tau$. It follows that if $\tau > 1$, then $f(t/\tau)$ is a stretched version of $f(t)$, while $F(\tau\nu)$ is a compressed version of $F(\nu)$. The Fourier transform of $f(-t)$ is $F(-\nu)$.
- *Time Translation.* If $f(t)$ has a Fourier transform $F(\nu)$, the Fourier transform of $f(t-\tau)$ is $\exp(-j2\pi\nu\tau)F(\nu)$. Hence, delay by a time τ corresponds to multiplying the Fourier transform by a phase factor $\exp(-j2\pi\nu\tau)$.
- *Frequency Translation.* If $F(\nu)$ is the Fourier transform of $f(t)$, the Fourier transform of $f(t)\exp(j2\pi\nu_0 t)$ is $F(\nu-\nu_0)$. Hence, multiplication by a harmonic function of frequency ν_0 corresponds to shifting the Fourier transform to a higher frequency ν_0.
- *Symmetry.* If $f(t)$ is real, then $F(\nu)$ has Hermitian symmetry, indicating that $F(-\nu) = F^*(\nu)$. If $f(t)$ is real and symmetric, then so too is $F(\nu)$.
- *Convolution Theorem.* If the Fourier transforms of $f_1(t)$ and $f_2(t)$ are $F_1(\nu)$ and $F_2(\nu)$, respectively, the inverse Fourier transform of the product

$$F(\nu) = F_1(\nu)F_2(\nu) \tag{A.1-3}$$

  is

$$\boxed{f(t) = \int_{-\infty}^{\infty} f_1(\tau)f_2(t-\tau)\,d\tau.} \tag{A.1-4}$$

  Convolution

  The operation defined in (A.1-4) is known as the **convolution** of $f_1(t)$ with $f_2(t)$. Convolution in the time domain is thus equivalent to multiplication in the frequency domain.
- *Correlation Theorem.* The **correlation** between two complex functions is defined as

$$\boxed{f(t) = \int_{-\infty}^{\infty} f_1^*(\tau)f_2(t+\tau)\,d\tau.} \tag{A.1-5}$$

  Correlation

  The Fourier transforms of $f_1(t)$, $f_2(t)$, and $f(t)$ are related by

$$F(\nu) = F_1^*(\nu)F_2(\nu). \tag{A.1-6}$$

  If $f_2(t) = f_1(t)$, the integral in (A.1-5) is called the **autocorrelation function**.
- *Parseval's Theorem.* The energy associated with $f(t)$, which is the time integral of the power $|f(t)|^2$, is equal to the frequency integral of the energy spectral density $|F(\nu)|^2$, i.e.,

$$\boxed{\int_{-\infty}^{\infty} |f(t)|^2\,dt = \int_{-\infty}^{\infty} |F(\nu)|^2\,d\nu.} \tag{A.1-7}$$

  Parseval's Theorem

Table of Fourier Transforms

The Fourier transforms of selected functions are set forth in Table A.1-1. The properties of linearity, scaling, delay, and frequency translation enable the Fourier transforms of other functions to be determined.

The following definitions are used in Table A.1-1:

- $\text{rect}(t) \equiv 1$ for $|t| \leqslant 1/2$, and $= 0$ elsewhere, indicating that it is a pulse of unit height and unit width centered at $t = 0$.
- $\delta(t)$ is the impulse function (also called the Dirac delta function), which is defined as $\delta(t) \equiv \lim_{\alpha \to \infty} \alpha\, \text{rect}(\alpha t)$. It is therefore the limit of a rectangular pulse of unit area as its width approaches 0 and its height approaches ∞.
- $\text{sinc}(t) \equiv \sin(\pi t)/(\pi t)$ is a symmetric function with a peak value of unity at $t = 0$ and zeros at $t = \pm 1, \pm 2, \ldots$.

Table A.1-1 Fourier transforms of selected functions.

Function	$f(t)$	$F(\nu)$
Uniform	1	$\delta(\nu)$
Impulse	$\delta(t)$	1
Rectangular	rect (t)	sinc (ν)
Exponential [a]	exp $(-\lvert t\rvert)$	$\frac{2}{1+(2\pi\nu)^2}$
Gaussian	exp $(-\pi t^2)$	exp $(-\pi\nu^2)$
Hyperbolic secant	sech (πt)	sech $(\pi\nu)$
Chirp [b]	exp $(j\pi t^2)$	$e^{j\pi/4}$ exp $(-j\pi\nu^2)$
$M = 2S+1$ Impulses	$\sum_{m=-S}^{S} \delta(t-m)$	$\frac{\sin(M\pi\nu)}{\sin(\pi\nu)}$
Comb	$\sum_{m=-\infty}^{\infty} \delta(t-m)$	$\sum_{m=-\infty}^{\infty} \delta(\nu-m)$

[a] The double-sided exponential function is illustrated. The Fourier transform of the single-sided exponential function, $f(t) = \exp(-t)$ with $t \geqslant 0$, is $F(\nu) = 1/(1 + j2\pi\nu)$; its magnitude is given by $1/\sqrt{1+(2\pi\nu)^2}$.
[b] The functions $\cos(\pi t^2)$ and $\cos(\pi\nu^2)$ are displayed.

A.2 TEMPORAL AND SPECTRAL WIDTHS

It is often useful to specify the width of a function, e.g., the time duration of a function of time $f(t)$ or the spectral extent (bandwidth) of its Fourier transform $F(\nu)$. Many different measures of width are in use.

In accordance with the scaling property of the Fourier transform, however, all width definitions for a Fourier-transform pair share the property that the spectral width is inversely proportional to the temporal width.

The three classes of measures defined below, the **root-mean-square width**, the **power-equivalent width**, and the **full-width at half-maximum**, are widely used in photonics.

Root-Mean-Square Width

The normalized **root-mean-square (RMS) temporal width** σ_t of a nonnegative, real function $f(t)$ is defined as

$$\sigma_t = \sqrt{\frac{\int_{-\infty}^{\infty}(t-\bar{t})^2 f(t)\,dt}{\int_{-\infty}^{\infty} f(t)\,dt}} \quad \text{with} \quad \bar{t} = \frac{\int_{-\infty}^{\infty} t f(t)\,dt}{\int_{-\infty}^{\infty} f(t)\,dt}\,. \tag{A.2-1}$$

If $f(t)$ is a probability density function, then $\bar{t}$ and σ_t represent its **mean** and **standard deviation**, respectively. As an example, the *Gaussian function* $f(t) = \exp(-t^2/2\sigma_t^2)$ has an RMS temporal width σ_t. Its Fourier transform, $F(\nu) = (1/\sqrt{2\pi}\,\sigma_\nu)\exp(-\nu^2/2\sigma_\nu^2)$, has an **RMS spectral width** given by $\sigma_\nu = 1/2\pi\sigma_t$, so that

$$\sigma_t\sigma_\nu = 1/2\pi\,. \tag{A.2-2}$$

Power-RMS Width. A measure analogous to that specified in (A.2-1), but one that also accommodates negative and complex functions, is provided by the RMS width of the absolute-square of $|f(t)|^2$, i.e.,

$$\sigma_t = \sqrt{\frac{\int_{-\infty}^{\infty}(t-\bar{t})^2|f(t)|^2\,dt}{\int_{-\infty}^{\infty}|f(t)|^2\,dt}} \quad \text{with} \quad \bar{t} = \frac{\int_{-\infty}^{\infty} t|f(t)|^2\,dt}{\int_{-\infty}^{\infty}|f(t)|^2\,dt}\,. \tag{A.2-3}$$

This version of σ_t is sometimes referred to as the **power-RMS width**.

With the help of the Schwarz inequality, it can be readily shown that multiplying the **temporal power-RMS width** for an arbitrary function $f(t)$ by the width of its Fourier transform $F(\nu)$, the **spectral power-RMS width**, gives rise to a duration–bandwidth reciprocity relation expressible as

$$\boxed{\sigma_t\sigma_\nu \geqslant 1/4\pi.} \tag{A.2-4}$$

Duration–Bandwidth Reciprocity Relation

The spectral width σ_ν specified in (A.2-4), defined in analogy with (A.2-3), is given by

$$\sigma_\nu = \sqrt{\frac{\int_{-\infty}^{\infty}(\nu-\bar{\nu})^2|F(\nu)|^2\,d\nu}{\int_{-\infty}^{\infty}|F(\nu)|^2\,d\nu}} \quad \text{with} \quad \bar{\nu} = \frac{\int_{-\infty}^{\infty}\nu|F(\nu)|^2\,d\nu}{\int_{-\infty}^{\infty}|F(\nu)|^2\,d\nu}\,. \tag{A.2-5}$$

Pursuant to (A.2-4), it is apparent that the time duration and bandwidth cannot simultaneously be made arbitrarily small. The *Gaussian function* $f(t) = \exp(-t^2/4\sigma_t^2)$, which has a power-RMS temporal width given by σ_t, provides an example. Its Fourier transform is also a Gaussian function, $F(\nu) = (1/2\sqrt{\pi}\,\sigma_\nu)\exp(-\nu^2/4\sigma_\nu^2)$, which has a power-RMS spectral width given by $\sigma_\nu = 1/4\pi\sigma_t$. The product of the power-RMS temporal and spectral widths for this Gaussian function therefore becomes

$$\sigma_t\sigma_\nu = 1/4\pi, \tag{A.2-6}$$

revealing that it assumes the minimum possible value of the duration–bandwidth product specified in (A.2-4).

Heisenberg Uncertainty Relation. Since angular frequency and frequency are related by $\omega = 2\pi\nu$, the time–frequency reciprocity relation in (A.2-4) can be rewritten in the form

$$\sigma_t\sigma_\omega \geqslant 1/2\,. \tag{A.2-7}$$

Moreover, since the energy of a photon is given by $E = \hbar\omega$, (A.2-7) in turn dictates a limit on the precision with which the time t and energy E of a photon can be simultaneously determined, as specified by the Heisenberg time–energy uncertainty relation set forth in (3.3-8).

In quantum mechanics, the position x of a particle is described by the wavefunction $\psi(x)$, while the spatial angular frequency (wavenumber) k is described by the function $\phi(k)$, which is the Fourier transform of $\psi(x)$. The uncertainties of x and k, denoted σ_x and σ_k, respectively, are therefore the RMS widths of the probability densities $|\psi(x)|^2$ and $|\phi(k)|^2$. The variables x and k (rad/m) are thus analogous to the variables t and ω that describe time and angular frequency (rad/s) in (A.2-7), respectively. In short, we arrive at

$$\sigma_x\sigma_k \geqslant 1/2\,, \tag{A.2-8}$$

which is analogous to (A.2-7). Moreover, the particle momentum is given by $p = \hbar k$ (where $\hbar \equiv h/2\pi$ and h is Planck's constant), which provides $\sigma_p = \hbar\sigma_k$, so that (A.2-8) becomes

$$\boxed{\sigma_x\sigma_p \geqslant \hbar/2\,.} \tag{A.2-9}$$

Heisenberg Position–Momentum Uncertainty Relation

Equation (A.2-9), known as the **Heisenberg position–momentum uncertainty relation**, is analogous to the Heisenberg time–energy uncertainty relation stated in (3.3-8).

Power-Equivalent Width

The **power-equivalent temporal width** of the function $f(t)$ is its associated energy divided by the peak power. In particular, if $f(t)$ has its peak value at $t = 0$, the power-equivalent temporal width is defined as

$$\tau = \int_{-\infty}^{\infty} \frac{|f(t)|^2}{|f(0)|^2}\,dt. \tag{A.2-10}$$

As examples, the *double-sided exponential function* $f(t) = \exp(-|t|/\tau)$ has a power-equivalent temporal width τ, as does the Gaussian function $f(t) = \exp(-\pi t^2/2\tau^2)$.

This measure was used in Sec. 2.7 to define the coherence time of light τ_c in terms of the complex degree of temporal coherence $g(\tau)$ via (2.7-10).

The **power-equivalent spectral width** is analogously defined as

$$\mathrm{B} = \int_{-\infty}^{\infty} \frac{|F(\nu)|^2}{|F(0)|^2}\, d\nu. \tag{A.2-11}$$

If $f(t)$ is real, then $|F(\nu)|^2$ is symmetric; if its peak is also located at $\nu = 0$, it is convenient to replace the power-equivalent spectral width B by the **positive-frequency power-equivalent spectral width** B, i.e.,

$$B = \int_{0}^{\infty} \frac{|F(\nu)|^2}{|F(0)|^2}\, d\nu. \tag{A.2-12}$$

As an example, if $F(\nu) = \tau/(1 + j2\pi\nu\tau)$, as for a single-sided exponentially decaying time function, carrying out the integration in (A.2-12) leads to

$$B = 1/4\tau. \tag{A.2-13}$$

Using Parseval's theorem (A.1-7), together with the relation $F(0) = \int_{-\infty}^{\infty} f(t)\, dt$, allows (A.2-12) to be rewritten in the form

$$B = 1/2T\,, \tag{A.2-14}$$

where

$$T = \frac{\left[\int_{-\infty}^{\infty} f(t)\, dt\right]^2}{\int_{-\infty}^{\infty} f^2(t)\, dt} \tag{A.2-15}$$

is yet another definition of the temporal width, namely, the square of the area under $f(t)$ divided by the area under $f^2(t)$. For these particular definitions of width, Parseval's theorem takes the form

$$TB = {}^1\!/_2\,. \tag{A.2-16}$$

Full-Width at Half-Maximum (FWHM), 3-dB, and 1/e Widths

The third class of width measures that we consider is the duration (or bandwidth) of a function at a prescribed fraction of its maximum value, e.g., ${}^1\!/_2$, ${}^1\!/_{\sqrt{2}}$, ${}^1\!/_e$, or ${}^1\!/_{e^2}$. Either the half-width or the full width on both sides of the peak may be used in the definition. Two particularly common measures of spectral width are the full-width at half-maximum (FWHM) and the half-width at ${}^1\!/_{\sqrt{2}}$-maximum, also called the 3-dB width. We provide three examples:

- The *double-sided exponential function* $f(t) = \exp(-|t|/\tau)$ has a half-width at ${}^1\!/_e$-maximum given by $\Delta t_{1/e} = \tau$. Its Fourier transform, which is given by $F(\nu) = 2\tau/[1 + (2\pi\nu\tau)^2]$, is known as the **Lorentzian distribution** and has a full-width at half-maximum (FWHM) expressible as

  $$\Delta\nu_{\mathrm{FWHM}} = 1/\pi\tau\,. \tag{A.2-17}$$

  The Lorentzian distribution characterizes the spectrum of certain sources of light (Sec. 4.6), and is usually cast in the form $F(\nu) = (\Delta\nu/2\pi)/[\nu^2 + (\Delta\nu/2)^2]$, where $\Delta\nu = \Delta\nu_{\mathrm{FWHM}}$.

- The *exponential function* $f(t) = \exp(-t/\tau)$ for $t \geqslant 0$, and $f(t) = 0$ for $t < 0$, which describes the temporal response of many photonic and electronic systems, is characterized by a width at $1/e$-maximum given by $\Delta t_{1/e} = \tau$. Its Fourier transform, $F(\nu) = \tau/(1 + j2\pi\nu\tau)$, has magnitude $\tau/[1 + (2\pi\nu\tau)^2]^{1/2}$ and hence a half-width at $1/\sqrt{2}$-maximum (3-dB width) expressible as

$$\Delta\nu_{3-\mathrm{dB}} = 1/2\pi\tau\,. \tag{A.2-18}$$

- The *Gaussian function* $f(t) = \exp(-t^2/2\tau^2)$ has a full-width at $1/e$-maximum $\Delta t_{1/e} = 2\sqrt{2}\,\tau$. Its Fourier transform $F(\nu) = \sqrt{2\pi}\,\tau \exp(-2\pi^2\tau^2\nu^2)$ has a full-width at $1/e$-maximum

$$\Delta\nu_{1/e} = \sqrt{2}\,/\pi\tau\,, \tag{A.2-19}$$

and a full-width at half-maximum

$$\Delta\nu_{\mathrm{FWHM}} = \sqrt{2\ln 2}\,/\pi\tau, \tag{A.2-20}$$

so that

$$\Delta\nu_{\mathrm{FWHM}} = \sqrt{\ln 2}\;\Delta\nu_{1/e} \approx 0.833\,\Delta\nu_{1/e}\,. \tag{A.2-21}$$

The Gaussian also describes the spectrum of light emitted by certain optical sources as well as the spatial distribution of so-called Gaussian light beams.

BIBLIOGRAPHY

J. V. Stone, *The Fourier Transform: A Tutorial Introduction*, Sebtel Press, 2021.

B. G. Osgood, *Lectures on the Fourier Transform and its Applications*, American Mathematical Society, 2019.

L. F. Chaparro and A. Akan, *Signals and Systems Using* MATLAB, Academic/Elsevier, 3rd ed. 2018.

B. P. Lathi and R. Green, *Linear Systems and Signals*, Oxford University Press, 3rd ed. 2017.

J. F. James, *A Student's Guide to Fourier Transforms: With Applications in Physics and Engineering*, Cambridge University Press, 3rd ed. 2011.

S. Haykin and B. Van Veen, *Signals and Systems*, Wiley, 2nd ed. 2003.

G. R. Cooper and C. D. McGillem, *Probabilistic Methods of Signal and System Analysis*, Oxford University Press, 3rd ed. 1999.

A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and Systems*, Pearson, 2nd ed. 1997.

SYMBOLS AND UNITS

Roman Symbols and Acronyms

a = Radius of an aperture or fiber [m]; also, Radius of a circle [m]; also, Distance between locations [m]; also, Lattice constant [m]
$\mathfrak{a}$ = Complex amplitude or magnitude of an optical wave
A = Complex envelope of a monochromatic plane wave
$A(\mathbf{r})$ = Complex envelope of a monochromatic wave
$\mathcal{A}$ = Area [m^2]
A_{eff} = Effective area [m^2]
$\mathcal{A}$ = Einstein $\mathcal{A}$ coefficient [s^{-1}]
$\mathcal{A}$ = Ratio of yellow-to-blue optical power in a white phosphor-conversion LED
$\mathfrak{A}$ = Alkali metal such as Li, Na, K, Rb, Cs, or combination thereof
AC = Alternating current
ACS = American Chemical Society
AM = Amplitude modulation
A19 = Designator for the bulb shape of a classic incandescent lamp

b = Proportionality constant in brightness–luminance relation
$\mathfrak{b}$ = Wien's constant [2.90×10^6 nm $\cdot$ K]
$\mathcal{B}$ = Magnetic flux density vector [$\mathrm{Wb} \cdot \mathrm{m}^{-2}$ or T]
$\mathbf{B}$ = Magnetic flux-density complex amplitude vector [$\mathrm{Wb} \cdot \mathrm{m}^{-2}$ or T]
B = Bandwidth [Hz]; also, Bandwidth of an electrical circuit [Hz]
$\mathfrak{B}$ = Power-equivalent spectral width [Hz]
$\mathcal{B}$ = Einstein $\mathcal{B}$ coefficient [$\mathrm{m}^3 \cdot \mathrm{J}^{-1} \cdot \mathrm{s}^{-2}$]
B_{v} = Psychophysical magnitude estimate of luminance
BB = Broadband

c = Speed of light [$\mathrm{m} \cdot \mathrm{s}^{-1}$]; also, Phase velocity [$\mathrm{m} \cdot \mathrm{s}^{-1}$]
c_0 = Speed of light in free space [2.9979×10^8 m $\cdot$ s^{-1}]
C = Electrical capacitance [F]
C = Psychophysical magnitude estimate of cold
CCT = Correlated color temperature
CFL = Compact fluorescent lamp
CIE = Commission Internationale de l'Éclairage

LED Lighting: Devices and Colorimetry. Malvin Carl Teich.
Google Books. Published 2024.

CIP = Commission Internationale de Photométrie
CLE = Current luminous efficacy η_{CLE} (also called current efficiency) $[\mathrm{cd/A}]$
CMLED = Color-mixing light-emitting diode
CMOS = Complementary metal-oxide-semiconductor
COB = Chip-on-board light-emitting diode
COG = Chip-on-glass construction for LED-filament lamps
cQD = Colloidal quantum dot
CRI = Color rendering index
CT = Color temperature
CVD = Color-vision deficiency
CW = Continuous-wave
$\mathfrak{CMY}$ = Primaries associated with CMYK color space

$d\mathbf{r}$ = Incremental volume $[\mathrm{m}^3]$
ds = Incremental length $[\mathrm{m}]$
d = Distance, Length, Thickness $[\mathrm{m}]$
D = Diameter $[\mathrm{m}]$
$\mathfrak{D}$ = Electric flux density vector $[\mathrm{C}\cdot\mathrm{m}^{-2}]$
$\mathbf{D}$ = Electric flux density complex amplitude vector $[\mathrm{C}\cdot\mathrm{m}^{-2}]$
DC = Direct current
DH = Double-heterostructure
DIP = Dual-inline package
DSPP = Doubly stochastic Poisson process
DUV = Deep ultraviolet, stretching from 200 to 300 nm
DWL = Dominant wavelength
0D = Zero-dimensional
1D = One-dimensional
2D = Two-dimensional
3D = Three-dimensional

e = Elementary electronic charge $[1.6022\times10^{-19}\ \mathrm{C}]$
$\mathfrak{E}$ = Electric-field vector $[\mathrm{V}\cdot\mathrm{m}^{-1}]$
$\mathbf{E}$ = Electric-field complex amplitude vector $[\mathrm{V}\cdot\mathrm{m}^{-1}]$
E = Energy [J]; also, Optical energy (or radiant energy) [J]
E_A = Acceptor energy level [J]
E_c = Energy at the bottom of the conduction band [J]
E_D = Donor energy level [J]
E_f = Fermi energy [J]
E_{fc} = Quasi-Fermi energy for the conduction band [J]
E_{fv} = Quasi-Fermi energy for the valence band [J]
E_g = Bandgap energy [J]
E_k = Kinetic energy [J]
E_q = Energy of the qth mode [J]
E_v = Energy at the top of the valence band [J]
E_ν = Energy spectral density $[\mathrm{J}\cdot\mathrm{Hz}^{-1}]$
E_{v} = Luminous energy $[\mathrm{lm}\cdot\mathrm{s}]$
$\mathfrak{E}$ = Alkaline-earth element such as Mg, Ca, Sr, Ba, Zn, or combination thereof
ECE = Energy-conversion efficiency η_{PCE} (also called power-conversion efficiency, PCE)
ELLED = Electroluminescent light-emitting diode
eQD = Epitaxial quantum dot

EQE = External quantum efficiency η_{EQE}
E.U. = European Union
EUV = Extreme-ultraviolet, stretching from 10 to 100 nm
E12 = (Edison) screw base designator for an incandescent candle lamp (U.S.)
E26 = (Edison) screw base designator for a classic incandescent lamp (U.S.)
E27 = (Edison) screw base designator for a classic incandescent lamp (E.U.)

f = Focal length of a lens $[\mathrm{m}]$; also, Frequency $[\mathrm{Hz}]$
$f(E)$ = Fermi function
$f_c(E)$ = Fermi function for the conduction band
$f_v(E)$ = Fermi function for the valence band
f_a = Probability that absorption condition is satisfied
f_e = Probability that emission condition is satisfied
f_g = Fermi inversion factor
f = Phosphor-absorption fraction
F = Focal point of an optical system; also, Complex amplitude of a harmonic function
F_{P} = Purcell factor
$\mathcal{F}$ = Force $[\mathrm{kg \cdot m \cdot s^{-2}}]$
FIR = Far infrared, stretching from 20 to 300 μm
FUV = Far ultraviolet, stretching from 100 to 200 nm
FWHM = Full-width at half-maximum

$g(\nu)$ = Lineshape function of a transition $[\mathrm{Hz^{-1}}]$
$\overline{g}(\nu)$ = Average lineshape function $[\mathrm{Hz^{-1}}]$
$g_{\nu 0}(\nu)$ = Electron–photon collisionally broadened lineshape function in a semiconductor $[\mathrm{Hz^{-1}}]$
$g(\tau)$ = Complex degree of temporal coherence
g = Degeneracy factor
g = Red phosphor absorption fraction
$G(\tau)$ = Temporal coherence function $[\mathrm{W \cdot m^{-2}}]$
G_0 = Rate of thermal electron–hole generation in a semiconductor $[\mathrm{m^{-3} \cdot s^{-1}}]$
GE = General Electric Company
GRIN = Graded-index

h = Planck's constant $[6.6261 \times 10^{-34}\ \mathrm{J \cdot s}]$
$\hbar$ = Reduced Planck constant $(\hbar \equiv h/2\pi)$ $[1.0546 \times 10^{-34}\ \mathrm{J \cdot s}]$
$h(t)$ = Impulse response function of a linear system
$h(x, y)$ = Impulse response function of a two-dimensional linear system
$\mathcal{H}$ = Magnetic-field vector $[\mathrm{A \cdot m^{-1}}]$
$\mathbf{H}$ = Magnetic-field complex amplitude vector $[\mathrm{A \cdot m^{-1}}]$
$H(\chi)$ = Transfer function of a linear electrical system
HCL = Human-centric lighting
HD = High definition
HOMO = Highest occupied molecular orbital
HVPE = Hydride vapor-phase epitaxy
HVS = Human visual system

i = Electric current $[\mathrm{A}]$; also, Integer; also, $\sqrt{-1}$
I = Irradiance (also called optical intensity) $[\mathrm{W \cdot m^{-2}}]$
I_ν = Spectral irradiance (also called spectral intensity) $[\mathrm{W \cdot m^{-2} \cdot Hz^{-1}}]$
I_{v} = Illuminance $[\mathrm{lx}]$

$\mathfrak{I}$ = Moment of inertia $[\mathrm{kg} \cdot \mathrm{m}^2]$
I = Radiant intensity $[\mathrm{W} \cdot \mathrm{sr}^{-1}]$
I_v = Luminous intensity $[\mathrm{cd}]$
IEEE = Institute of Electrical and Electronics Engineers
INL = Inner nuclear layer
IPL = Inner plexiform layer
IQE = Internal quantum efficiency $\eta_\mathtt{IQE}$
IR = Infrared
IRE = Institute of Radio Engineers
ITO = Indium tin oxide

j = Integer; also $\sqrt{-1}$
J = Electric current density $[\mathrm{A} \cdot \mathrm{m}^{-2}]$
$J_0(u)$ = Bessel function of the first kind and zeroth order
J = Total atomic overall angular momentum quantum number
$\boldsymbol{\mathcal{J}}$ = Electric current density vector $[\mathrm{A} \cdot \mathrm{m}^{-2}]$

k = Wavenumber $[\mathrm{m}^{-1}]$; also, Integer; also, Spatial angular frequency $[\mathrm{rad} \cdot \mathrm{m}^{-1}]$
k_0 = Free-space wavenumber $[\mathrm{m}^{-1}]$
k_x, k_y, k_z = Wavevector components in x, y, and z directions $[\mathrm{m}^{-1}]$; also, Spatial angular frequency $[\mathrm{rad} \cdot \mathrm{m}^{-1}]$
$\mathbf{k}$ = Wavevector $[\mathrm{m}^{-1}]$
$\mathbf{k_q}$ = Wavevector for mode $\mathbf{q}$ $[\mathrm{m}^{-1}]$
k = Boltzmann's constant $[1.3807 \times 10^{-23}\ \mathrm{J} \cdot \mathrm{K}^{-1}]$
k, k' = Normalization constant for XYZ tristimulus values $[\mathrm{W}^{-1} \cdot \mathrm{m}]$
$\mathit{K}_{1,2,3}$ = Mixed-light chromaticity-coordinate weight functions
KSF = Potassium fluorosilicate doped with manganese ions (K_2SiF_6:Mn^{4+}, also called PFS)

l = Thickness $[\mathrm{m}]$
l_c = Coherence length $[\mathrm{m}]$
ℓ = Orbital angular-momentum quantum number
L = Length $[\mathrm{m}]$; also, Radiance $[\mathrm{W} \cdot \mathrm{sr}^{-1} \cdot \mathrm{m}^{-2}]$
L_v = Luminance $[\mathrm{cd} \cdot \mathrm{m}^{-2}]$
L_0 = Reference luminance value $[\mathrm{cd} \cdot \mathrm{m}^{-2}]$
L = Total atomic orbital angular momentum quantum number
L-cone = Long-wavelength (red) retinal cone (wavelength sensitivity: 560–565 nm)
LASER = Light amplification by stimulated emission of radiation
LCD = Liquid-crystal display
LD = Laser diode
LED = Light-emitting diode
LEP = Light-emitting polymer material
LER = Luminous efficacy of radiation $\eta_\mathtt{LER}$ $[\mathrm{lm/W}]$
LG = LG Chem
LGN = Lateral geniculate nucleus
LHP = Lead-halide perovskite
LMS = Color space associated with the cone fundamentals
LPE = Liquid-phase epitaxy
LPS = Low-pressure sodium lamp
LPW = Lumens per watt (lm/W)
LUC = Luminaire wall-plug luminous efficiency $\eta_\mathtt{LUC}$

LUM = Luminaire wall-plug luminous efficacy $\eta_{\mathtt{LUM}}$ $[\mathrm{lm/W}]$
LUMO = Lowest unoccupied molecular orbital
LWIR = Long-wavelength infrared, stretching from 8 to 14 μm

m = Mass of a particle $[\mathrm{kg}]$
m_c = Effective mass of a conduction-band electron $[\mathrm{kg}]$
m_r = Reduced effective mass of an electron–hole pair in a semiconductor $[\mathrm{kg}]$
m_v = Effective mass of a valence-band hole $[\mathrm{kg}]$
m_0 = Free electron mass $[9.1094 \times 10^{-31}\ \mathrm{kg}]$
m = Photon number; also, Photoelectron number
M = Magnification of an imaging system
$M_{\mathtt{V}}$ = Luminous exitance (luminous emittance) [lx]
$\mathfrak{M}$ = Magnetization density vector $[\mathrm{A \cdot m^{-1}}]$
M = Mass of an atom or molecule $[\mathrm{kg}]$
M_r = Reduced mass of an atom or molecule $[\mathrm{kg}]$
$\mathsf{M}(\nu)$ = Density of modes in a cavity [$\mathrm{m^{-3} \cdot Hz^{-1}}$ for 3D cavity; $\mathrm{m^{-1} \cdot Hz^{-1}}$ for a 1D cavity]
$\mathfrak{M}$ = Mulliken molecular term symbol
M-cone = Middle-wavelength (green) retinal cone (wavelength sensitivity: 530–535 nm)
MBE = Molecular-beam epitaxy
MCOB = Multiple chip-on-board
MES = Medium Edison screw
MHP = Metal-halide perovskite
MIR = Mid infrared, stretching from 2 to 20 μm
MIT = Massachusetts Institute of Technology
MMF = Multimode fiber
MOCVD = Metalorganic chemical vapor deposition (also called MOVPE)
MOVPE = Metalorganic vapor phase epitaxy (also called MOCVD)
MQD = Multiquantum dot
MQW = Multiquantum well
MQWLED = Multiquantum-well light-emitting diode
MUV = Mid ultraviolet, stretching from 200 to 300 nm
MWIR = Medium-wavelength infrared, stretching from 3 to 5 μm

n = Refractive index; also, Integer
n_s = Refractive index of a medium butt-coupled to an optical fiber
$n(\mathbf{r})$ = Refractive index of an inhomogeneous medium
n = Photon number
$\overline{\mathsf{n}}$ = Mean photon number
$\overline{\mathsf{n}}_\nu$ = Spectral photon number $[\mathrm{Hz^{-1}}]$
$\mathtt{n}$ = Concentration of electrons in a semiconductor $[\mathrm{m^{-3}}]$
$\mathtt{n}_{\mathrm{i}}$ = Concentration of electrons/holes in an intrinsic semiconductor $[\mathrm{m^{-3}}]$
$\mathtt{n}_0$ = Equilibrium concentration of electrons in a semiconductor $[\mathrm{m^{-3}}]$
$\mathfrak{n}$ = Principal quantum number of an atomic shell
N = Integer; also, Group index; also, Number of atoms or particles
N = Number density $[\mathrm{m^{-3}}]$; also, Number of subintervals
N_A = Number density of ionized acceptor atoms in a semiconductor $[\mathrm{m^{-3}}]$
N_D = Number density of ionized donor atoms in a semiconductor $[\mathrm{m^{-3}}]$
N_c = Constant associated with carrier density in the conduction band $[\mathrm{m^{-3}}]$
N_v = Constant associated with carrier density in the valence band $[\mathrm{m^{-3}}]$
NA = Numerical aperture

NB = Narrowband
NC = Nanocrystal
NSF = Sodium fluorosilicate doped with manganese ions (Na_2SiF_6:Mn^{4+})
NIR = Near infrared, stretching from 0.760 to 2 μm
NL = Nonlinear
NUV = Near ultraviolet, stretching from 300 to 390 nm

OLED = Organic light-emitting diode
ONL = Outer nuclear layer
OSA = Optical Society of America

p = Probability; also, Momentum $[\mathrm{kg \cdot m \cdot s^{-1}}]$
p_{ab} = Probability density for absorption (mode containing one photon) $[\mathrm{s^{-1}}]$
p_{sp} = Probability density for spontaneous emission (into one mode) $[\mathrm{s^{-1}}]$
p_{st} = Probability density for stimulated emission (mode containing one photon) $[\mathrm{s^{-1}}]$
$p(n)$ = Photon-number distribution
$\mathtt{p}$ = Concentration of holes in a semiconductor $[\mathrm{m^{-3}}]$
$\mathtt{p}_0$ = Equilibrium concentration of holes in a semiconductor $[\mathrm{m^{-3}}]$
P = Pressure $[\mathrm{kg \cdot m^{-1} \cdot s^{-2}}]$
P_{ab} = Probability density for absorption (mode containing many photons) $[\mathrm{s^{-1}}]$
P_{sp} = Probability density for spontaneous emission (into any mode) $[\mathrm{s^{-1}}]$
P_{st} = Probability density for stimulated emission (mode containing many photons) $[\mathrm{s^{-1}}]$
$\mathcal{P}$ = Electric polarization density vector $[\mathrm{C \cdot m^{-2}}]$
P = Optical power (also radiant flux and radiant power) $[\mathrm{W}]$
$\mathsf{P}_{\mathrm{col}}$ = Collected optical power $[\mathrm{W}]$
$\mathsf{P}_{\mathtt{EL}}$ = Electrical drive power $[\mathrm{W}]$
P_0 = Output (or emitted) optical power $[\mathrm{W}]$
P_ν = Spectral power $[\mathrm{W \cdot Hz^{-1}}]$
$\mathsf{P}_{\mathtt{V}}$ = Luminous flux $[\mathrm{lm}]$
PC = Phosphor-conversion
PCB = Printed circuit board
PCE = Power-conversion efficiency $\eta_{\mathtt{PCE}}$ (also called energy-conversion efficiency, ECE)
PCLED = Phosphor-conversion light-emitting diode
PeLED = Perovskite-based light-emitting diode
PFM = Pulse-frequency modulation
PFS = Potassium fluorosilicate doped with manganese ions (K_2SiF_6:Mn^{4+}, also called KSF)
PLED = Polymer light-emitting diode
PLQD = Photoluminescence quantum defect $\eta_{\mathtt{PLQD}}$
PLQY = Photoluminescence quantum yield $\eta_{\mathtt{PLQY}}$
P-OLED = Polymer light-emitting diode
PPV = Poly(p-phenylene vinylene)
PWL = Peak wavelength
PWM = Pulse-width modulation

q = Integer mode index
$\mathbf{q}$ = Mode index
Q = Quality factor of an optical cavity or resonant circuit
QCSE = Quantum-confined Stark effect
QD = Quantum dot
QLED = Quantum-dot light-emitting diode

QW = Quantum well

r = Radial distance $[\mathrm{m}]$
$\mathbf{r}$ = Position vector $[\mathrm{m}]$
$\mathfrak{r}$ = Electron–hole recombination coefficient $[\mathrm{m}^3 \cdot \mathrm{s}^{-1}]$
$\mathfrak{r}_{\mathrm{nr}}$ = Nonradiative electron–hole recombination coefficient $[\mathrm{m}^3 \cdot \mathrm{s}^{-1}]$
$\mathfrak{r}_{\mathrm{r}}$ = Radiative electron–hole recombination coefficient $[\mathrm{m}^3 \cdot \mathrm{s}^{-1}]$
$r_{\mathrm{ab}}(\nu)$ = Rate of photon absorption in a semiconductor $[\mathrm{s}^{-1} \cdot \mathrm{m}^{-3} \cdot \mathrm{Hz}^{-1}]$
$r_{\mathrm{sp}}(\nu)$ = Rate of spontaneous emission from a semiconductor $[\mathrm{s}^{-1} \cdot \mathrm{m}^{-3} \cdot \mathrm{Hz}^{-1}]$
$r_{\mathrm{st}}(\nu)$ = Rate of stimulated emission from a semiconductor $[\mathrm{s}^{-1} \cdot \mathrm{m}^{-3} \cdot \mathrm{Hz}^{-1}]$
$\mathrm{rect}(\cdot)$ = Pulse of unit height and unit width centered about the point 0
R = Radius of curvature $[\mathrm{m}]$; also, Radius $[\mathrm{m}]$; also, Distance $[\mathrm{m}]$; also, Electrical resistance $[\Omega]$
R_a = Average CRI over the Munsell color samples R1–R8
R_{L} = Lambertian ratio
$\mathcal{R}$ = Intensity, power, or energy reflectance
R = Electron–hole injection rate in a semiconductor $[\mathrm{s}^{-1} \cdot \mathrm{m}^{-3}]$
$\mathfrak{R}$ = Responsivity of a photon source $[\mathrm{W} \cdot \mathrm{A}^{-1}]$
RC = Resistor–capacitor combination
RC = Resonant-cavity
RCLED = Resonant-cavity light-emitting diode
rg = Chromaticity coordinates associated with Red/Green/Blue color space
rgb = Normalized tristimulus values associated with Red/Green/Blue color space
$\bar{r}\bar{g}\bar{b}$ = Color matching functions associated with Red/Green/Blue color space
RGB = Basis vectors and designation for Red/Green/Blue color space
RGB = Tristimulus values associated with Red/Green/Blue color space
$\mathfrak{RGB}$ = Primaries associated with Red/Green/Blue color space
RGC = Retinal ganglion cell
RMS = Root-mean square
RoHS = Regulation on Hazardous Substances (E.U.)
R1–R8 = Standard Munsell color samples
R9–R15 = Special Munsell color samples

s = Length or distance $[\mathrm{m}]$
$\mathrm{sinc}(\cdot)$ = Symmetric function with peak value of unity at 0 $[\mathrm{sinc}(t) \equiv \sin(\pi t)/(\pi t)]$
$\mathcal{S}$ = Spin angular momentum (helicity) $[\mathrm{J} \cdot \mathrm{s}$ if circularly polarized$]$
S = Spin angular-momentum quantum number
S = Transition strength (oscillator strength) $[\mathrm{m}^2 \cdot \mathrm{Hz}]$
$\mathcal{S}$ = Poynting vector $[\mathrm{W} \cdot \mathrm{m}^{-2}]$
$\mathbf{S}$ = Poynting vector complex amplitude $[\mathrm{W} \cdot \mathrm{m}^{-2}]$
$S_\lambda(\lambda_0)$ = Wavelength-based power spectral density $[\mathrm{W} \cdot \mathrm{m}^{-1}]$;
$S(\nu)$ = Intensity spectral density $[\mathrm{W} \cdot \mathrm{m}^{-2} \cdot \mathrm{Hz}^{-1}]$; also, Frequency-based power spectral density $[\mathrm{W} \cdot \mathrm{Hz}^{-1}]$
$\mathfrak{S}$ = Si, Ge, Sn, Ti, Zr, Al, Ga, In, Sc, Hf, Y, La, Nb, Ta, Bi, Gd, or combination thereof
S = Singlet state
SCF = Single-core fiber
S-cone = Short-wavelength (blue) retinal cone (wavelength sensitivity: 420–430 nm)
SFS = Sodium fluorosilicate doped with manganese ions (Na_2SiF_6:Mn^{4+})
SI = International system of units; also, Step-index
SMD = Surface-mounted device
SMF = Single-mode fiber

SMOLED = Small-molecule organic light-emitting diode
SNR = Signal-to-noise ratio
SPAD = Single-photon avalanche diode
SPE = Single-photon emitter
SPIE = The International Society for Optical Engineering
SQW = Single quantum well
sRGB = Standard Red/Green/Blue

t = Time [s]
t_{sp} = Spontaneous lifetime [s]; also, Effective spontaneous lifetime [s]
t = Complex amplitude transmittance
T = Temperature [K]
$\mathcal{T}$ = Intensity or power transmittance
$\mathcal{T}$ = Period of an optical wave ($\mathcal{T} = 1/\nu$ where $\nu = \mathrm{frequency}$) [s]; also, Counting time [s]; also, Temporal width [s]
$\mathcal{T}_2$ = Electron–phonon collision time [s]
T = Triplet state
TADF = Thermally activated delayed fluorescence
TEM = Transverse electromagnetic
TIR = Total internal reflection
TM = Transverse magnetic
TMD = Transition-metal dichalcogenide
TOLED = Transparent organic light-emitting diode
TPD = Triphenyl diamine derivative

$u(\mathbf{r}, t)$ = Wavefunction of an optical wave
$\mathfrak{u}$ = Number of electrons in an atomic subshell
$U(\mathbf{r}, t)$ = Complex wavefunction of an optical wave
$U(\mathbf{r})$ = Complex amplitude of a monochromatic optical wave
UCS = Uniform color space
uv = Chromaticity coordinates associated with CIE 1960 UCS color space
$u'v'$ = Chromaticity coordinates associated with CIE 1976 UCS color space
U.S. = United States
UV = Ultraviolet
UVA = Ultraviolet-A band, stretching from 315 to 400 nm
UVB = Ultraviolet-B band, stretching from 280 to 315 nm
UVC = Ultraviolet-C band, stretching from 100 to 280 nm

$\mathcal{v}$ = Velocity of an atom or object [$\mathrm{m} \cdot \mathrm{s}^{-1}$]; also, Fermi velocity [$\mathrm{m} \cdot \mathrm{s}^{-1}$]
V = Volume [m^3]; also, Modal volume [m^3]; also, Voltage [V]
V_0 = Built-in potential difference in a p–n junction [V]
$\mathcal{V}(\mathbf{r}, t)$ = Potential energy [J]
$\mathcal{V}_0$ = Rectangular barrier height [J]; also, Energy depth of a quantum well [J]
$V(\nu)$ = Fourier transform of the complex wavefunction of an optical pulse
$V(\lambda_0)$ = Photopic luminous efficiency function (photopic luminosity function)
VPE = Vapor-phase epitaxy
VUV = Vacuum ultraviolet, stretching from 10 to 200 nm

$\mathcal{w}$ = Integrated intensity in units of photon number
$\mathcal{W}$ = Electromagnetic energy density [$\mathrm{J} \cdot \mathrm{m}^{-3}$]

W_i = Probability density for absorption and stimulated emission $[\mathrm{s}^{-1}]$
W = Psychophysical magnitude estimate of warmth
WCG = Wide color gamut
WGM = Whispering-gallery mode
WOLED = White organic light-emitting diode
WPC = Wall-plug luminous efficiency $\eta_{\mathtt{WPC}}$ (also called wall-plug luminous coefficient)
WPE = Wall-plug luminous efficacy $\eta_{\mathtt{WPE}}$ (also called luminous efficacy of the source and overall luminous efficacy) $[\mathrm{lm/W}]$

x = Position coordinate $[\mathrm{m}]$; also, Displacement $[\mathrm{m}]$; also, Compositional mixing ratio in a compound semiconductor
$\hat{\mathbf{x}}$ = Unit vector in the x direction in Cartesian coordinates
XTE = Extraction efficiency $\eta_{\mathtt{XTE}}$ (also called transmission efficiency)
XUV = Extreme ultraviolet
xy = Chromaticity coordinates associated with CIE 1931 xyY color space
xyz = Normalized tristimulus values associated with CIE 1931 XYZ color space
$\bar{x}\bar{y}\bar{z}$ = Color matching functions associated with CIE 1931 XYZ color space
XYZ = Basis vectors and designation for CIE 1931 XYZ color space
XYZ = Tristimulus values associated with CIE 1931 XYZ color space
$\mathfrak{X}\mathfrak{Y}\mathfrak{Z}$ = Imaginary primaries associated with CIE 1931 XYZ color space

y = Position coordinate $[\mathrm{m}]$; also, Compositional mixing ratio in a compound semiconductor
$\hat{\mathbf{y}}$ = Unit vector in the y direction in Cartesian coordinates
YAG = Yttrium aluminum garnet

z = Position coordinate (Cartesian or cylindrical coordinates) $[\mathrm{m}]$
$\hat{\mathbf{z}}$ = Unit vector in the z direction in Cartesian coordinates
Z = Atomic number; also, Electronic-circuit impedance $[\Omega]$
ZPL = Zero-phonon line

Greek Symbols

α = Apex angle of a prism; also, Absorption or attenuation coefficient $[\mathrm{m}^{-1}]$; also, Designation for a particular phase of a material

β = Label for homogeneously broadened subset of atoms; also, Designation for a particular phase of a material
$\mathfrak{b}$ = Power-law exponent in the brightness–luminance relation

$\gamma_0(\nu)$ = Net gain coefficient for stimulated emission and absorption $[\mathrm{m}^{-1}]$
Γ = Spectral width of a uniform band of optical frequencies $[\mathrm{Hz}]$

δ = Designation for a particular phase of a material
$\delta(\cdot)$ = Delta function (impulse function)
Δ = Thickness of a thin optical component $[\mathrm{m}]$; also, Fractional refractive-index change in an optical fiber $[\Delta \approx (n_1 - n_2)/n_1]$
Δx = Width, change, spread, shift, increment, interval, or uncertainty of generic variable x
$\Delta\mathfrak{n}$ = Concentration of excess electron–hole pairs $[\mathrm{m}^{-3}]$
$\Delta\lambda$ = Wavelength spectral width or linewidth $[\mathrm{m}]$
$\Delta\lambda_{\mathrm{FWHM}}$ = Full-width-at-half-maximum wavelength spectral width $[\mathrm{m}]$
$\Delta\nu$ = Frequency spectral width or linewidth $[\mathrm{Hz}]$

$\Delta\nu_c$ = Frequency spectral width that is inverse of coherence time $(\Delta\nu_c = 1/\tau_c)$ [Hz]
$\Delta\nu_{\mathrm{FWHM}}$ = Full-width-at-half-maximum frequency spectral width [Hz]

ϵ = Electric permittivity of a medium $[\mathrm{F}\cdot\mathrm{m}^{-1}]$
ϵ_0 = Electric permittivity of free space $[8.8542\times10^{-12}\ \mathrm{F}\cdot\mathrm{m}^{-1}]$

ζ = Intermediate parameter for CCT calculation using McCamy's method

$\eta_{A,B}$ = Transmission efficiency along A and B ray directions
$\eta_{\mathtt{CLE}}$ = Current luminous efficacy (CLE) (also called current efficiency) $[\mathrm{cd/A}]$
$\eta_{\mathtt{EQE}}$ = External quantum efficiency (EQE)
$\eta_{\mathtt{IQE}}$ = Internal quantum efficiency (IQE)
$\eta_{\mathtt{LER}}$ = Luminous efficacy of radiation (LER) $[\mathrm{lm/W}]$
$\eta_{\mathtt{LUC}}$ = Luminaire wall-plug luminous efficiency
$\eta_{\mathtt{LUM}}$ = Luminaire wall-plug luminous efficacy $[\mathrm{lm/W}]$
$\eta_{\mathtt{PCE}}$ = Power-conversion efficiency (PCE) (also called energy-conversion efficiency (ECE) and overall efficiency)
$\eta_{\mathtt{PLQD}}$ = Photoluminescence quantum defect (PLQD)
$\overline{\eta}_{\mathtt{PLQD}}$ = Complementary photoluminescence quantum defect $(\overline{\eta}_{\mathtt{PLQD}} = 1-\eta_{\mathtt{PLQD}})$
$\eta_{\mathtt{PLQY}}$ = Photoluminescence quantum yield (PLQY)
$\eta_{\mathtt{WPC}}$ = Wall-plug luminous efficiency (also called wall-plug luminous coefficient, WPC)
$\eta_{\mathtt{WPE}}$ = Wall-plug luminous efficacy (WPE) (also called luminous efficacy of the source and overall luminous efficacy) $[\mathrm{lm/W}]$
$\eta_{\mathtt{XTE}}$ = Extraction efficiency (XTE) (also called transmission efficiency)
$\eta_{1,2,3}$ = Transmission efficiency along particular directions in an LED structure
η = Impedance of a dielectric medium $[\Omega]$
η_0 = Impedance of free space $[376.73\ \Omega]$

θ = Angle; also, Half vertex angle (half radiation angle) measured from emission-plane normal; also, Polar angle in a spherical coordinate system
$\overline{\theta}$ = Complement of angle θ $(90^\circ-\theta)$
θ_a = Acceptance angle
θ_{B} = Brewster angle
θ_c = Critical angle
$\overline{\theta}_c$ = Complement of critical angle θ_c $(90^\circ-\theta_c)$
θ_d = Deflection angle of a ray or wave imparted by a prism
θ_q = Deflection angles associated with diffraction from a thin grating
θ_s = Acceptance angle of fiber butt-coupled to a medium
$\theta_{1/2}$ = Half-angle from emission-plane normal at which intensity decreases to half its maximum
$2\theta_{1/2}$ = Viewing angle (50%-power angle)

κ = Elastic constant of a molecular harmonic oscillator $[\mathrm{J}\cdot\mathrm{m}^{-2}]$
κ = Overall photoluminescence decay rate $[\mathrm{s}^{-1}]$
κ_{nr} = Nonradiative portion of photoluminescence decay rate $[\mathrm{s}^{-1}]$
κ_{r} = Radiative portion of photoluminescence decay rate $[\mathrm{s}^{-1}]$

λ = Wavelength $[\mathrm{m}]$
λ_A = Wavelength associated with valence-band to acceptor-level transition $[\mathrm{m}]$
λ_g = Bandgap wavelength (long-wavelength limit) of a semiconductor material $[\mathrm{m}]$
λ_0 = Free-space wavelength $[\mathrm{m}]$

λ_p = Wavelength of maximum interband absorption in thermal equilibrium [m]
λ_{p} = Wavelength of maximum blackbody energy density [m]
λ_p = Peak wavelength [m]; also, Wavelength of maximum injection-electroluminescence spectral density [m]
$\overline{\lambda}$ = Median (or mean) wavelength [m]
$\lambda_{\mathtt{MIN}}$ = Short-wavelength cutoff of wavelength-based spectral density [m]
$\lambda_{\mathtt{MAX}}$ = Long-wavelength cutoff of wavelength-based spectral density [m]
Λ = Spatial period of thickness variation in a thin transparent grating [m]

μ = Magnetic permeability of a medium [$\mathrm{H \cdot m^{-1}}$]
μ_0 = Magnetic permeability of free space [1.2566×10^{-6} $\mathrm{H \cdot m^{-1}}$]
μ = Carrier mobility in a medium [$\mathrm{m^2 \cdot s^{-1} \cdot V^{-1}}$]

ν = Frequency [Hz]
ν_F = Mode spacing (free spectral range) [Hz]
ν_g = Bandgap frequency [Hz]
ν_0 = Central frequency [Hz]; also, Resonance frequency [Hz]; also, Frequency of a monochromatic wave [Hz]
ν_p = Frequency of maximum interband absorption in thermal equilibrium [Hz]
ν_{p} = Frequency of maximum blackbody energy density [Hz]
ν_p = Frequency of maximum injection-electroluminescence spectral density [Hz]
$\nu_{\mathbf{q}}$ = Frequency of $\mathbf{q}$th mode [Hz]

ξ = Optical absorbance of graphene monolayer [$\mathrm{C^2 \cdot J^{-1} \cdot m^{-1}}$]

ρ = Radial distance in spherical coordinate system [m]
$\varrho(k)$ = Wavenumber modal density (photons and electrons) [$\mathrm{m^{-2}}$]
$\varrho(\nu)$ = Optical joint density of states [$\mathrm{m^{-3} \cdot Hz^{-1}}$]
$\varrho_c(E)$ = Density of states near conduction band edge in a bulk semiconductor [$\mathrm{m^{-3} \cdot J^{-1}}$]
$\varrho_v(E)$ = Density of states near valence band edge in a bulk semiconductor [$\mathrm{m^{-3} \cdot J^{-1}}$]
$\varrho_\lambda(\lambda)$ = Spectral energy density (wavelength parameterization) [$\mathrm{J \cdot m^{-3} \cdot nm^{-1}}$]
$\varrho_\nu(\nu)$ = Spectral energy density (frequency parameterization) [$\mathrm{J \cdot m^{-3} \cdot Hz^{-1}}$]

$\sigma(\nu)$ = Transition cross section [$\mathrm{m^2}$]
$\overline{\sigma}$ = Average transition cross section [$\mathrm{m^2}$]
σ_0 = Peak transition cross section [$\mathrm{m^2}$]
$\sigma_{\max}$ = Maximum transition cross section [$\mathrm{m^2}$]
σ_x = RMS width of a generic function of x; also, Standard deviation of random variable x
σ_x^2 = Variance of generic random variable x; also, Variance of random variable x
σ_E = Energy uncertainty [J]
σ_t = Time duration of a function [s]; also, RMS temporal width [s]; also, Power-RMS temporal width [s]
σ_ν = Spectral width of a function [Hz]; also, RMS spectral width [Hz]; also, Power-RMS spectral width [Hz]
σ_{SB} = Stefan–Boltzmann constant [5.67×10^{-8} $\mathrm{W \cdot m^{-2} \cdot K^{-4}}$]
σ = Conductivity of a material [$\mathrm{\Omega^{-1} \cdot m^{-1}}$]

τ = Lifetime, decay time, delay time, risetime, intraband relaxation time, electron–hole recombination lifetime; also, Power-equivalent temporal width of a function [s]; also, Temporal scaling factor [s]
τ_c = Coherence time [s]

τ_{nr} = Nonradiative electron–hole recombination lifetime [s]
τ_{r} = Radiative electron–hole recombination lifetime [s]
τ_{RC} = RC time constant [s]
τ_{21} = Lifetime of a transition between energy levels 2 and 1 [s]

ϕ = Angle; also, Azimuthal angle in spherical or cylindrical coordinate system; also, Mean photon-flux density $[\mathrm{m}^{-2} \cdot \mathrm{s}^{-1}]$
$\phi(k)$ = Spatial angular-frequency wavefunction [Fourier transform of particle position wavefunction $\psi(x)$] $[\mathrm{m}^{1/2}]$
ϕ_ν = Spectral photon-flux density $[\mathrm{m}^{-2} \cdot \mathrm{s}^{-1} \cdot \mathrm{Hz}^{-1}]$
Φ = Mean photon flux $[\mathrm{s}^{-1}]$; also, Internal photon flux $[\mathrm{s}^{-1}]$
Φ_0 = External photon flux $[\mathrm{s}^{-1}]$
Φ_ν = Spectral photon flux $[\mathrm{s}^{-1} \cdot \mathrm{Hz}^{-1}]$
φ = Phase, phase shift, phase difference

χ = Angular frequency of a harmonic electrical signal $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$

ψ = Angle
$\psi(x)$ = Particle position wavefunction $[\mathrm{m}^{-1/2}]$

ω = Angular frequency $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$; also, Angular velocity $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$
Ω = Solid angle [sr]

Mathematical Symbols

$\overline{x}$ = Mean of x
$\langle x \rangle$ = Ensemble average over x
d = Differential
∂ = Partial differential
$\delta\{\cdot\}$ = Variation of a quantity (e.g., optical pathlength)
∇ = Gradient operator
$\nabla\cdot$ = Divergence operator
$\nabla\times$ = Curl operator
∇^2 = Laplacian operator ($\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ in Cartesian coordinates)

AUTHOR

Malvin Carl Teich is Professor Emeritus in Columbia University and Boston University. He is also a consultant to government, academia, and private industry, where he serves as an advisor in intellectual-property conflicts. He has authored or coauthored more than 400 peer-reviewed journal articles/book chapters, presented some 500 conference talks/lectures, and holds 6 patents. He is the coauthor of *Fundamentals of Photonics* (Wiley, 3rd Ed. 2019, with B. E. A. Saleh) and of *Fractal-Based Point Processes* (Wiley, 2005, with S. B. Lowen). Over the course of his career, he has spent sabbatical leaves at the University of Colorado at Boulder, the University of California at San Diego, and the University of Central Florida at Orlando.

Education. Teich's academic credentials include an S.B. degree in physics from the Massachusetts Institute of Technology, an M.S. degree in electrical engineering from Stanford University, and a Ph.D. degree from Cornell University. His bachelor's thesis, written jointly with Paul J. Schweitzer and supervised by Professor Theos J. Thompson, investigated the total neutron cross section of palladium using the fast chopper at the MIT Nuclear Reactor Laboratory, and identified a new resonance at 3 eV. His studies at Stanford included a course on Lasers and Masers taught by Professor Anthony E. Siegman, which galvanized his interest in quantum photonics. In carrying out his doctoral dissertation at Cornell, supervised by Professor George J. Wolga, he made use of the then-new gallium-arsenide laser diode developed at MIT Lincoln Laboratory to observe the nonlinear two-photon photoelectric effect in metallic sodium, and determined the two-quantum yield for the process.

Career. Teich assumed his first professional affiliation in January 1966 at MIT Lincoln Laboratory, as a member of the research group directed by Robert J. Keyes and Robert H. Kingston. In September 1967, he was recruited by Jacob Millman to join the faculty of Columbia University, where he served as a member of the Electrical Engineering Department (as Chairman from 1978 to 1980), the Applied Physics and Applied Mathematics Department, the Columbia Radiation Laboratory (founded and directed by I. I. Rabi) in the Department of Physics, and the Fowler Memorial Laboratory (directed by Shyam M. Khanna) in the Department of Otolaryngology at the Columbia University Medical Center. In 1996, he was appointed Professor Emeritus of Engineering Science and Applied Physics. In 1995, concurrently with his Emeritus Status at Columbia, he

LED Lighting: Devices and Colorimetry. Malvin Carl Teich.
Google Books. Published 2024.

was urged by Dean Charles deLisi to join Boston University as a faculty member in the Department of Electrical & Computer Engineering (as Director of the Quantum Photonics Laboratory and as a member of the Boston University Photonics Center), the Department of Biomedical Engineering (as a member of the Graduate Program for Neuroscience and the Hearing Research Center), and the Department of Physics. In 2011, he was appointed Professor Emeritus of Electrical & Computer Engineering, Biomedical Engineering, and Physics in Boston University.

Over the course of his career, his efforts in quantum photonics have been directed toward exploring the properties, behavior, and applications of classical and nonclassical light, including its generation, characterization, modulation, transmission, propagation, amplification, detection, and frequency-conversion. In computational neuroscience, he established the role of fractal stochastic processes in neural information processing. He has also concentrated on codifying the detection laws of vision and audition, an enterprise that lies at the interface of quantum photonics and computational neuroscience. He has mentored more than 30 doctoral students and a collection of postdoctoral fellows at Columbia University and Boston University.

Honors and Awards.

- Sigma Xi (1968).
- IEEE Browder J. Thompson Memorial Prize Award for the paper "Infrared Heterodyne Detection," published in *Proceedings of the IEEE* (1969). Presented from 1945 through 1997, this award recognized the best paper in any IEEE publication by an author under thirty years of age.
- John Simon Guggenheim Memorial Fellowship (1973).
- Fellow of Optica (1983).
- Fellow of the American Physical Society (1988).
- Fellow of the American Association for the Advancement of Science (1989).
- Fellow of the Institute of Electrical and Electronics Engineers (1989).
- Tau Beta Pi (1989).
- Commemorative Medal of Palacký University, Olomouc, Czech Republic (1992).
- Fellow of the Acoustical Society of America (1994).
- IEEE Morris E. Leeds Award (1997). Presented from 1958 through 2000, this award honored outstanding contributions to the field of electrical measurement.
- Life Fellow of the Institute of Electrical and Electronics Engineers (2005).
- Charles DeLisi Award of Boston University (2009).
- Fellow of SPIE–The International Society for Optics and Photonics (2011).

INDEX

Absorption, 97
  broadband light, 104, 107
  coefficient, 182
  indirect-bandgap semiconductor, 176
  occupancy probability, 177
  photon stream, 103, 107
  polychromatic narrowband light, 107
  probability density, 98, 107
  semiconductors, 173
  transition rate, 180
Activators
  term symbol, 318
Akasaki, Isamu, 306
Amano, Hiroshi, 306
Angle
  apex, 10
  Brewster, 49
  critical, 9, 19, 49
  deflection, 10
  negative, 6, 12
  of incidence, 3, 4, 8, 9
  of reflection, 3, 4
  of refraction, 8, 9
  solid, 4, 10
Antibunching, photon, 194
Auger recombination, 148, 149
Autocorrelation function, 53, 364
Axicon, 11

Bandgap
  direct, 134
  energy, 128, 130, 136
  indirect, 134
  wavelength, 136, 172
Bardeen, John, 125
Beam director, 11
Beamsplitter
  random partitioning, 79
  rays, 11
  single photon, 67
  waves, 67, 80
Binomial distribution
  count signal-to-noise ratio, 80
  mean photon number, 80
  photon-number variance, 80
Blackbody radiation
  Planckian locus, 293
  rate equations, 112
  Rayleigh–Jeans formula, 114
  spectrum, 113, 115, 293
  Stefan–Boltzmann law, 116
  ultraviolet catastrophe, 114
  Wien's law, 116
Bloch
  modes, 130
Boltzmann, Ludwig, 84
  Boltzmann distribution, 90
  Boltzmann's constant, 87
Bonding
  covalent, 127, 165
  ionic, 127
  metallic, 127
  van der Waals, 127, 165
Bose–Einstein condensate, 87, 93, 113
Bose–Einstein distribution, 79, 92–94
  count signal-to-noise ratio, 94
  mean photon number, 93, 94, 112
  normalization, 94
  photon-number variance, 94
Boundary, planar
  reflection, 8, 35
  refraction, 8, 35
Bragg grating
  distributed Bragg reflector, 206
Brattain, Walter H., 125
Brillouin
  zone, 132

Casimir effect, 65
Catadioptric system, 13
Cavity
  electromagnetic optics, 63
  frequencies, 100
  modal density, 101
  photon optics, 64
  planar, 100
Chromaticity diagrams, 286, 292
  xy, 292
  hue, 292
  luminance, 292
  Planckian locus, 292
  saturation, 292
Coherence
  degree of temporal, 54, 110

effective, 55
length, 55, 58
time, 54, 58
Collimator, 5, 13, 17
Color
complementary, 272
model, 286
opponent, 249, 273
shade, 286
tint, 286
tone, 286
vision, 236–265
Color spaces, 292
1976 CIELAB, 269, 276
1976 CIELUV, 269, 276, 278
CAM16, 276
CAM16-UCS, 269
chromaticity diagrams, 286, 292
Grassmann's laws, 272
LMS, 279
LUV, 292
metameric white light, 274
saturation, 288
tristimulus values, 279, 280
UCS, 292
xyY, 282
XYZ, 280
Color temperature, 300
Planckian locus, 293
Color-matching function, 280
Color-rendering index, 300, 301, 359
Colorimetry, 266–305
Cone
fundamentals, 246
Convolution, 364
Correlated color temperature, 300
cool white, 299
daylight, 299
warm white, 299
Correlation, 364
Craford, M. George, 333
Critical angle, 9, 19, 49
Cross section, 97, 99, 103, 106
peak, 109

Diffraction grating, 42
spectrum analyzer, 43
Dipole wave, 48
Distributed Bragg reflector, 206

Edison, Thomas, 85
Efficacy
current, 261
wall-plug luminous, 260
Efficiency
droop, 217, 219, 230, 345
energy-conversion, 207
external, 206
internal, 151, 202
wall-plug luminous, 261, 295
Einstein, Albert, 61
A and B coefficients, 105
emission and absorption, 85
photon, 63
Electroluminescence, 183, 186
injection, 183, 186
Electromagnetic waves, *see* Waves, electromagnetic
Emissivity
graybody, 117, 118
incandescent lamp, 118, 295
wavelength-dependent, 119
Energy levels, 89
bandgap energy, 128–130, 136
conduction band, 128
crystal-field theory, 319
degeneracy, 91
forbidden band, 128
ligand field theory, 319
occupation, 90
quantum dots, 163
valence band, 128
Energy, optical, 4
Entropy, 113
Equipartition theorem, 89
chilled-out DOFs, 114
degrees-of-freedom (DOFs), 89
failure in quantum systems, 114
frozen-out DOFs, 89
Excitons
bulk semiconductors, 172
organic semiconductors, 221
quantum dots, 163
quantum-confined, 194

Fermat, Pierre de, 1
Fermat's principle, 2
Fermi
–Dirac distribution, 91, 143
energy, 91
function, 91, 142
inversion factor, 182
level, 143, 144
tail, 144
velocity, 140
Fermion
Dirac-, 140
Fiber, optical
acceptance angle, 20, 21
butt-coupled, 22
cladding, 18
coupling efficiency, 21
graded-index, 49
meridional ray, 19
multimode, 49
numerical aperture, 20–22
silica-glass, 21
single-mode, 49

skewed ray, 20
step-index, 49
uncladded, 21
Flicker-fusion threshold, 218
Fluorescence, 221
Fourier transform, 363–369
definition, 363
inverse, 363
pair, 363
properties, 364
table, 365
Frequency
infrared, 34, 46
optical, 34, 46
ultraviolet, 34, 46
visible, 34, 46
Fresnel
approximation of spherical wave, 32
biprism, 39
equations, 49
Huygens–Fresnel principle, 24
lens, 17
reflection, 204

Graphene photonics, 139
Grassmann, Hermann, 266, 271
Grassmann's laws, 272
Graybody
emissivity, 117, 118, 295
incandescent lamp, 118, 295
radiation, 117, 118
Group-IV photonics, 135, 139
2D materials, 140
allotropes, 139
graphene photonics, 139
SiC Schottky diode, 232
silicon photonics, 231
transition-metal dichalcogenides, 140
Guild, John, 266

Hero's principle, 3
Heterostructures
organic semiconductors, 221
Holonyak, Nick, 333
Huygens, Christiaan, 24
Huygens–Fresnel principle, 24

Illuminance, 256
Imaging
aberration, 13
aberration-free, 14
equation, 7, 15, 42
single-photon, 67
Incandescent lamp, 118, 295
carbon-filament, 118
gas-mantle, 119, 268
halogen, 119, 335
tungsten-filament, 118, 294, 335
Insulators
band structure, 128
Integrated
optics, 231
photonics, 231
Ionization
energy of a donor electron, 139
energy of H, 139
Irradiance, 26, 254, 255
Isotropic
radiator, 256

Kelvin, Lord, 84
temperature scale, 86
Kirchhoff, Gustav
blackbody, 85, 111, 116

Lambertian radiator, 209, 256
Laser
silicon Raman, 231
Laser diodes (LDs), 229
LED Lighting
Color spaces, 277
phosphor-conversion LEDs, 306–332
LED lighting, 333–362
additive color mixing, 340
array devices, 341
COB devices, 329
complementary colors, 311
discrete LEDs, 308
Eiffel Tower, 339
electronic circuitry, 339
Empire-state building, 339
hybrid, 344
luminaire luminous efficacy, 354
luminaires, 354
organic, 355
performance comparison, 359
Philips L-prize lamp, 344, 347
relative merits, 335
retrofit lamps, 346
retrofit lamps, adjustable CCT, 350
retrofit lamps, adjustable color, 351
retrofit lamps, adjustable hue, 351
retrofit lamps, wireless, 357
rope lighting, adjustable color, 353
salutary features, 337
strip lighting, adjustable color, 353
surface-mounted devices, 312
traditional technologies, 335
white, 309
YAG:Ce^{3+} phosphor, 311
Lens
aberrations, 16
aspheric, 14, 16
biconcave, 16
biconvex, 14, 16, 40
compound, 16
concave, 16
converging, 16

convex, 16
cylindrical, 16
diverging, 16
dome, 209, 312, 341
double-convex, 14
focal length, 15, 16, 39, 40
focusing, 41
Fresnel, 17
Fresnel biprism, 17
glass, 14
graded-index, 18, 42
imaging, 16, 41
LED, 209
meniscus, 16
plano-concave, 16
plano-convex, 16, 39
simple, 16
spherical, 14, 40
thin, 15
Light, speed of, 2, 26, 45, 63
Light-emitting diodes (LEDs), 200–235
bioinspired, 205
blue, 213
characteristics, 201–211
current modulationi format, 218
device structures, 211
die geometries, 205
discrete, 329
drive circuitry, 216
edge-emitting, 214
energy-conversion efficiency, 207
external efficiency, 206
extraction efficiency, 203
green, 213
green gap, 344
illumination applications, 334
indication applications, 201
internal efficiency, 202, 207
light–current curve, 208
materials, 211
optics, 13
optics for, 209
orange, 210, 213, 214, 333
organic, 220
output photon flux, 206
overall efficiency, 207
perovskite, 223
photonic-crystal, 206
plasmonic, 203
power-conversion efficiency, 207
quantum-dot, 219
red, 210, 213, 214, 333
resonant-cavity, 207
response time, 210
responsivity, 207
roughened-surface, 206
spatial pattern, 208
spectral distribution, 209
surface-emitting, 214
trapping of light, 10
violet, 213
wall-plug efficiency, 207
WOLED, 223
YAG:Ce^{3+} phosphor, 311
yellow, 210, 213, 214, 333
Line broadening
Gaussian, 369
homogeneous, 110
inhomogeneous, 110
lifetime, 108, 110
lineshape function, 99, 103
Lorentzian, 108–110, 368
spectral packet, 110
Linewidth, 57, 99
relation to coherence time, 58
Losev, Oleg V., 170
Luminance, 256, 282
Luminescence
electroluminescence, 183
fluorescence, 221
phosphorescence, 221
photoluminescence, 309
Luminous
efficacy, 256, 356, 359
efficiency, 256
flux, 255, 356, 359
intensity, 209, 255
wall-plug efficacy, 359

Magnification
lens, 16
spherical boundary, 13
spherical mirror, 7
Maxwell, James Clerk, 24
Maxwell's equations, 44, 46, 50
Mean, 76, 366
Medium
homogeneous, 3, 26
inhomogeneous, 2, 27, 47
Metals
band structure, 128
conductivity, 128
Metameric white light, 274
Miniband, 160, 194
Mirror
collimating, 5
elliptical, 5
focal length, 5–7
focal point, 5
focusing, 5, 7
imaging, 5, 7
paraboloidal, 5, 6
planar, 5, 35
radius of curvature, 6
reflection, 3
spherical, 5
Multiquantum
well, 159

Nakamura, Shuji, 306
Negative-binomial distribution, 95, 121
  mean photon number, 95
  photon-number variance, 96
Newton, Isaac, 1
  laws of motion, 87
Neyman Type-A distribution, 310
Nobel laureates, 125, 306, 307, 332
  Nobel lectures, 167, 168, 235, 332

Optical coherence tomography, 244
Optical materials
  2D, 140
  TMDs, 140
Optical pathlength, 2, 4, 36
Optics
  aspheric, 14
  classical, 25
  electromagnetic, 44
  electroweak theory, 62
  first-order, 6
  Gaussian, 6
  geometrical, 2
  nonclassical, 25
  photon, 62
  quantum, 62
  ray, 2
  statistical, 51
  theories of, 62
  wave, 26
Organic semiconductors, 165

Paraboloidal wave, 32, 48
Paraxial
  approximation, 6
  Helmholtz equation, 33
  wave, 33, 48
Parseval's theorem, 364, 368
Pauli exclusion principle, 91, 131
Periodic table
  semiconductors, 134
Phosphorescence, 221
Photoluminescence, 309
  quantum dots, 164
Photometry
  color-matching function, 280
  current luminous efficacy, 261
  illuminance, 255, 256
  luminance, 255, 256, 282
  luminous efficacy, 256, 260
  luminous efficiency, 256, 261
  luminous energy, 255
  luminous flux, 255
  luminous intensity, 209, 255
  photopic luminous efficiency function, 247, 255
Photon, 63
  antibunching, 194
  at a beamsplitter, 67
  energy, 63, 65, 69
  frequency, 65
  imaging, 67
  interactions with atom, 96
  mode, 64, 66, 69
  momentum, 63, 69
  monochromatic, 68, 69
  orbital angular momentum, 63
  period, 66
  polarization, 69
  polychromatic, 68
  position, 66
  position and time, 68, 69
  rest mass, 63
  spin angular momentum, 63, 69
  time, 67
  wave-particle duality, 66
  wavelength, 66
  wavelike character, 63
  wavepacket, 68
Photon streams, 69
  absorption, 103, 107
  energy spectral density, 72
  intensity spectral density, 72
  photon flux, 71, 73
  photon number, 72, 73
  photon-flux density, 71, 73
  photon-number statistics, 75
  power spectral density, 72
  randomly partitioned, 79
  randomness, 73
  registration locations, 74
  registration times, 74
  spectral photon flux, 72
  spectral photon number, 72
  spectral photon-flux density, 72
  stimulated emission, 103, 107
  time-varying light, 73
Photon-number distribution
  Bernoulli, 79
  binomial, 80
  Bose–Einstein, 79, 93
  doubly stochastic, 78
  geometric, 79, 93
  negative-binomial, 95, 121
  Neyman Type-A, 310
  Poisson, 75
Photon-number statistics, 75
  chi-square integrated intensity, 121
  coherent light, 75
  count signal-to-noise ratio, 77
  count standard deviation, 76
  counting distribution, 75
  counting statistics, 76
  counting time, 75
  doubly stochastic, 78, 121
  exponential integrated intensity, 79
  luminescence, 310
  mean photon number, 75, 76, 78

partially coherent light, 78, 121
photon-number distribution, 75
photon-number variance, 76, 78
Poisson, 75
Poisson transform, 78
randomly deleted, 79
randomly partitioned, 79, 80
thermal light, multimode, 95, 121
thermal light, single-mode, 92
Photonic
integrated circuits, 231
Photopic luminous efficiency function, 247, 255
Planck, Max, 61
blackbody radiation, 111, 118
Planck spectrum, 63, 85, 118
Planck's constant, 65, 157, 367
Planckian locus, 292, 293
radiation law, 113, 118
weak illumination image, 70
Plane wave, 30, 47
Poisson distribution, 75
count signal-to-noise ratio, 77
derivation, 76
mean photon number, 77
normalization, 77
photon-number variance, 77
randomly partitioned, 81
Polarization, 49, 50
Poynting vector, 47
Prism, 10, 38
apex angle, 10
axicon, 39
biprism, 11, 39
deflection angle, 10
Fresnel biprism, 11, 39
thin, 10

Quantum dots, 162
artificial atoms, 163
core–shell, 164, 217
excitons, 163
fabrication, 162
LEDs, 324
photoluminescence, 164
self-assembly, 163
silicon photonics, 231
single-photon emitter, 194
synthesis, 163
Quantum mechanics, 89, 367
Quantum well, 156–161
Quantum wire, 161
Quantum-confined
excitons, 194
structures, 156–166
Quist, Robert J., 199
Quist, Theodore, 199

Radiance, 254
Radiometry
irradiance, 26, 255
radiance, 255
radiant energy, 255
radiant flux, 255
radiant intensity, 255
Raman
silicon laser, 231
Random waves, *see* Waves, random
Rays, 1–23
caustic, 6, 16
convergence, 17
divergence, 17
paraxial, 6
scattered, 17
tracing, 8
Reflectance
power, 49
role of polarization, 49
Reflection
critical angle, 9, 19
law of, 3
planar boundary, 4, 35
spherical boundary, 12
total internal, 9–11, 13, 18, 49
Refraction
dielectric boundary, 13
external, 8
internal, 8
planar boundary, 4, 8, 35
Snell's law, 4, 9, 12
spherical boundary, 12
Refractive index, 2, 26, 45
Responsivity
LED, 207
Retina
OCT imaging, 244
Round, Henry J., 170

Scalar waves, *see* Waves, scalar
Sellmeier equation, 195
Semiconductors
k-selection rule, 174
p–n junction, 153
p–n junction, biased, 154
absorption, 172, 182
AlGaAs, 213
AlInGaN, 214
AlInGaP, 213
allotropes, 139
Auger recombination, 148, 149
bandgap energy, 128, 130, 136
bandgap wavelength, 136, 172
Brillouin zone, 132
bulk, 129, 171–183
carborundum, 135
carrier concentrations, 144, 147
carrier generation, 148
carrier injection, 149
carrier recombination, 148

carriers, 131
degenerate, 147
density of states, 141
density of states, joint, 175
depletion layer, 153
direct-bandgap, 134
dopants, 138
effective mass, 132
electroluminescence, 186–193
elemental, 135
energy bands, 128, 129
energy–momentum relations, 131
excitons, 172, 221, 222
extrinsic, 138
Fermi function, 142
Fermi inversion factor, 182
fundamentals, 125–166
GaAs, 212
GaAsP, 212
GaAsP:N, 213
gain coefficient, 181
GaN, 214
heterojunction, 155
II–VI materials, 137–138
III–nitride materials, 135, 213–214
III–V materials, 135–137, 211–214
indirect-bandgap, 134
InGaN, 213
internal efficiency, 151
intrinsic, 138
Kronig–Penney model, 129
law of mass action, 146
minibands, 160, 194
multiquantum-well, 159
nanocrystals, 162
nonradiative recombination, 148
occupancy probabilities, 142, 177
organic, 165
periodic table, 134
quantum dots, 162
quantum wells, 156–161
quantum wires, 161
quantum-confined, 156, 193
quasi-equilibrium, 147
recombination coefficient, 149
recombination lifetime, 150
refractive index, 195
Shockley equation, 155
SiC, 135
silicon photonics, 231
superlattice, 160, 194
transition probabilities, 178
Semimetals
band structure, 128, 137, 138
graphene, 140
massless Dirac fermions, 140
Shockley, William B., 125
Silicon photonics, 231
direct-mounting integration, 231
flip-chip integration, 231
heteroepitaxy, 232
heterogeneous integration, 232
hybrid approach, 232
PIC, 231
Snell's law, 4, 36
paraxial, 9
Solids
covalent, 127
ionic, 127
metallic, 127
molecular, 127
van der Waals, 127, 141
Spatial
LED emission pattern, 208
Spectral density, 55
Spectral width, *see* Linewidth
Spectrum, *see* Spectral density
Spherical
boundary, imaging, 12
wave, 31, 48
Spin
-allowed transitions, 221
-forbidden transitions, 221
-orbit coupling, 222
electron, 142
photon, 142
singlet state, 221
triplet state, 221
Spontaneous emission, 96
frequency distribution, 103
into a band of modes, 107
into a prescribed mode, 97, 107
into all modes, 102
into any mode, 99, 107
lifetime, 102, 103
lifetime, effective, 104
occupancy probability, 177
probability density, 97, 102
Purcell factor, 195, 203, 207
semiconductors, 173
spectral density, 180
transition rate, 180
Standard deviation, 76, 366
Stefan–Boltzmann law
blackbody radiation, 116
graybody radiation, 117
Stefan–Boltzmann constant, 116
thermal radiation, 119
Stimulated emission, 98
broadband light, 104, 107
occupancy probability, 177
photon stream, 103, 107
polychromatic narrowband light, 107
probability density, 98, 107
semiconductors, 173
semiconductors, indirect-gap, 176
transition rate, 180
Superlattice, 160, 194

Swan, Joseph, 85

Tail
band, 147
Fermi, 144
Urbach, 183
Temperature
absolute zero, 86
blackbody spectrum, 114
Celsius scale, 86
correlated color, 297, 300
examples, 86, 87
Fahrenheit scale, 86
ideal gas law, 87
internal energy, 87, 88
Kelvin scale, 86
kinetic theory of gases, 88
thermographic images, 120
Temperature, color, 292
Temporal coherence function, 53
Thermal
equilibrium, 87
light, 116, 119
mode average energy, 95, 113
quasi-equilibrium, 87
radiation, 116, 119
Thermography, 119
applications, 120
hyperspectral, 120
thermal camera, 119
Transition
excitonic, 172, 194
free-carrier, 172
impurity-to-band, 172
interband, 172, 194
intersubband, 194
intraband, 172
miniband, 194
phonon, 172
strength, 99, 102
Transmittance
axicon, 39
biconvex lens, 40
biprism, 39
diffraction grating, 42
graded-index lens, 42
plano-convex lens, 39
power, 49
prism, 38
spherical lens, 40
transparent plate, 36
Transparent plate
arbitrary, 36
fixed thickness, 36
varying refractive index, 37
varying thickness, 38

Uncertainty relation
position–momentum, 69, 195, 367
position–spatial-frequency, 367
time–angular-frequency, 68, 367
time–energy, 68, 69, 367
time–frequency, 68, 366
Units, radiometric and photometric, 253
current luminous efficacy, 261
illuminance, 255, 256
irradiance, 26, 255
luminance, 255, 256, 282
luminous efficacy, 256, 260
luminous efficiency, 256, 261
luminous energy, 255
luminous flux, 255
luminous intensity, 209, 255
radiance, 255
radiant energy, 255
radiant flux, 255
radiant intensity, 255

Vacuum field, *see* Zero-point energy
Velocity
phase, 31
von Helmholtz, Hermann, 236

Wavefunction
electron, 131
Wavelength, 31
bandgap, 136, 172
de Broglie, 157
infrared, 34, 46
optical, 34, 46
ultraviolet, 34, 46
visible, 34, 46
Waves, 24–60
Waves, electromagnetic, 44
boundary counditions, 45
cavity, 63
complex envelope, 47
dipole, 48
electric permittivity, 45
elementary, 47
energy density, 45
energy in a mode, 64
Helmholtz equation, 47
impedance, 48
intensity, 45, 47, 67, 73
magnetic permeability, 45
Maxwell's equations, 44, 46, 50
monochromatic, 46
paraboloidal, 48
paraxial, 48
plane, 47
power, 45, 47
power reflectance, 49
Poynting vector, 45
relation to scalar waves, 50
spherical, 48
superposition, 45
TEM, 47

wave equation, 44, 50
Waves, random, 51
aurocorrelation function, 53
coherence length, 55
coherence time, 54
degree of temporal coherence, 54
effective coherence, 55
ensemble average, 52
ergodic process, 53
instantaneous intensity, 52
intensity, 52
linewidth, 57
power spectral density, 55
random intensity, 52
spectral density, 55
spectral radiant flux, 55
spectral width, 57
spectrum, 55
stationarity, 52
temporal coherence function, 53
wavelength spectral density, 55
Wiener–Khinchin theorem, 56
Waves, scalar, 26
coherent, 28
complex amplitude, 28
complex envelope, 30
complex wavefunction, 28
deterministic, 28
elementary, 30
energy, 26, 27
Helmholtz equation, 29
intensity, 26, 29
monochromatic, 27
paraboloidal, 32
paraxial, 33
plane, 30
power, 26, 27
relation to electromagnetic waves, 50
spherical, 31
spherical-wave wavefunction, 32
superposition, 26
wave equation, 26, 28
wavefronts, 29
wavefunction, 26, 27
wavenumber, 31
wavevector, 30
Width of a function
$1/e$, 368
$1/\sqrt{2}$, 368
3-dB, 368
duration–bandwidth reciprocity, 366
FWHM, 368
measures of, 366
positive-frequency, 368
power-equivalent, 367
power-RMS, 366
RMS, 366
spectral, 366
temporal, 366
Wien's law
blackbody radiation, 116
graybody radiation, 118
Wien's constant, 116
Wiener–Khinchin theorem, 56
WOLED, 223
Wright, W. David, 266
Wärmestrahlung, 117

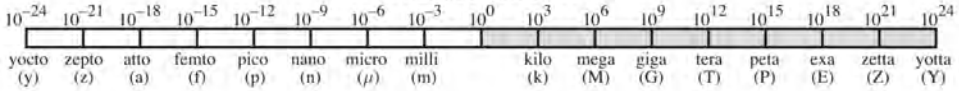
Young, Thomas, 236

Zero-point energy, 65, 93, 98

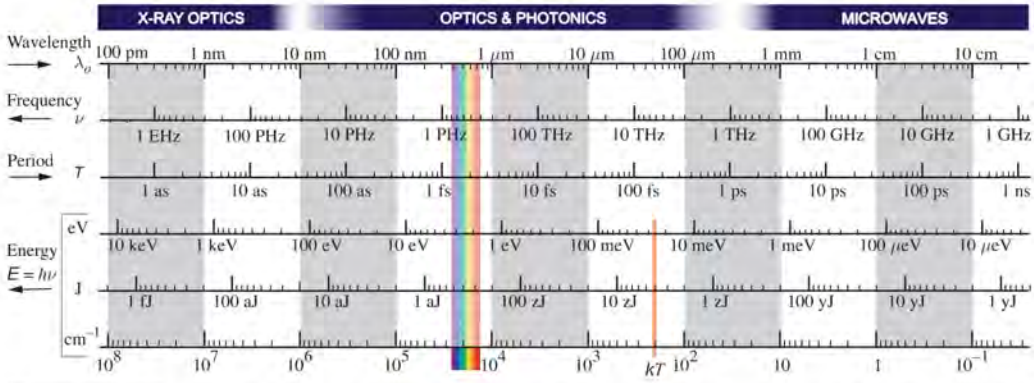
USEFUL CONSTANTS

Speed of light in free space	c_o	2.9979×10^{8}	m/s	Planck's constant	h	6.6261×10^{-34}	J · s
Permittivity of free space	ϵ_o	8.8542×10^{-12}	F/m	Electron charge	e	1.6022×10^{-19}	C
Permeability of free space	μ_o	1.2566×10^{-6}	H/m	Electron mass	m_0	9.1094×10^{-31}	kg
Impedance of free space	η_o	376.73	Ω	Boltzmann's constant	k	1.3807×10^{-23}	J/°K

PREFIXES FOR UNITS

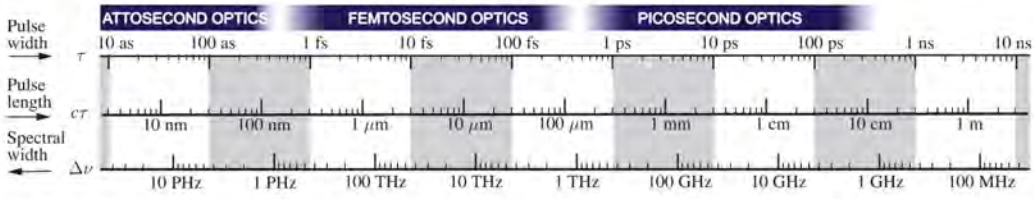


THE PHOTON



A photon of free-space wavelength $\lambda_O = 1\ \mu$m has frequency $\nu = 300$ THz, period $T = 3.33$ fs, and energy $E = 1.24$ eV $= 199$ zJ $= 10^4$ cm^{-1}. At room temperature ($T = 300^o$ K), the thermal energy $kT = 26$ meV $= 4.14$ zJ $= 209$ cm^{-1}.

OPTICAL PULSES



PHOTONIC STRUCTURES

LED LIGHTING

Devices and Colorimetry

ABOUT THIS BOOK: ***LED Lighting*** is a self-contained and introductory-level book featuring a blend of theory and applications that thoroughly covers this important interdisciplinary area. Building on the underlying fields of optics, photonics, and vision science, it comprises four parts: **PART I** is devoted to fundamentals. The behavior of light is described in terms of rays, waves, and photons. Each of these approaches is best suited to a particular set of applications. The properties of blackbody radiation, thermal light, and incandescent light are derived and explained. The essentials of semiconductor physics are set forth, including the operation of junctions and heterojunctions, quantum wells and quantum dots, and organic and perovskite semiconductors. **PART II** deals with the generation of light in semiconductors, and details the operation and properties of III–V semiconductor devices (MQWLEDs & μLEDs), quantum-dot devices (QLEDs & WQLEDs), organic semiconductor devices (OLEDs, SMOLEDs, PLEDs, & WOLEDs), and perovskite devices (PeLEDs, PPeLEDs, QPeLEDs, & PeWLEDS). **PART III** focuses on vision and the perception of color, as well as on colorimetry. It delineates radiometric and photometric quantities as well as efficacy and efficiency measures. It relays the significance of metrics often encountered in LED lighting, including the color rendering index (CRI), color temperature (CT), correlated color temperature (CCT), and chromaticity diagram. **PART IV** is devoted to LED lighting, focusing on its history and salutary features, and on how this modern form of illumination is deployed. It describes the principal components used in LED lighting, including white phosphor-conversion LEDs, chip-on-board (COB) devices, color-mixing LEDs, hybrid devices, LED filaments, retrofit LED lamps, LED luminaires, and OLED light panels. It concludes with a discussion of smart lighting and connected lighting. Each chapter contains highlighted equations, color-coded figures, practical examples, and reading lists.

ABOUT THE AUTHOR: ***Malvin Carl Teich, PhD,*** is Professor Emeritus at Columbia University and Boston University. He is the coauthor of ***Fundamentals of Photonics*** (with B.E.A. Saleh) and ***Fractal-Based Point Processes*** (with S.B. Lowen). He is a Fellow of the AAAS, APS, ASA, IEEE, OPTICA, SPIE, and the Guggenheim Foundation. He is the recipient of the IEEE Browder Thompson Memorial Prize, the IEEE Morris Leeds Award, the Commemorative Medal of Palacký University, and the Charles DeLisi Award of Boston University.

ISBN 979-8-9901705-0-6

90000>

9 798990 170506