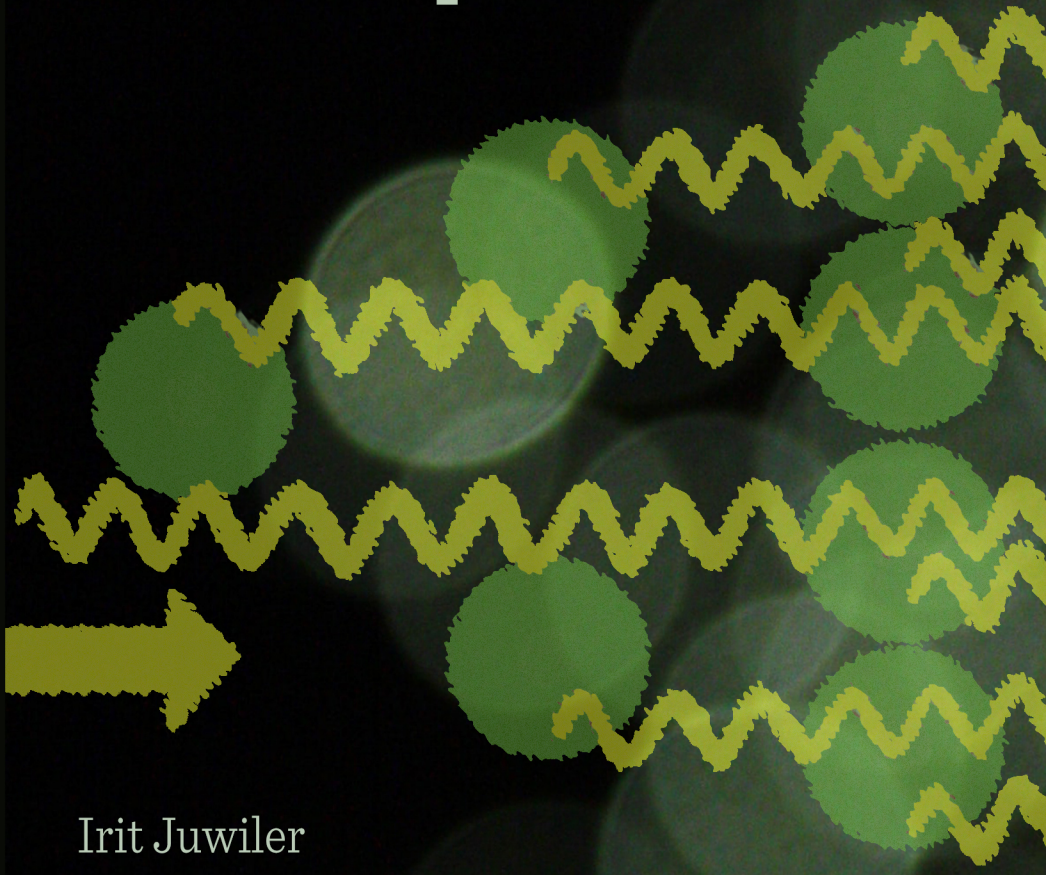
The Physical Fundamentals of Electro-Optics



Irit Juwiler
Nathan Blaunstein

The Physical Fundamentals of Electro-Optics

The Physical Fundamentals of Electro-Optics

By

Irit Juwiler and Nathan Blaunstein

Cambridge
Scholars
Publishing

The Physical Fundamentals of Electro-Optics

By Irit Juwiler and Nathan Blaunstein

This book first published 2022

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

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

ISBN (10): 1-5275-8476-3
ISBN (13): 978-1-5275-8476-1

CONTENTS

PREFACE

This book is intended to appeal to any practicing optical scientist and optical engineer who is concerned with the design, operation, and service of wired (fiberoptic) and wireless optical systems for resolving both the direct and the inverse problems of optical communication and optical location, namely of LIDAR. It will be very useful for students of all three degree levels, B.Sc., M.Sc., and Ph.D., who are concerned with the performance of mathematical algorithms, theoretical and applied models, as well as with the design, construction, and servicing of different optical devices: from various kinds of laser, photodetectors, light emitters and diodes, and optical amplifiers, wired waveguide optical structures, such as fiberoptic structures – 2D and 3D – with applications in wireless (atmospheric) networks, to LIDAR applications.

It should be mentioned that during the last 20–30 years a lot of optical elements, devices, and systems have been developed and constructed to satisfy the continually increasing demands of modern optical engineering and photonics for wired and wireless communications and LIDAR applications, including wide spectra – visual, infrared (IR) ultraviolet (UV) – sensors, devices, and systems. And, if for wireless (atmospheric) communication systems numerous excellent monographs have been published (see bibliography in Refs. [1-4]), the role of fiberoptic communication has been weakly illuminated. Moreover, the foundational books regarding photonics, physical aspects of laser and optical detectors operation, and photodiodes were published twenty to thirty years ago and paid attention only to some specific fields of wide spectra optical physics applications [5–8], accounting mostly either for a wide description of solid materials, such as dielectrics and semiconductors and their role in different optical sources and detectors operation, or fiberoptic communication, ignoring basic aspects of such elements as optical emitters (lasers) and detectors. During recent decades, many articles about all of these aspects have been published, including articles by the authors of this book, but general views on the problems of fiberoptic communication and lasers and detectors as basic terminals of any wireless and wired optical communication system or network were absent. Moreover, even such books and papers that were recently published mostly paid attention to aspects of

signal coding and decoding and signal modulation, and less so the physical layers of devices, transmitters, and receivers of optical information [9–11].

We created this book to bring together all layers, where the first "layer" deals with a wide spectrum of electronic devices and circuits, which needs a careful and very transparent explanation of the physical processes occurring in the basic elements of optical emitters; lasers, detectors of optical radiation, various amplifiers, and so forth. The second "layer" illuminated in our book regards the presentation of optical signals in the channels and elements of their modulation during signal processing of the information passing through fiberoptic and wireless channels. The last "layer" deals with the physical nature of all kinds of noise occurring in each element of wired and wireless communication links – from the light emitter consisting of optical lasers to detectors consisting of laser and photodiodes. To unify all these "layers" and to create a "bridge" between them, it is important to introduce an additional layer, which the authors call the "physical and mathematical layer." Therefore all aspects described in this book regarding electro-optic engineering start from the physical explanation of the matter and then, by entering into other engineering problems of these three "layers" mentioned above, where each engineering aspect is accompanied by corresponding examples, give the reader the chance to use the obtained information for application in the performance and design of modern devices and systems for optical communication and optical location (LIDAR) applications.

At the same time, the book does not enter into technical details of how to produce different kinds of lasers, emitters, diodes, amplifiers, and optical waveguides, nor how to design different kinds of electronic devices based on semiconducting materials, assuming that for the reader it is more important to obtain fundamental knowledge about all above-mentioned elements of electro-optical engineering based on the common and joint physical "layers" on whole spectra of these elements, without entering into individual technical details and schemes.

The main goal of this book is to illuminate those questions and aspects of modern electro-optical engineering and optical physics, which were only partly illuminated in the existing literature. The authors enjoyed sharing their knowledge of teaching undergraduate and postgraduate students the physical fundamentals of classical and applied optics and photonics, optical emitters and detectors fundamentals, different aspects of wired (fiber optic) and wireless engineering, and fundamentals in optical waves propagation in fiberoptic 2-D and 3-D structures.

The book comprises ten chapters. Chapter 1 presents an introduction to the subjects that will be discussed and explained in chapters 2 to 10. It

gives the reader information on optical spectra, from UV to IR, as a part of the full electromagnetic spectra, with a general explanation of the similarity of optical and radio waves. This fundamental similarity of radio and optical waves allows Chapter 2 to present all electromagnetic aspects of optical wave propagation via the general laws of Maxwell, of optical waves via plane electromagnetic waves, their propagation in free space, the intersection between two media, and finally, in various kinds of media – from dielectric to conductive. In Chapter 3, the main laws of classical and quantum physics based on corpuscular theory and on wave-corpuscular dualism are discussed using a simple explanation of the subject with clearly presented illustrations. In this manner, the structure of a simple atom, molecules, and crystals are described based on elements of quantum mechanics and wave theory in such a manner so as not to complicate the text of the book with mathematical descriptions of differential equations and integral presentation of the basic characteristics and functions describing each element's own structure. Then, Chapter 4 describes the basic physical principles of photonics, optical emitters, and laser operation based on the quantum presentation of their structure via linearly distributed discrete spectra of each element, emitter, or detector, and based on the interaction between holes and electrons inside various kinds of semiconductors as materials of such optical elements. In Chapter 5, laser diodes (LDs), *p*-*n*- and *p*-*i*-*n*-type photodiodes, and the avalanche photodiode (APD) are described, acting as emitted sources (e.g., lasers) or receiving detectors, which have found importance in electronics, photonics, and in optoelectronic diodes, as well as in solar cells. Their operational parameters and characteristics were described in a unified manner based on the physical knowledge illustrated in Chapters 3 and 4. Chapters 3 and 5 are filled with corresponding examples to aid the reader in understanding the matter and using the obtained knowledge in practice.

In Chapter 6, different types of noise occurring in light sources (lasers) and detectors (diodes), as the initial and the later terminals of any optical communication link, whether wired (fiber optic) or wireless (atmospheric), are described in a unified manner. Chapter 7 explains the principles of operation of optical amplifiers based on various kinds of emission – stimulated and spontaneous – which compete with absorption in any semiconducting material. It is shown what type of emission gives the main impact in terms of noise and plays a positive role in the amplification of optical signals with data transmission along the link. An example of an optical amplifier based on an Erbium fiberoptic amplifier is fully described, with estimations of its full noise interference via the corresponding examples having practical meaning. In Chapter 8, types of optical signals –

continuous and discrete – are fully described with their mathematical explanation for practical applications. Chapter 9 deals with the types of modulation, analog and discrete, most practically used in optical communications. Here, more precisely, is given the relationship between the spectral presentation of analog signals after amplitude and frequency modulation to show the reader advantages of angular-frequency and phase modulation with respect to amplitude. At the end of this chapter, the corresponding examples are shown to prove these advantages. In Chapter 10, which is short but informative, the basic types and characteristics of optical waveguides – 2D and 3D – as well as of fiber optic cylindrical structures are presented, and the propagation of optical wave modes in various kinds of optical guiding structures is fully analyzed with illustrations and computing plots. Then, the main factors of dispersion – waveguide, modal, material, and polarization – are described in a unified manner accounting for the specific impact of each of these factors on noises and fading occurring in optical communication links, both wired and wireless. At the end of this chapter, the examples, having practical meaning, will be important for the reader to accumulate the knowledge obtained in previous chapters and introduce them in practice.

Bibliography

[1] Marz, R. 1995. *Integrated Optics: Design and Modeling*. Norwood, MA: Artech House.
[2] Morthier, G., and P. Vankwikelberge. 1997. *Handbook of Distributed Feedback Laser Diodes*. Norwood, MA: Artech House.
[3] Palais, J. C. 1998. *Fiber Optic Communications*, 4th Ed. New Jersey: Prentice-Hall.
[4] Johnston, W. 1997. *Erbium Doped Fiber Amplifiers,* Student Manual. Glasgow: Optoelectronic Systems.
[5] Palais, J. C. 2006. "Optical Communications." In Handbook: *Engineering Electromagnetics Applications*, edited by R. Bansal., New York: Taylor and Frances.
[6] Johnston, W. 1995–2011. *Fiber Optic Communications,* Student Manual. Glasgow: Optoelectronic Systems.
[7] Saleh, B. E. A., and M. C. Tech. 2012. *Fundamentals of Photonics,* 2nd Ed. Norwood, MA: Artech House.
[8] Krouk, E. and S. Semenov, eds. 2011. *Modulation and Coding Techniques in Wireless Communications*. NJ: John Wiley & Sons.
[9] Bansal, R. ed. 2006. Handbook: *Engineering Electromagnetics Applications*. New York: Taylor and Frances.

[10] N. Blaunstein, Sh. Arnon, A. Zilberman, and N. Kopeika. 2010. *Applied Aspects of Optical Communication and LIDAR*. New York: CRC Press, Taylor &Francis Group.
[11] Blaunstein, N., S. Engelberg, E. Krouk, and M. Sergeev. 2019. *Fiber Optic and Atmospheric Optical Communication*. Hoboken NJ: Wiley.

LIST OF ABBREVIATIONS

2-D – two dimensional

3-D – three dimensional

PDF – probability density function

CDF – cumulative distribution function

UV – ultraviolet

VIS – visible light

THz – terahertz-band wave

MW – microwaves

RW – radiowaves

IR – Infrared spectrum

IRA – near infrared

IRB – middle infrared

IRC – far infrared

ASK – amplitude shift keying

FSK – frequency shift keying

PSK – phase shift keying

RZ – return to zero

NRZ – non return to zero

$\mathbf{A(r,}\ t\mathbf{)}$ – vector of any electromagnetic component

$\mathbf{D(r,}\ t\mathbf{)}$ – electric flux induced in the medium by the electric field, in coulombs/m^3

$\mathbf{E(r,}\ t\mathbf{)}$ – electric field strength vector, in volts per meter (V/m)

$\mathbf{B(r,}\ t\mathbf{)}$ – magnetic flux induced by the magnetic field, in Webers/m^2

$\mathbf{H(r,t)}$ – magnetic field strength vector, in amperes per meter (A/m)

$\mathbf{k}$ – vector of electromagnetic field propagation

$\mathbf{j(r,t)}$ – vector of electric current density, in Amperes/m^2

$\rho(\mathbf{r},t)$ – charge density in Coulombs/m^2

$\mathbf{S} = \mathbf{E} \times \mathbf{H}$ – vector of density of wave power flow, the absolute value of which is in Watt/m^2

$\nabla\times$ – curl operator, as a measure of field rotation

$\varepsilon(\mathbf{r})$ – permittivity

$\mu(\mathbf{r})$ – permeability

$\sigma(\mathbf{r})$ – conductivity

$\varepsilon_0 = 8.854 \bullet 10^{-12} = \frac{1}{36\pi}10^{-9}$, in Farad/meter (F/m) – permittivity in free space

$\mu_0 = 4\pi \bullet 10^{-7}$, in Henry/meter (H/m) – permeability of free space

$c = \frac{1}{\sqrt{\varepsilon_0\mu_0}} = 3 \bullet 10^8$, in m/c – light velocity in free space

$\varepsilon = \varepsilon' - j\varepsilon''$ – complex form of permittivity presentation; ε' – its real part, ε'' – its imaginary part.

$n = n' - jn''$ – complex refractive index; $n' = \sqrt{\varepsilon'/\varepsilon_0}$ – its real part, $n'' = \sqrt{\varepsilon''/\varepsilon_0}$ – its imaginary part

$\lambda = c/f$, f - wave frequency in Hz=s^{-1}

$\omega = 2\pi f$ – angular frequency of wave

$\Psi(\mathbf{r})$ – scalar component of electromagnetic field

$\gamma = \alpha + j\beta$ – propagation parameter of wave

α – attenuation factor, in Neper/km

β – phase velocity factor in Radian/m

$v_{ph} = v_{ph}(\omega)$ – wave phase velocity

1 eV – electron-volt = $1.6 \bullet 10^{-19}$ Joule

$h = 6.625 \cdot 10^{-34} J \cdot s$ – Planck's constant

$P_n = h/\lambda \cdot n$ – impulse of electron of number n inside atom; λ – wavelength

$Ekn = (1/2)Pn^2/m_0$ – kinetic energy of electron of number n inside atom

$m_0 = 9.11 \cdot 10^{-31}$ kg – mass of electron in free space

$n = 1, 2, 3, \ldots, N$ – main number, related to radius of the orbit r

$l = 0, 1, 2, \ldots, n\text{-}1$ – orbital number related to the longitudinal angle θ

m_l – momentum of rotation number related to the azimuthal angle φ

$s = -1/2$ and $s = +1/2$ – left-hand and right-hand orbital spin momentum, respectively

(r, θ, φ) – spherical system of coordinate

Si – silicon

GaAs – gallium-arsenide

GaN – gallium-nitrogen

Sb – stibium

Ge – germanium

SbGe – stibium-germanium

In – indium

InGe – indium-germanium

p-n – positive-negative junction

A_{ji}, B_{ij} and B_{ji} – Einstein coefficients

D_n and τ_n – coefficient of diffusion and life-time of n-type particles (electrons), respectively

D_p and τ_p – coefficient of diffusion and life-time of p-type particles (holes), respectively

σ_n and σ_p – conductivity of electrons and holes, respectively

ρ_n and ρ_p – resistivity of electrons and holes, respectively

μ_n and μ_p – mobility of electrons and holes, respectively

e – $1.6\ 10^{-19}$ C – charge of electron (the same with + for hole)

η_i – quantum efficiency

LED – light emitting diode

LD – laser diode

E_g – depletion zone energy (energetic width of gap)

λ_g – cutoff frequency

k_B – $1.38 \cdot 10^{-23}$ *J/K* - Boltzmann constant

$\acute{R}$ – coefficient of reflection

α_r – total attenuation coefficient

P-N – positive-negative diode

Φ – flux of photons entering into the diode

R – photodetector responsivity, in Ampere per Watt (A/W)

ζ – fraction of electron-hole pairs that successfully contribute to the detector current

PiN – positive-intrinsic-negative diode

W_j – width of junction

v_e and v_h – velocities of electrons and holes

i_{ph} – photocurrent

R_L – load resistance of detector

B_ω – bandwidth of detector

APD – Avalanche photodiode

α_e and α_h – ionization coefficients for electrons and holes, respectively, as ionization probabilities per unit length, in cm^{-1}

M – multiplication factor

SAM APD – separate-absorption-multiplication APD

G – gain of APD

InGaAs APD – indium-gallium-arsenide APD

σ_I^2 – standard deviation of wave intensity

SNR (S/R) – signal-to-noise ratio

σ_i^2 – standard deviation of photocurrent noise

N_T – thermal noise

σ_q – circuit-noise parameter

η – Poisson random variable

q – Gaussian random variable

$<\eta>$ – mean square variance of Poisson noise

σ_q^2 – mean square variable of Gaussian noise

F – excess – noise factor

B – bandwidth of receiver (detector)

$<i_{nd}^2> / B$ – noise current spectral density for PiN detector

$<e_{nd}^2> / B$ – noise voltage spectral density for PiN detector

$<i_{nd}^2> / B$ – noise current spectral density for avalanche photodiode

$<e_{nd}^2> / B$ – noise voltage spectral density for avalanche photodiode

$<i_{nd}^2> / B$ – noise current spectral density for photoconductor

$<I_c^2>$ – preamplifier noise contribution

I_b and I_c – basic and collector current inside bipolar transistor

FET – field-effect transistor

EDFA – erbium doped fiber amplifier

σ_{se} – stimulated emission cross-section

$\gamma(\upsilon)$ – signal gain of amplifier

ASE – amplified spontaneous emission

$d\Omega$ – solid angle

PASE – power of Amplified Spontaneous Emission (ASE)

ρ_{ASE} – spectral power density of the Amplified Spontaneous Emission (ASE)

B_o – optical bandwidth of amplifier

$<i^2>$ – mean square current fluctuations

σ_S – thermal noise

σ_{ASE} – ASE short noise

$\sigma_{S\text{-}ASE}$ – signal-ASE beat noise

$\sigma_{ASE\text{-}ASE}$ – ASE-ASE beat noise

η_{ASE} – quantum efficiency of ASE amplifier

$PASE^{B0}$ – single polarized signal optical power at the optical bandwidth

Bo – optical bandwidth

B_e – receiver bandwidth

$PASE^{Be}$ – single polarized signal power at the receiver bandwidth

SNR_{out} – signal to noise ratio of optical amplifier

NF – noise figure

$\gamma_o(\upsilon)$ –- small signal gain coefficient

$G_0(\upsilon)$ – overall gain of erbium doped fiber amplifier (EDFA)

CW – continuous narrowband (NB) signals

WB – wideband signals

AM – amplitude modulation

FM – frequency modulation

PM – phase modulation

OF – optical frequency

km – amplitude modulation index

β_f – frequency modulation index

k_θ – phase sensitivity

β_θ – phase modulation index

$(SNR)_{in}$ – signal to noise ratio at the input of receiver

$(SNR)_{out,FM}$ – signal to noise ratio at the output of receiver for FM

$(SNR)_{out,AM}$ – signal to noise ratio at the output of receiver for AM

BPSK – binary phase shift keying

QPSK – quadrature phase shift keying

MSK – minimum shift keying

GMSK – Gaussian minimum shift keying

FSK – frequency shift keying

NA – numerical aperture

2-D – two-dimensional fiber optic structure (slab)

3-D – three-dimensional fiber optic structure (cylindrical cable)

TE – transverse electric field (vertical polarization)

TM – transverse magnetic field (horizontal polarization)

LHS – left-hand-side

RHS – right-hand-side

V – normalized frequency parameter

$J_l(hr)$ – Bessel function of the first kind

$K_l(hr)$ – modified Hankel function

LP – linearly polarized (mode)

D_M – material dispersion parameter

$\Delta\tau_g$ – time spread between two modes

$\Delta\omega$ – spectral range of the total signal

n_g – index of transmission of energy via cable

D_W – waveguide dispersion parameter

Dp – polarization mode dispersion (PMD) parameter

TIR – total intrinsic reflection

CHAPTER 1

FUNDAMENTAL ASPECTS OF ELECTRO-OPTICS

Electro-optical engineering, as a subject of analysis and discussion, covers many basic aspects which should be understood and explained to the reader, such as [1–16]:

- electromagnetic nature of light,
- similarity of optical and electromagnetic waves,
- corpuscular nature of light,
- electromagnetic aspects of optical wave propagation in various environments,
- elements of photonics,
- optical lasers – emitters of light,
- optical detectors of light – laser diodes and photo diodes,
- optical amplifiers,
- optical signals presentation – analog and digital,
- types of modulation of optical signals,
- types of noise occurring in optical elements and devices,
- optical guiding structures and fiber optic engineering aspects,
- dispersion and noises, occurring in optical guiding structures, etc.

In this chapter, we will try to introduce the reader to the most important aspects of electro-optical engineering, including the technical and technological aspects of optical elements and component fabrication, their material description, applied aspects of optical links fabrication, basic aspects of optical radar (called LIDAR) operation, and so forth, since the fine details are out of the scope of this book. The main goal of this book is to introduce the reader to fundamental aspects of electro-optical engineering based on basic physical fundamental questions which future engineers, technicians and researchers will meet during the design and development of basic elements and devices for optical communication and LIDAR.

1.1 Spectrum of Optical Waves

An optical communication system, either wired (i.e., fiber optics) or wireless (i.e., atmospheric or LIDAR), transmits analog and digital information from one place to another, using high carrier frequencies in the range of 100 THz to 1000 THz in the visible and infrared (IR) region of the electromagnetic spectrum [1–16]. As for microwave systems, they operate at carrier frequencies that are five orders of magnitude smaller, from 1 GHz to 50 GHz.

As a narrow band of the whole electromagnetic spectrum, the light wavelength band spreads from the ultraviolet spectral band to the far infrared (IR) spectral band, passing through the visible band, to the middle- and far infrared bands, as illustrated in Fig. 1.1, since most fiber optic cables, optical detectors and sources operate in these spectral bands.

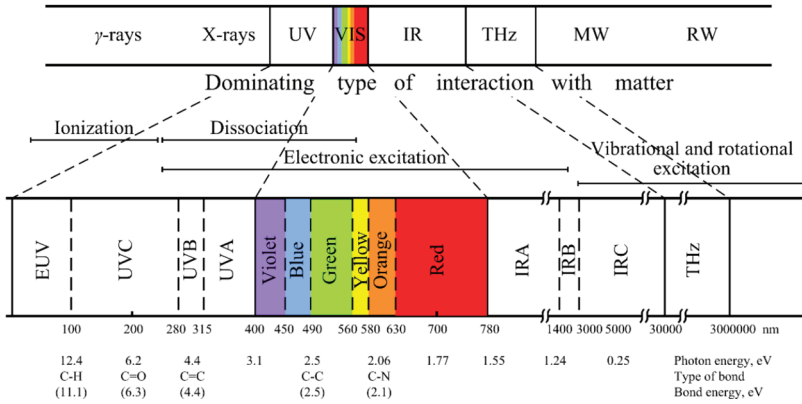


Figure 1.1. Electromagnetic spectrum and types of interaction with matter, indicated in the top panel by: UV – ultraviolet, VIS– visible light, IR – infrared, THz – terahertz-band wave, MW – microwaves, RW – radio waves (modified from [12-16]).

In electro-optics the large spectrum usually used – from the UV-band to the THz–band (see middle panel), which can be divided for practical applications into: UV, with UVC – far, UVB – middle, and UVA – near ultraviolet; VIS, divided from violet to red, as vividly shown by the corresponding color; and IR, with IRA – near, IRB – middle, and IRC – far infrared, as illustrated by the middle panel. The bottom panel presents along the horizontal axis the corresponding wavelengths for each narrow band [in

nanometers, 1nm = 10^{-9} m], the type of light and the band energy [in eV, 1 eV = $1.6 \bullet 10^{-19}$ J].

We notice that the relationship between wavelength (λ) and frequency (*f*) is: $\lambda = c / f$, where $c = 3 \cdot 10^8 \, m / s$ is the velocity of light in free space. As an example, a wavelength of light from the near IR band equals $\lambda = 1.5 \mu m$; it corresponds to a frequency of $f = 2 \cdot 10^{14} \, Hz = 2 \cdot 10^2 \, THz$ (with a period of oscillation equal to $T = 0.5 \cdot 10^{-14} \, s$).

The main goal of modern electro-optical engineering, photonics and optical electronics is to find the lowest energy and bandwidth losses of the corresponding materials during fabrication of the optical elements and devices [1–7, 11–16]. Thus, optical fiber systems operating in the 0.65–0.67 μm bandwidth with a plastic intrinsic surface have losses of 120–160 dB/km, whereas those operating in the 0.8–0.9 μm bandwidth have losses of 3–5 dB/km, and those operating in the 1.25–1.35 μm and 1.5–1.6 μm bandwidths, based on a glass surface, have losses of 0.5 to 0.25 dB/km, respectively [13]. We notice that the decibel [dB] is a measure called “path loss” denoted by *L* and defined as *L* = 10log*E*, where *E* is the energy of the optical wave in Joules [J].

Sufficiently wide frequency bands of light have allowed the increase of the bit rate (in bit/second, bps) – distance (in km) ratio during a period of about 150 years from $\sim 10^2$ *bps/km* to $\sim 10^{15}$ *bps*/*km* (summarized from [7, 12, 14–16]).

1.2 Fiber Optic Links

Below, we will give a definition of the optical link, both for a fiber optical link, as a "wire" communication link, and for an atmospheric link, a "wireless" communication link. The atmospheric links are outside of the scope of this book because they are fully presented in [15, 16].

As for wired optical communication links via fiber optics, they can be considered as a finishing optical communication mono-network, consisting of one fiber optic link, as shown in Fig. 1.2 rearranged from [16]. The message passing such a link is assumed to be available in electronic form, usually as a current. The transmitter is a light source that is modulated so that the optical beam carries the message.

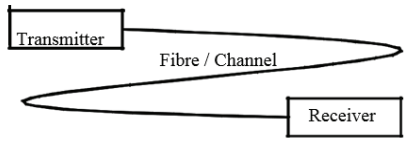


Figure 1.2. Scheme of optical communication link connected by fiber optics.

As an example, for a digital signal, the light beam is electronically turned on (for binary ones) and off (for binary zeros). Here, the optical beam is the carrier of the digital message. As a source fiber optic links usually take the light emitting diode and the laser diode. Several characteristics of the light source determine the behavior of the propagating optical waves [1–6]. The corresponding modulated light beam (i.e., the message with the carrier) is coupled into the transmission fiber.

1.3. Main Elements and Devices in Electro-Optics

The input to each optical channel is the optical signal from the optical transmitter, which emits optical signals, and the output of the channel is the input to the receiver, which detects optical signals. The receiver amplifies these optical signals, converts them to an electronic signal, and extracts the information. At the receiver, the signals are collected by a photodetector, which converts the information back into electrical form.

The photodetectors do not affect the propagation properties of the optical wave but certainly must be compatible with the rest of the optical system (Chapter 5). The transmitter includes a modulator, a driver, a light source, and optics (Fig. 1.3). The modulator converts the information bits to an analog signal that represents a symbol stream. The driver provides the required current to the light source based on the analog signal from the output of the modulator. The light sources are a light emitting diode (LED) and a pure laser, which is a coherent source and the subject of Chapter 5.

The source converts the electronic signal to an optical signal [6, 12]. The optics focuses and directs the light from the output of the source in the direction of the receiver.

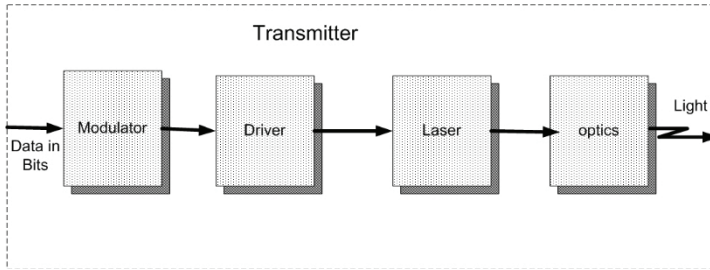


Fig. 1.3. The light source (transmitter) scheme.

The receiver includes optics, a filter, a polarizer, a detector, a trans-impedance amplifier, a clock recovery unit, and a decision device (see Fig. 1.4). The optics concentrate the received signal power onto the filter. Only light at the required wavelength propagates through the filter to the polarizer. The polarizer only enables light at the required polarization to propagate through to the detector. The detector, in most cases, is a semiconductor device such as a positive-negative (PN) or positive-intrinsic-negative (PiN) photodiode, which converts the optical signal to an electronic signal (see Chapter 5).

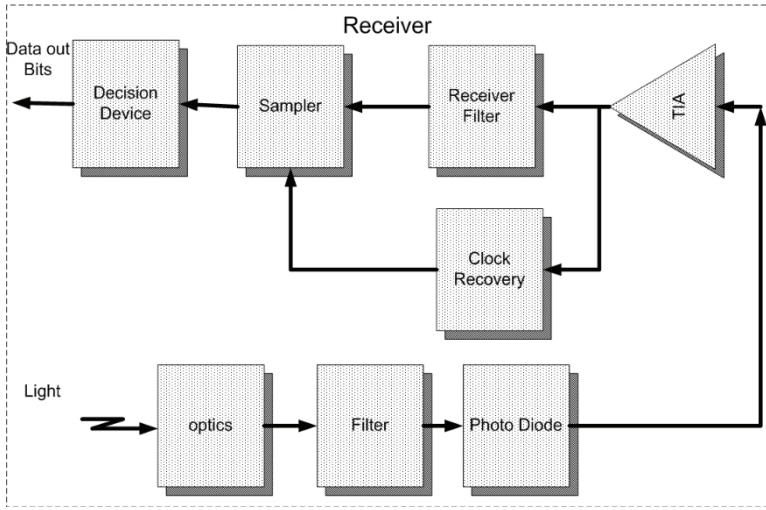


Figure 1.4. The light detector (receiver) scheme.

The amplifier increases the amplitude of the electronic signal from the detector. The clock recovery unit provides a synchronization signal to the decision device based on the signal from the output of the trans-impedance amplifier. The decision device estimates the received information based on the electronic signal from the trans-impedance amplifier and synchronization signal.

1.4. Noise in Optical Emitters and Detectors

In wired fiber optic and wireless (atmospheric) links, when the data stream is guided through them, they can be affected by noise occurring in each element of the optical emitters and detectors; the corresponding types of noise are discussed in Chapter 6.

1.5. Presentation of Signals in Electro-Optics

In electro-optics, the information carried by the optical signal can be presented both in analog and digital form, as shown in Figure 1.5. The analog form is a harmonically presented form of the signal in the time and frequency domains

$$s(t) = a(t)\exp\{j[\phi(\mathrm{t}) + 2\pi f t]\} \tag{1.1}$$

via its amplitude $a(t)$, phase $\phi(\mathrm{t})$ and frequency f (see Chapter 8).

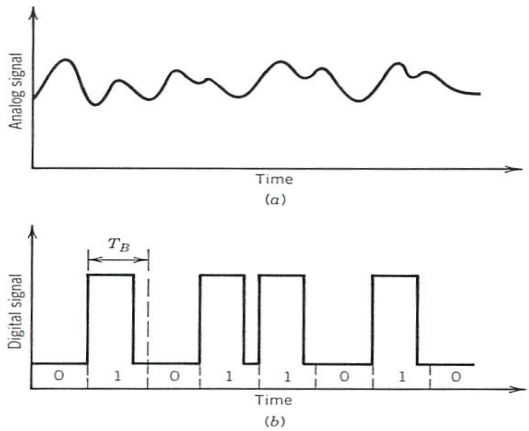


Figure 1.5. Presentation of information in optical communication links in the form of a) analog signal and b) digital (e.g., pulse) signal.

So, the upper set of blocks shown in Fig. 1.4 operate with a set of digital signals that were obtained by converting an analog signal, presented in harmonic form [see Eq. (1.1) above], into a digital signal via quantization of the analog optical signal and presentation of the flux of optical quants as a discrete sequence of codes, 0 and/or 1, as shown in Figure 1.6.

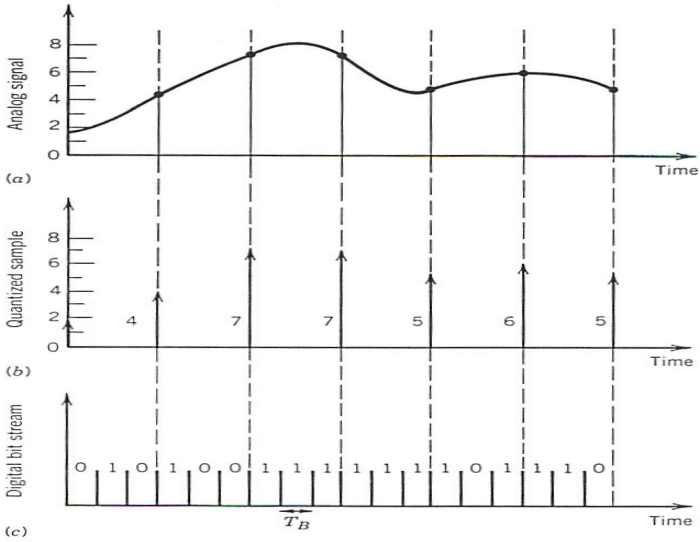


Figure 1.6. a) Sampling, b) quantization, and c) coding.

1.6. Types of Modulation of Optical Signals

As for types of modulation, it also depends on the type of optical signal – analog (or continuous wave (CW)) or digital. To understand the further presented material, we refer the reader to a simple explanation of CW modulation usually used in optical devices to obey different kinds of signals.

For such kinds of optical signals, we deal with three kinds of modulation/demodulation: *amplitude* or *intensity*, *phase* and *frequency*. Thus, each CW signal can be presented in exponential form (1.1). As follows from (1.1), there are three possible kinds of modulation/demodulation of CW optical signals:

a) via changes of carrier optical signal amplitude or intensity by the influence of the modulating signal (usually called the *message*),

b) via changes in phase of the modulated carrier optical signal by

mixing it with modulating signal, and

c) via changes in frequency of the modulated carrier optical signal by mixing it with modulating signal frequency.

All these aspects will be discussed in Chapter 9, where some examples of practical application will also be presented.

As for digital modulation/demodulation [10, 16], this also can be divided into three types according to changes of amplitude (called amplitude shift keying, ASK), phase (phase shift keying, PSK), and frequency (frequency shift keying, FSK), as shown schematically in Fig. 1.7.

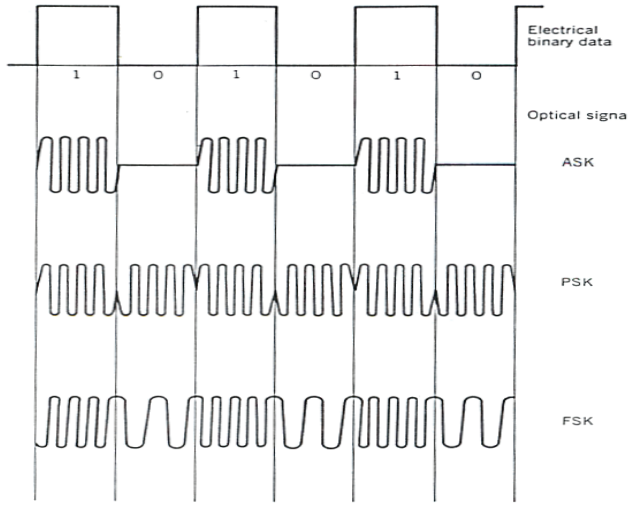


Figure 1.7. Three types of modulation of digital (pulse) optical signal: amplitude (ASK), phase (PSK), and frequency (FSK).

The most common type of digital coding and encoding is On-Off Keying (OOK) [10]. This can be presented in two basic formats (see Fig. 1.8): a) Return-to-Zero (RZ), and b) Non-Return-to-Zero (NRZ) [10, 16].

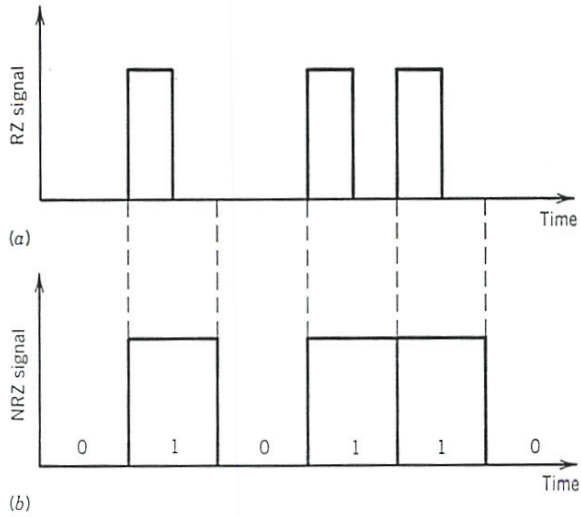


Figure 1.8. a) RZ and b) NRZ format of digital optical signals (according to [10, 16]).

Now we will briefly describe the basic elements of the optical communication channel, including the transmitter, as a source of light and the receiver, as the detector of light. On this topic, the basic electronic elements usually used in electro-optics and photonics, in optical sources and detector fabrication, will be mentioned in Chapters 5–7.

1.7. Wired (Fiber Optic) Fundamentals

Propagation of light in wired links is fully characterized by propagation in guiding structures, as in 2-D slab or 3-D fiber optical cable. In this case, the electromagnetic approach, based on the wave nature of light, illuminates all the peculiarities of wave mode propagation within such kinds of guiding structures, the interaction of these modes, losses, time spreading processes, and so on. All these aspects will be discussed in Chapter 10.

Bibliography

[1] Jenkins, F. A., and H. E. White. 1953. *Fundamentals of Optics*. New York: McGraw-Hill.
[2] Born, M., and E. Wolf. 1964. *Principles in Optics*. New York: Pergamon Press.
[3] Fain, V. N., Ya. N. Hanin. 1965. *Quantum Radiophysics*. Moscow: Sov. Radio, (in Russian).
[4] Akhamov, S. A., and R. V. Khohlov. 1965. *Problems of Nonlinear Optics*. Moscow: Fizmatgiz, (in Russian).
[5] Lipson, S. G., and H. Lipson. 1969. *Optical Physics*. Cambridge: University Press.
[6] Akhamov, S. A., R. V. Khohlov, and A. P. Sukhorukov. 1972. *Laser Handbook*. North Holland: Elsevier.
[7] Marcuse, O. 1972. *Light Transmission Optics*. New York: Van Nostrand-Reinhold Publisher.
[8] Kapany, N. S., and J. J. Burke. 1972. *Optical Waveguides*. Chapter 3, New York: Academic Press.
[9] Fowles, G. R. 1975. *Introduction in Modern Optics*. New York: Holt, Rinehart, and Winston Publishers.
[10] Krouk, E., and S. Semenov, eds. 2011. *Modulation and Coding Techniques in Wireless Communications*. NJ: John Wiley & Sons.
[11] Hecht, E.1987. *Optics*. MA: Addison-Wesley, Reading.
[12] Dakin, J., and B. Culshaw, eds. 1988. *Optical Fiber Sensors: Principles and Components*. Artech House, Boston-London.
[13] Culshaw, B., and J. Dakin, eds. 1989. *Optical Fiber Sensors: Systems and Applications*, Vol. 2. Norwood Ma.
[14] Bansal, R. ed. 2006. Handbook: *Engineering Electromagnetics Applications*. New York: Taylor and Frances.
[15] Blaunstein, N., Sh. Arnon, A. Zilberman, and N. Kopeika. 2010. *Applied Aspects of Optical Communication and LIDAR*. New York: CRC Press, Taylor & Francis Group.
[16] Blaunstein, N., S. Engelberg, E. Krouk, and M. Sergeev. 2019. *Fiber Optic and Atmospheric Optical Communication*. Hoboken NJ: Wiley.

CHAPTER 2

ELECTROMAGNETIC NATURE OF LIGHT

2.1. Optical Wave Electromagnetic Fundamentals

The theoretical analysis of optical wave propagation, as a part of the whole electromagnetic spectrum [1–6] (see Paragraph 1.1, Chapter 1), is based on Maxwell's equations [10–16]. In vector notation and in the SI-units system, the optical wave electromagnetic features can be presented in the uniform macroscopic form [1–6]:

$$\nabla \times \boldsymbol{E}(\mathbf{r},t) = -\frac{\partial}{\partial t}\boldsymbol{B}(\mathbf{r},t), \tag{2.1a}$$

$$\nabla \times \boldsymbol{H}(\mathbf{r},t) = \frac{\partial}{\partial t}\boldsymbol{D}(\mathbf{r},t) + \boldsymbol{j}(\mathbf{r},t), \tag{2.1b}$$

$$\nabla \cdot \boldsymbol{B}(\mathbf{r},t) = 0, \tag{2.1c}$$

$$\nabla \cdot \boldsymbol{D}(\mathbf{r},t) = \rho(\boldsymbol{r},t) \tag{2.1d}$$

Here, $\boldsymbol{E}(\boldsymbol{r},t)$ is the electric field strength vector, in volts per meter (V/m); $\boldsymbol{H}(\boldsymbol{r},t)$ is the magnetic field strength vector, in amperes per meter (A/m); $\boldsymbol{D}(\boldsymbol{r},t)$ is the electric flux induced in the medium by the electric field, in coulombs/m^3 (this is why, in the literature, sometimes it is called an "induction" of an electric field); $\boldsymbol{B}(\boldsymbol{r},t)$ is the magnetic flux induced by the magnetic field, in webers/m^2 (it is also called an "induction" of a magnetic field); $\boldsymbol{j}(\boldsymbol{r},t)$ is the vector of electric current density, in amperes/m^2; $\rho(\boldsymbol{r},t)$ is the charge density in coulombs/m^2. The curl operator $\nabla \times$ is a measure of field rotation, and the divergence operator $\nabla \cdot$ is a measure of the total flux radiated from the desired point.

It should be noted that for a time-varying EM-wave field, equations (2.1c) and (2.1d) can be derived from (2.1a) and (2.1b), respectively. In fact, taking the divergence of (2.1a) (by use of the divergence operator $\nabla \cdot$) one can immediately obtain (2.1c). Similarly, taking the divergence of (2.1b) and using the well-known continuity equation [1–3, 10–13]

$$\nabla \cdot \boldsymbol{j}(\boldsymbol{r},\mathrm{t}) + \frac{\partial \rho(\boldsymbol{r},\mathrm{t})}{\partial t} = 0 \qquad (2.2)$$

one can arrive at (2.1d). Hence, only two equations (2.1a) and (2.1b) are independent.

Equation (2.1a) is the well-known Faraday law and indicates that a time-varying magnetic flux generates an electric field with rotation; (2.1b) without the term $\frac{\partial D}{\partial t}$ (displacement current term [10–13]) limits to the well-known Ampere law and indicates that a current or a time-varying electric flux (displacement current [10–13]) generates a magnetic field with rotation.

Because one now has only two independent equations (2.1a) and (2.1b), which describe the four unknown vectors $\boldsymbol{E}, \boldsymbol{D}, \boldsymbol{H}, \boldsymbol{B}$, three more equations relating to these vectors are needed. To do this, we introduce relations between $\boldsymbol{E}$ and $\boldsymbol{D}$, $\boldsymbol{H}$ and $\boldsymbol{B}$, $\boldsymbol{j}$ and $\boldsymbol{E}$, which are known in electrodynamics. In fact, for isotropic media, which are usually considered in problems of land-atmospheric optical propagation, the electric and magnetic fluxes are related to the electric and magnetic fields, and the electric current is related to the electric field, via the constitutive relations [10–13]:

$$\boldsymbol{B} = \mu(\boldsymbol{r})\boldsymbol{H} \qquad (2.3)$$

$$\boldsymbol{D} = \varepsilon(\boldsymbol{r})\boldsymbol{E} \qquad (2.4)$$

$$\boldsymbol{j} = \sigma(\boldsymbol{r})\boldsymbol{E} \qquad (2.5)$$

It is important to emphasize that relations (2.3) to (2.5) are valid only for propagation processes in regular isotropic media, which are characterized by the three scalar functions of any point $\mathbf{r}$ in the medium:

- permittivity $\varepsilon(\boldsymbol{r})$,
- permeability $\mu(\boldsymbol{r})$, and
- conductivity $\sigma(\boldsymbol{r})$.

In relations (2.3) to (2.5), it was assumed that the medium is inhomogeneous. In a homogeneous medium, the functions $\varepsilon(\boldsymbol{r})$, $\mu(\boldsymbol{r})$, and $\sigma(\boldsymbol{r})$ transform to simple scalar values ε, μ and σ.

In free space, these functions simply are constants, i.e., $\varepsilon = \varepsilon_0 = 8.854 \cdot 10^{-12} = \frac{1}{36\pi}10^{-9}$ Farad/meter (F/m), while $\mu = \mu_0 = 4\pi \cdot 10^{-7}$ Henry/meter (H/m). The constant $c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}}$ is the velocity of light, which

has been measured very accurately and is $3 \cdot 10^8$ m/s.

The system (2.1) can be further simplified if we assume that the fields are time harmonic. If the field time dependence is not harmonic then, using the fact that equations (2.1) are linear, we may treat these fields as sums of harmonic components and consider each component separately. In this case, the time harmonic field is a complex vector and can be expressed via its real part as [10–13, 16]

$$\boldsymbol{A}(\boldsymbol{r},t) = Re[\boldsymbol{A}(\boldsymbol{r})e^{-i\omega t}], \tag{2.6}$$

where $i = \sqrt{-1}$, ω is the angular frequency in radians per second, $\omega = 2\pi f$, f is the radiated frequency (in Hz = s^{-1}), and $\boldsymbol{A}(\boldsymbol{r},t)$ is the complex vector ($\boldsymbol{E},\boldsymbol{D},\boldsymbol{H},\boldsymbol{B}$, or $\boldsymbol{j}$). The time dependence $\sim e^{-i\omega t}$ is commonly used in the literature regarding electrodynamics and wave propagation. If $\sim e^{i\omega t}$ is used, then one must substitute $-i$ for i and i for $-i$, in all equivalent formulations of Maxwell's equations. In (2.6) $e^{-i\omega t}$ presents the harmonic time dependence of any complex vector $\boldsymbol{A}(\boldsymbol{r},t)$, which satisfies the relationship:

$$\frac{\partial}{\partial t}\mathbf{A}(\mathbf{r},t) = Re[-i\omega\boldsymbol{A}(\boldsymbol{r})e^{-i\omega t}] \tag{2.7}$$

Using this transformation, one can easily obtain from the system (2.1):

$$\nabla \times \boldsymbol{E}(\boldsymbol{r}) = i\omega\boldsymbol{B}(\boldsymbol{r}) \tag{2.8a}$$

$$\nabla \times \boldsymbol{H}(\boldsymbol{r}) = -i\omega\boldsymbol{D}(\boldsymbol{r}) + \boldsymbol{j}(\boldsymbol{r}) \tag{2.8b}$$

$$\nabla \cdot \boldsymbol{B}(\boldsymbol{r}) = 0 \tag{2.8c}$$

$$\nabla \cdot \boldsymbol{D}(\boldsymbol{r}) = \rho(\boldsymbol{r}) \tag{2.8d}$$

It can be observed that system (2.8) was obtained from system (2.1) by replacing $\frac{\partial}{\partial t}$ with $-i\omega$. Alternatively, the same transformation can be obtained by the use of the Fourier transform of system (2.1) with respect to time [1, 2, 10–16]. In (2.8a-d) all vectors and functions are actually the Fourier transforms with respect to the *time domain*, and the fields $\boldsymbol{E},\boldsymbol{D},\boldsymbol{H}$, and $\boldsymbol{B}$ are functions of frequency as well, we call them *phasors* of time domain vector solutions. They are also known as the *frequency domain solutions* of the EM field according to system (2.8). Conversely, the solutions of system (2.1) are the *time domain solutions* of the EM field. It is

more convenient to work with system (2.8) instead of system (2.1) because of the absence of the time dependence and time derivatives in it.

2.2. Propagation of Optical Waves in Free Space

The mathematical tool presented above shows that light can be fully described mathematically by Maxwell's unified theory [1–3, 10–16], according to which optical waves have the same nature as electromagnetic waves, being their own part in frequency (or wavelength) domain (see Fig. 1.1, Chapter 1). So, we may start with a physical explanation of electromagnetic waves based on Maxwell's unified theory [1, 2, 10–13], which postulates that an electromagnetic field could be represented as a wave. The coupled wave components, electric and magnetic fields, are depicted in Fig. 2.1, from which it follows that the electromagnetic (EM) wave travels in a direction perpendicular to both EM field components. In Fig. 2.1 this direction is denoted as the z-axis in the Cartesian coordinate system by the wave vector **k**. In their orthogonal space-planes, the magnetic and electric oscillatory components repeat their waveform after a distance of one wavelength along the *y*-axis and *x*-axis, respectively (see Fig. 2.1).

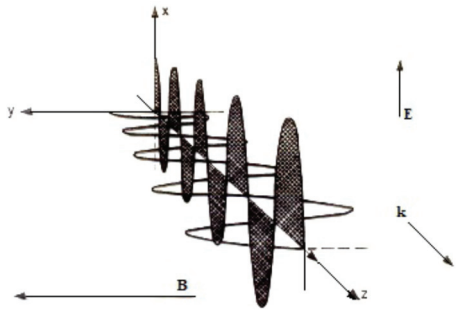


Fig. 2.1. Optical wave as an electromagnetic wave with its electrical and magnetic components, wavefront, and direction of propagation presentation.

Both components of the EM wave are in phase in the time domain but not in the space domain [1, 2, 10–13]. Moreover, the magnetic component value of the EM field is closely related to the electric component value, from which one can obtain the radiated power of the EM wave propagating along the z-axis (see Fig. 2.1).

At the same time, using Huygen's principle, well-known in electrodynamics [10–13], one can show that the optical wave is the electromagnetic wave propagating only straightforward from the source, as rays with the minimum loss of energy and with minimum time for propagation (according to Fermat's Principle postulated in classical optics [2, 7, 15]) in free space, as an unbounded homogeneous medium without sources, obstacles and discontinuities.

Thus, if we present Huygen's concept, as it is shown in Fig. 2.2, the ray from each point propagates in all forward directions and forms many elementary spherical wavefronts, which Huygens called *wavelets*.

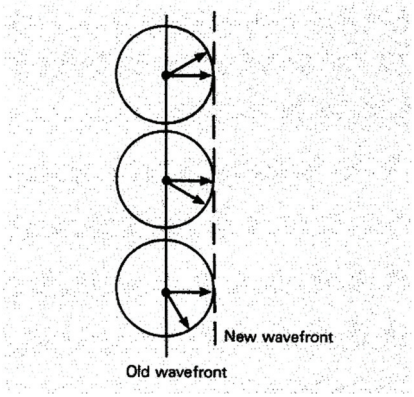


Figure 2.2. Huygens principle for proof of straight propagation of waves as rays.

The envelope of these wavelets forms the new wavefront. In other words, each point on the wavefront acts as a source of secondary elementary spherical waves, described by Green's function (see Refs. [10–13]). These elementary waves combined together produce a new wavefront in the direction of wave propagation in a straight manner (see Fig. 2.2). As we will show below, each wavefront can be represented by the plane, which is normal to the wave vector **k** (e.g., wave energy transfer). Moreover, propagating forward along straight lines normal to their wavefront, any wave propagates as light rays in optics, spending minimum energy for passing from the source to detector, that is, the maximum energy of the ray is observed in a straight direction normal to the wavefront (as is seen from Fig. 2.2). The first person who mathematically showed this principle was Kirchhoff, based on a general Maxwell's unified theory.

Let us now assess all propagation phenomena theoretically using Maxwell's unified theory. Mathematically, optical wave propagation phenomena can be described by the use of both the scalar and vector wave

equation presentations. Because most problems of optical wave propagation in wireless and wired communication links are considered in unbounded, homogeneous, source-free isotropic media, we can present the environmental and material functions simply, as numbers, $\varepsilon(\boldsymbol{r}) \equiv \varepsilon$, $\mu(\boldsymbol{r}) \equiv \mu, \sigma(\boldsymbol{r}) \equiv \sigma$, and finally obtain from general wave equations:

$$\nabla \times \nabla \times \boldsymbol{E}(\boldsymbol{r}) - \omega^2 \varepsilon \mu \boldsymbol{E}(\boldsymbol{r}) = 0$$

$$\nabla \times \nabla \times \boldsymbol{H}(\boldsymbol{r}) - \omega^2 \varepsilon \mu \boldsymbol{H}(\boldsymbol{r}) = 0 \tag{2.9}$$

Because both equations are symmetric, one can use one of them, namely that for $\boldsymbol{E}$, and by introducing the vector relation $\nabla \times \nabla \times \boldsymbol{E} = \nabla(\nabla \cdot \boldsymbol{E}) - \nabla^2 \boldsymbol{E}$ and taking into account that $\nabla \cdot \boldsymbol{E} = 0$, finally obtain

$$\nabla^2 \boldsymbol{E}(\boldsymbol{r}) + k^2 \boldsymbol{E}(\boldsymbol{r}) = 0 \tag{2.10}$$

where $k^2 = \omega^2 \varepsilon \mu$.

In special cases of a homogeneous, source-free, isotropic medium, the three dimensional wave equation reduces to a set of scalar wave equations. This is because in Cartesian coordinates, $\boldsymbol{E}(\boldsymbol{r}) = E_x \boldsymbol{x}_0 + E_y \boldsymbol{y}_0 + E_z \boldsymbol{z}_0$, where $\boldsymbol{x}_0$, $\boldsymbol{y}_0$, $\boldsymbol{z}_0$ are unit vectors in the directions of the x, y, z coordinates, respectively. Hence, the equation (2.10) consists of three scalar equations such as

$$\nabla^2 \Psi(\boldsymbol{r}) + k^2 \Psi(\boldsymbol{r}) = 0 \tag{2.11}$$

where $\Psi(\boldsymbol{r})$ can be either E_x, E_y, or E_z. This equation fully describes the propagation of optical waves in free space.

2.3 Propagation of Optical Waves Through the Boundary of Two Media

2.3.1 Boundary conditions

The simplest case of wave propagation over the intersection between two media is that where the intersection surface can be assumed to be flat and perfectly conductive.

If so, for a perfectly conductive flat surface the total electric field vector is equal to zero, i.e., $\boldsymbol{E} = 0$ [1–3, 10–13, 16]. In this case, the tangential component of the electric field vanishes at the perfectly conductive flat surface, that is,

$$E_\tau = 0 \tag{2.12}$$

Consequently, as follows from Maxwell's equation $\nabla \times \boldsymbol{E}(\boldsymbol{r}) = i\omega \boldsymbol{B}(\boldsymbol{r})$, (see above for the case of $\mu = 1$ and $\boldsymbol{B} \equiv \boldsymbol{H}$), at such a flat, perfectly conductive surface, the normal component of the magnetic field also vanishes, i.e.,

$$H_n = 0 \tag{2.13}$$

As also follows from system (2.1) of Maxwell's equations, the tangential component of the magnetic field does not vanish because of its compensation by the surface electric current. At the same time, the normal component of the electric field is also compensated by pulsing electrical charge at the intersection surface. Hence by introducing the Cartesian coordinate system, one can present the boundary conditions at the flat perfectly conductive intersection surface as follows:

$$E_x(x, y, z = 0) = E_y(x, y, z = 0) = H_z(x, y, z = 0) = 0 \tag{2.14}$$

2.3.2 Main formulations of reflection and refraction coefficients

As was shown above, the influence of a flat material surface on optical wave propagation leads to phenomena such as reflection. Because all kinds of waves can be represented by means of the concept of the plane waves [1–3, 10–13], let us obtain the main reflection and refraction formulas for a plane wave that incidents on a plane surface between two media, as shown in Fig. 2.3. The media have different dielectric properties, which are described above and below the boundary plane z = 0 by the permittivity and permeability ε_1, μ_1 and ε_2, μ_2, respectively, for each medium.

Without reducing the general problem, let us consider an optical wave with wave vector $\boldsymbol{k}_i$ and frequency $\omega = 2\pi f$ incident from a medium described by parameter n_1. The reflected and refracted waves are described by wave vectors $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$, respectively. Vector $\boldsymbol{n}$ is a unit normal vector directed from a medium with the refractive index n_1 into a medium with refractive index n_2, where $\varepsilon_1 = n_1^2$ and $\varepsilon_2 = n_2^2$. Here, we should notice that in optics, usually the designers of optical systems deal with non-magnetized materials, putting the normalized dimensionless permeability of the two media to equal the unit, that is, $\mu_1 = \tilde{\mu}_1/\mu_0 = 1$ and $\mu_2 = \tilde{\mu}_2/\mu_0 = 1$, as well as using the normalized dimensionless permittivity for each medium, $\varepsilon_1 = \tilde{\varepsilon}_1/\varepsilon_0$, and $\varepsilon_2 = \tilde{\varepsilon}_2/\varepsilon_0$, accounting for the above presented

relations: $\varepsilon_1 = n_1^2$ and $\varepsilon_2 = n_2^2$. We notice that these parameters for free space were defined and introduced above.

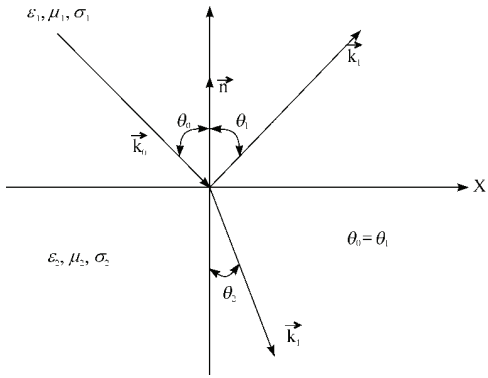


Fig. 2.3. Reflection and refraction of optical wave at the boundary of two media.

According to the relations between electrical and magnetic components, which follow from Maxwell's equations (see system (2.1)), one can easily obtain the expressions for the coefficients of reflection and refraction (see, for example, [1–3]). A physical meaning of the reflection coefficient is the follows:

It defines the ratio of the reflected electric field component of the optical wave to its incident electric field component.

The same physical meaning is of the refractive coefficient:

It defines the ratio of the refractive electric field component to the incident electric field component of the optical wave.

Before presenting these formulas, let us show two important laws usually used in classical optics. As follows from Maxwell's laws, from the boundary conditions and geometry presented in Fig. 2.3, the values of the wave vectors are related by the following expressions [16]:

$$|\mathbf{k}_0| = |\mathbf{k}_1| \equiv k = \frac{\omega}{c} n_1, \quad |\mathbf{k}_2| \equiv k_2 = \frac{\omega}{c} n_2 \tag{2.15}$$

From the boundary conditions described earlier by (2.12) to (2.14), one can easily obtain the condition of the equality of phase for each wave at the plane $z = 0$:

$$(\boldsymbol{k}_0 \cdot \boldsymbol{x})_{z=0} = (\boldsymbol{k}_1 \cdot \boldsymbol{x})_{z=0} = (\boldsymbol{k}_2 \cdot \boldsymbol{x})_{z=0}, \tag{2.16}$$

which is independent of the nature of the boundary condition. Equation (2.16) describes the condition that all three wave vectors must lie in the same plane. From this equation it also follows that

$$k_0 \sin\theta_0 = k_1 \sin\theta_1 = k_2 \sin\theta_2 \tag{2.17}$$

which is the analog of the *second Snell's law*:

$$n_1 \sin\theta_0 = n_2 \sin\theta_2 \tag{2.18}$$

Moreover, because $|\boldsymbol{k}_0| = |\boldsymbol{k}_1|$, we find $\theta_0 = \theta_1$, i.e., the angle of incidence equals the angle of reflection. This is the *first Snell's law*.

In the literature which describes wave propagation aspects, the optical waves are called waves with *vertical* and *horizontal* polarization, depending on the orientation of the electric field component regarding the plane of propagation, perpendicular or parallel, respectively.

Without entrance into straight retinue computations, following classical electrodynamics, we will obtain the expressions for the complex coefficients of reflection (R) and refraction (T) for waves with vertical (denoted by index V) and horizontal (denoted by index H) polarization, respectively. For this purpose, we now introduce the relative dielectric parameter $\varepsilon_r = \varepsilon_2/\varepsilon_1$, that is, present it via dimensionless dielectric parameters of two media, $\varepsilon_1 = \tilde{\varepsilon}_1/\varepsilon_0$, and $\varepsilon_2 = \tilde{\varepsilon}_2/\varepsilon_0$ introduced above. Moreover, we will also account for the above introduced relations between the dimensionless dielectric permittivity and the refractive index for each medium: $\varepsilon_1 = n_1^2$ and $\varepsilon_2 = n_2^2$, and will use the 2nd Snell's law, $n_1 \sin\theta_0 = n_2 \sin\theta_2$. Finally, we will get:

For *vertical* polarization:

$$R_V = |R_V| e^{j\phi_V} = \frac{\cos\theta_0 - \sqrt{\varepsilon_r}\cos\theta_2}{\cos\theta_0 + \sqrt{\varepsilon_r}\cos\theta_2} \tag{2.19a}$$

$$T_V = |T_V| e^{j\phi_V^{\copyright}} = \frac{2\cos\theta_0}{\cos\theta_0 + \sqrt{\varepsilon_r}\cos\theta_2} \tag{2.19b}$$

For *horizontal* polarization:

$$R_H = |R_H| e^{j\phi_H} = \frac{\cos\theta_2 - \sqrt{\varepsilon_r}\cos\theta_0}{\cos\theta_2 + \sqrt{\varepsilon_r}\cos\theta_0} \tag{2.20a}$$

$$T_H = |T_H| e^{j\phi_H^{\copyright}} = \frac{2\cos\theta_0}{\cos\theta_2 + \sqrt{\varepsilon_r}\cos\theta_0} \tag{2.20b}$$

In the case of vertical polarization there is a special angle of incidence, called the *Brewster angle*, for which there is *no reflected wave*, only a *refractive wave*. For simplicity, we will assume that the condition $\mu_1 = \mu_2$ is valid. Then from (2.18) and (2.19a), it follows that the reflected wave limits to zero when the angle of incidence is equal to Brewster's angle

$$\theta_0 \equiv \theta_{Br} = tan^{-1}\left(\frac{n_2}{n_1}\right) \tag{2.21}$$

We should notice that the Brewster angle is only valid for the wave with the vertical polarization, which describes a situation with the absence of the reflected wave and the existence of the refractive wave only (the so-called effect of *total refraction*). For the case of $\mu_1 = \mu_2 = 1$, the reflected wave $\boldsymbol{E}_1$ limits to zero when the incident wave is under the Brewster angle and can be described by formula (2.21).

Another interesting phenomenon that follows from the presented formulas is called *total ray reflection*. It takes place when the condition $n_2 >> n_1$ is valid. In this case, from Snell's law (2.21) it follows that, if $n_2 >> n_1$, then $\theta_1 >> \theta_i$. Consequently, when $\theta_i >> \theta_c$ the reflection angle $\theta_1 = \frac{\pi}{2}$, where

$$\theta_c = sin^{-1}\left(\frac{n_2}{n_1}\right) \tag{2.22}$$

For waves incident at the surface under the critical angle $\theta_i \equiv \theta_c$ there is *no refracted wave* within the second medium; the refracted wave is propagated along the boundary between the first and second media and there is no energy flow across the boundary of these two media.

Therefore, this phenomenon is called in the literature *total internal reflection* (TIR), and the smallest incident angle θ_i for which we get TIR, is called the critical angle $\theta_i \equiv \theta_c$ defined by expression (2.22). The refraction of the wave in the second media is fully absent.

2.4. Total Intrinsic Reflection in Optics

We can rewrite Snell's law, presented above for $\theta_i = \theta_1$, as [1–6] (see also the geometry of the problem shown in Fig. 2.3):

$$n_1 \sin\theta_r = n_2 \sin\theta_t \tag{2.23}$$

or

$$sin\,\theta_i \equiv sin\,\theta_r = \frac{n_2}{n_1} sin\,\theta_t \tag{2.24}$$

If the second medium is less optically dense than the first medium and the incident ray has amplitude $|\boldsymbol{E}_i|$, that is, $n_1 > n_2$, from (2.24) it follows that

$$sin\,\theta_i > \frac{n_2}{n_1} \tag{2.25a}$$

or

$$\frac{n_1}{n_2} sin\,\theta_i > 1. \tag{2.25b}$$

The value of the incident angle θ_i for which (2.25) becomes true is known as a *critical angle*, which was introduced above. We now define its meaning by use of the ray concept [1–3]. If a critical angle is determined by (2.22), which we will rewrite in another way:

$$sin\,\theta_c = \frac{n_2}{n_1} \tag{2.26}$$

then for all values of incident angles $\theta_i > \theta_c$ the light is totally reflected at the boundary of the two media. This phenomenon is called in ray theory the *total internal reflection* (TIR) of rays, the effect which is very important in light propagation in fiber optics.

We can also introduce another main parameter usually used in optic communications. The *effective index of refraction* is defined as: $n_{eff} \equiv n_1\,sin\,\theta_i$. When the incident ray angle $\theta_i = 90^0$, $n_{eff} \equiv n_1$, and when $\theta_i = \theta_c$, $n_{eff} \equiv n_2$.

The guiding effect, which occurs in fiber optic structures (see Chapter 8), is based on the TIR phenomenon:

All energy transport occurs along the boundary of two media after TIR, without any penetration of light energy inside the intersection.

Moreover, we should notice that the totally internal reflected (TIR) wave undergoes a phase change, which depends on both the angle of incidence and the field polarization [15, 16].

Let us now explain the *total internal reflection* from another point of view based on discussions introduced in [16]. When the total internal reflection occurs, we should assume that there would be no electric field in the second medium. This is not the case, however. The boundary conditions presented above require that the electric field be continuous at the boundary,

that is, at the boundary the field in region 1 and region 2 must be equal. The exact solution shows that due to total internal reflection we have in region 1 standing waves caused by the interference of incident and fully reflected waves, whereas in region 2 a finite electric field decays exponentially away from the boundary and carries no power into the second medium. This wave is called an *evanescent field* (see Fig. 2.4). As shown on the left side of Fig. 2.4, the standing wave occurs as a result of interaction between two optical waves, the incident wave and the wave reflected from the interface of two media. We should notice that this picture is correct in situations when the refractive index of the first transparent medium is larger than that of the second transparent medium, that is, $n_1 > n_2$, and when total reflection from the intersection occurs, that is, for an incident angle exceeding the critical one, θ_c, defined by Eq. (2.26).

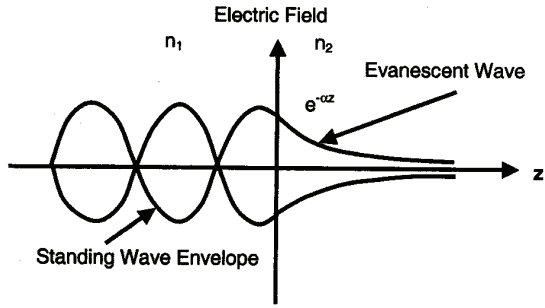


Fig. 2.4. Electric component of the optical wave at the boundary of two media forming standing wave due to reflection, and wave decay $\sim e^{-\alpha z}$ due to refraction.

This field attenuates away from the boundary as

$$E \propto exp\{-\alpha z\} \tag{2.27}$$

where the attenuation factor equals

$$\alpha = \frac{2\pi}{\lambda}\sqrt{n_1^2\, sin^2\, \theta_i - n_2^2} \tag{2.28}$$

It can be seen from (2.28), at the critical angle $\theta_i = \theta_c$ $\alpha \to 0$, attenuation increases as the incident angle increases beyond the critical angle defined by (2.26). Because α is so small near the critical angle, the evanescent fields penetrate deeply beyond the boundary but do so less and less as the angle increases.

However, the behavior of the main formulas (2.19) and (2.20) depends on boundary conditions. Thus, if the fields are continuous across the boundary, as required by Maxwell's equations, there must be a field disturbance of some kind in the second media (see Fig. 2.4). To investigate this disturbance, we can use Fresnel's formulas. We first of all rewrite, following [15, 16], as $cos\,\theta_t = (1 - sin^2\,\theta_t)^{1/2}$. For $\theta_t > \theta_c$ we can present $sin\,\theta_t$ by the use of an additional function $sin\,\theta_t = cosh\,\gamma$, which can be more than one unit. If so, $cos\,\theta_t = j(cosh^2\,\gamma - 1)^{1/2} = \pm j\,sinh\,\gamma$. Hence, we can write the field component in the second medium to vary as (for nonmagnetic materials $\mu_1 = \mu_2 = \mu_0$)

$$exp\left\{j\omega\left(t - n_2\frac{x\,cosh\,\gamma - jz\,sinh\,\gamma}{c}\right)\right\} \tag{2.29a}$$

or

$$exp\left(-\frac{\omega n_2 z\,sinh\,\gamma}{c}\right)exp\left\{j\omega\left(t - n_2\frac{x\,cosh\,\gamma}{c}\right)\right\}. \tag{2.29b}$$

The last formula represents a ray traveling in the z-direction in the second medium (that is, parallel to the boundary) with the amplitude decreasing exponentially in the z-direction (at right angles to the boundary). The rate of the amplitude decrease versus z can be written

$$exp\left(-\frac{2\pi z\,sinh\,\gamma}{\lambda_2}\right)$$

where λ_2 is the wavelength of the light in the second medium. The wave with the exponential decay is usually called an *evanescent wave* in the literature [15, 16]. As seen from Fig. 2.4, rearranged from [16], the wave attenuates significantly ($\sim e^{-1}$) over critical distances d_c of about λ_2. Another expression of the evanescent wave decay region, d_c, can be obtained, following [15, 16], by introducing the incident angle of light at the boundary of two media θ and both refractive indexes of the media, n_1 and n_2:

$$d_c = \frac{\lambda_2}{2\pi\left(n_1^2\,sin^2\,\theta - n_2^2\right)^{1/2}} \tag{2.30}$$

This critical depth of field exponential attenuation is similar to the characteristics of electromagnetic wave penetration into the material usually used in electrodynamics and electromagnetism and called the *skin layer* [11–13].

Even though the wave is propagating in the second medium, it transports no light energy in a direction normal to the boundary. All the light

is totally internally reflected (TIR) at the boundary.

2.5. Propagation of Optical Waves in Materials

As was shown above, each electrical field component (let us say the *x*-component) of the optical wave can be presented as a plane wave in any media in the following manner

$$\tilde{E}_x = Ae^{-\gamma z} + Be^{+\gamma z} \tag{2.31}$$

where *A* and *B* are constants that can be obtained from the corresponding boundary conditions; the propagation parameter is complex and can be written as

$$\gamma = \alpha + j\beta \tag{2.32}$$

Here, α describes the attenuation of the optical wave amplitude, that is, the wave energy losses, and β describes the phase velocity of the plane wave in the material media.

Now we can present the magnetic field phasor component in the same manner, as the electric field by use of [11–16]:

$$\tilde{H}_y = \frac{1}{\eta}(Ae^{-\gamma z} - Be^{+\gamma z}) \tag{2.33}$$

where η is the intrinsic impedance of the medium, which is also complex. Solutions (2.31) and (2.33) can be concretized by the use of the corresponding boundary conditions. But this is not a goal of our future analysis. We will show the reader how the properties of the material medium change propagation conditions within it. For this purpose, we analyze the propagation parameters γ [or α and β] and η associated with plane waves (2.31) and (2.33). After straightforward computations of the corresponding equations, following [11–13, 16], we can find for $\mu = 1$ that

$$\alpha = \frac{\omega\sqrt{2\varepsilon}}{2}\left[\sqrt{1 + \left(\frac{\sigma}{\omega\varepsilon}\right)^2} - 1\right]^{1/2} \tag{2.34a}$$

$$\beta = \frac{\omega\sqrt{2\varepsilon}}{2}\left[\sqrt{1 + \left(\frac{\sigma}{\omega\varepsilon}\right)^2} + 1\right]^{1/2} \tag{2.34b}$$

The phase velocity is described by the propagation parameter β along the direction of propagation, which is defined by (2.34b):

$$v_{ph} = \frac{\omega}{\beta} = \frac{\sqrt{2}}{\sqrt{\varepsilon}}\left[\sqrt{1+\left(\frac{\sigma}{\omega\varepsilon}\right)^2}+1\right]^{-1/2} \tag{2.35}$$

The dispersion properties follow from dependence on the frequency of the wave phase velocity $v_{ph} = v_{ph}(\omega)$. Thus, waves with different frequencies $\omega = 2\pi f$ travel with different phase velocities. In the same manner, the wavelength in the medium is dependent on the frequency of the optical wave:

$$\lambda = \frac{2\pi}{\beta} = \frac{\sqrt{2}}{f\sqrt{\varepsilon}}\left[\sqrt{1+\left(\frac{\sigma}{\omega\varepsilon}\right)^2}+1\right]^{-1/2} \tag{2.36}$$

We notice that the field variations with distances are not purely sinusoidal, as in free space. In other words, the wavelength is not exactly equal to the distance between two consecutive positive (or negative) extremes. It is equal to the distance between two alternative zero crossings.

We can now present formulas (2.33) and (2.34) using the general presentation of ε in the complex form, that is, $\varepsilon = \varepsilon' - j\varepsilon"$. If so,

$$\gamma^2 = (\alpha + j\beta)^2 = j\omega\mu(\sigma + j\omega\varepsilon") - \omega^2\mu\varepsilon' \tag{2.37}$$

where now

$$\alpha = \frac{\omega\sqrt{2\varepsilon'}}{2}\left[\sqrt{1+\left(\frac{\sigma+\omega\varepsilon"}{\omega\varepsilon'}\right)^2}-1\right]^{1/2} \tag{2.38a}$$

and

$$\beta = \frac{\omega\sqrt{2\varepsilon'}}{2}\left[\sqrt{1+\left(\frac{\sigma+\omega\varepsilon"}{\omega\varepsilon'}\right)^2}+1\right]^{1/2} \tag{2.38b}$$

From general formulas (2.37) to (2.38), there follow some special cases for different kinds of material media.

Imperfect Dielectric Medium. This medium is characterized by $\sigma \neq 0$,but $\sigma/\omega\varepsilon << 1$. Using the following expansion

$$(1+x)^m = 1 + mx + \frac{m(m-1)}{2!}x^2 + \ldots \tag{2.39}$$

we can easily obtain from (2.37) and (2.38) that

$$\alpha \approx \sqrt{\frac{1}{\varepsilon'}}\frac{\omega\varepsilon''}{2} \tag{2.40a}$$

$$\beta \approx \omega\sqrt{\varepsilon'}\left(1 + \frac{\varepsilon''^2}{8\varepsilon'^2}\right) \tag{2.40b}$$

Now, as has been done from the beginning, we will introduce the complex refractive index $n = n' - jn''$ in the above expressions instead of permittivity, ε, where now $n' = \sqrt{\varepsilon'/\varepsilon_0}$ and $n'' = \sqrt{\varepsilon''/\varepsilon_0}$ [5, 6]. Then, we will get in the case of a low-loss dielectric (or "imperfect" dielectric) that

$$\alpha \approx \frac{\omega n''}{c}, \tag{2.41a}$$

$$\beta \approx \frac{\omega}{c}\sqrt{n'}\left(1 + \frac{n''^2}{8n'^2}\right), \tag{2.41b}$$

and

$$n'' \approx n'\frac{\varepsilon''}{2\varepsilon'} \tag{2.42}$$

Good Conducting Medium. Good conductors are characterized by $\sigma/\omega\varepsilon >> 1$, the opposite of imperfect dielectrics. In this case, the so-called conductivity current component exceeds the polarized (dielectric) current component, that is, $|\boldsymbol{j}_c| \sim \sigma\tilde{E}_x >> |\boldsymbol{j}_d| \sim \omega\varepsilon\tilde{E}_x$. Finally, from (2.45a) and (2.45b), we get:

$$\alpha \approx \frac{\omega n''}{c} \approx \sqrt{\frac{\sigma\omega}{2}} \tag{2.43a}$$

and

$$n'' \approx \sqrt{\frac{\sigma}{2\omega\varepsilon}} \tag{2.43b}$$

Exercises

Exercise 1.

The plane optical wave falls under the angle θ_0 at the boundary of two media with the following parameters: $\varepsilon_{r1} = \varepsilon_{r2} = 3$ and $\mu_1 = \mu_2 = \mu_0$. The electric field of the incident wave equals $E_1(V/m)$.

Find: 1) Angle of refraction θ_2. 2) Amplitude E_2 of the second wave having entered the second medium.

Solution

1) From the beginning we find the relation between the permittivity of these two media

$$\frac{\varepsilon_2}{\varepsilon_1} = \frac{\varepsilon_{r2}\varepsilon_0}{\varepsilon_{r1}\varepsilon_0} = \frac{\varepsilon_{r2}}{\varepsilon_{r1}}.$$

Then, the angle of refraction can be defined as the following:

$$tan\,\theta_2 = \frac{\varepsilon_{r1}}{\varepsilon_{r2}} tan\,\theta_1 \text{ or } \theta_2 = tan^{-1}\left(\frac{\varepsilon_{r1}}{\varepsilon_{r2}} tan\,\theta_1\right)$$

2) Taking into account the boundary conditions described earlier, we get

- for the normal components of the incident and the refracted (entered into the second medium) we get:

$$E_{n1} = E_{n2} \text{ or } E_1\varepsilon_{r1}\,cos\,\theta_1 = E_2\varepsilon_{r2}\,cos\,\theta_2$$

- for the tangential components of the incident and the refracted (entered into the second medium) we get:

$$E_{\tau 1} = E_{\tau 2} \text{ or } E_1\,sin\,\theta_1 = E_2\,sin\,\theta_2$$

3) From the first equation we have:

$$E_2 = E_1 \frac{\varepsilon_{r1}\,cos\,\theta_1}{\varepsilon_{r2}\,cos\,\theta_2}$$

4) From the second equation we have:

$$sin\,\theta_2 = \frac{E_1}{E_2} sin\,\theta_1$$

And replacing E_2 on E_1, we finally get:

$$E_1 \frac{sin\,\theta_1}{sin\,\theta_2} = E_1 \frac{\varepsilon_{r1}\,cos\,\theta_1}{\varepsilon_{r2}\,cos\,\theta_2}$$

5) Then, the angle of refraction can be defined as follows:

$$tan\,\theta_2 = \frac{\varepsilon_{r1}}{\varepsilon_{r2}} tan\,\theta_1 \text{ or } \theta_2 = tan^{-1}\left(\frac{\varepsilon_{r1}}{\varepsilon_{r2}} tan\,\theta_1\right)$$

6) Finally, the refractive field amplitude will be defined from the expression written in item 3, that is,

$$E_2 = E_1 \frac{\varepsilon_{r1}\,cos\,\theta_1}{\varepsilon_{r2}\,cos\left[tan^{-1}\left(\frac{\varepsilon_{r1}}{\varepsilon_{r2}} tan\,\theta_1\right)\right]}, V/m$$

Exercise 2.

A plane optical wave falls under the angle of $\theta_0 = 60^0$ at the boundary of two media with the parameters $\varepsilon_{r1} = 1$, $\varepsilon_{r2} = 3$ and $\mu_{r1} = \mu_{r2} = 1$. The amplitude of the electric field of the wave equals $|E_0| = 3\ (V/m)$.

Find: 1) The coefficients of reflection and refraction for both types of wave polarization. 2) The corresponding amplitudes of the reflection and the refraction wave. 3) Check the obtained results are correct via the corresponding laws.

Solution

1). First of all we will find the relative permittivity

$$\varepsilon_r = \frac{\varepsilon_{r2}\varepsilon_0}{\varepsilon_{r1}\varepsilon_0} = \frac{3}{1} = 3.$$

Then for the incident angle of 60^0, we will get respectively:

$$|R_V| = \frac{\varepsilon_r\,cos\,\theta_0 - \sqrt{\varepsilon_r - sin^2\,\theta_0}}{\varepsilon_r\,cos\,\theta_0 + \sqrt{\varepsilon_r - sin^2\,\theta_0}} = 0$$

$$|R_H| = \left|\frac{cos\,\theta_0 - \sqrt{\varepsilon_r - sin^2\,\theta_0}}{cos\,\theta_0 + \sqrt{\varepsilon_r - sin^2\,\theta_0}}\right| = \frac{1}{2}$$

$$|T_V| = \frac{2\sqrt{\varepsilon_r}\,cos\,\theta_0}{\varepsilon_r\,cos\,\theta_0 + \sqrt{\varepsilon_r - sin^2\,\theta_0}} = 1$$

$$|T_H| = \frac{2\,cos\,\theta_0}{cos\,\theta_0 + \sqrt{\varepsilon_r - sin^2\,\theta_0}} = \frac{1}{2}$$

2) The corresponding components of the reflected and the refracted waves for both types of polarization equal:

$$|E_{1V}| = |R_V||E_0| = 0 \cdot 3 = 0\ (V/m)$$

$$|E_{1H}| = |R_H||E_0| = \frac{1}{2} \cdot 3 = \frac{3}{2}\quad (V/m)$$

$$|E_{2V}| = |T_V||E_0| = 1 \cdot 3 = 3\ (V/m)$$

$$|E_{2H}| = |T_H||E_0| = \frac{1}{2} \cdot 3 = \frac{3}{2}\quad (V/m)$$

3) We check these coefficients and find coincidence with the corresponding laws:

$$|R_H|+|T_H|=1,\ |R_V|+|T_V|=1.$$

Now we check the components of the incident, the reflected, and the refracted waves for both types of polarization that gives:

$$|E_{1V}| + |E_{2V}| = 0 + 3 = 3 \equiv |E_0|$$

$$|E_{1H}| + |E_{2H}| = \frac{3}{2} + \frac{3}{2} = 3 \equiv |E_0|$$

Thus, all above computations are fully correct.

Exercise 3.

A plane optical wave falls under the angle of $\theta_0$at the boundary of two media with the parameters $\varepsilon_{r1} = 4$, $\varepsilon_{r2} = 1$, $\mu_{r1} = \mu_{r2} = 1$, from the first to the second medium. The vector of the incident wave of the vertical polarization equals $\boldsymbol{E}_0 = 5 \cdot \left[cos\,\theta_0\,\boldsymbol{i}_x + sin\,\theta_0\,\boldsymbol{i}_y\right]$ (V/m).

Find: 1) The Brewster angle. 2) The wave field in the second medium in conditions where the incident angle equals the Brewster angle. 3) The critical angle of absence of refraction.

Solution

1) We find the Brewster angle as:

$$\theta_{Br} = tan^{-1}\left(\frac{\varepsilon_2}{\varepsilon_1}\right)^{1/2} = tan^{-1}\left(\frac{1}{4}\right)^{1/2} = 26.56^0$$

2) For incident angle equal to the Brewster angle we find from Snell's law that:

$$\theta_2 = sin^{-1}\left[sin\,\theta_{Br}\sqrt{\frac{\varepsilon_1}{\varepsilon_2}}\right] = sin^{-1}(sin\,2\,6.56^0 \cdot 2) = 63.4^0$$

3) Then the refractive coefficient for the wave in the second medium with vertical polarization equals:

$$|T_V| = \frac{2\sqrt{\varepsilon_{r2}}\,cos\,\theta_0}{\varepsilon_r\,cos\,\theta_0 + \sqrt{\varepsilon_{r2} - sin^2\,\theta_0}} = 2$$

4) Finally, the wave passed from the medium 1 to medium 2 equals:

$$\mathbf{E}_2 = \boldsymbol{E}_0 = 5 \cdot 2 \cdot \left[cos\,\theta_2\,\boldsymbol{i}_x + sin\,\theta_2\,\boldsymbol{i}_y\right] = 10 \cdot \left[0.448\boldsymbol{i}_x + 0.894\boldsymbol{i}_y\right]\ (V/m).$$

5) The critical angle equals:

$$\theta_{\boldsymbol{kr}} = sin^{-1}\left(\frac{\varepsilon_2}{\varepsilon_1}\right)^{1/2} = sin^{-1}\left(\frac{1}{4}\right)^{1/2} = 30^0$$

Exercise 4.

Find the expression of skin depth for copper and intrinsic impedance.

Solution

1) The skin depth for copper is equal to

$$\delta = \frac{1}{\sqrt{\pi f 4\pi \cdot 10^{-7}\left(\frac{1}{s}\right) \cdot 5.8 \cdot 10^{7}\left(\frac{S}{m}\right)}} = \frac{0.066}{\sqrt{f}}(m)$$

2) The amplitude of intrinsic impedance is equal to

$$|\eta| = \sqrt{\frac{2\pi f \cdot 4\pi \cdot 10^{-7}}{5.8 \cdot 10^{7}}} = 3.69 \cdot 10^{-7}\sqrt{f}, \quad \Omega$$

Exercise 5.

In seawater with $\sigma = 4\ S/m$, $\varepsilon' = 81\varepsilon_0$, the frequency of an e/m wave is 100 MHz.

Find: 1) The parameter$\alpha$and the attenuation (in dB/m) considering that transmission of the wave is proportional to $\tau = exp(-\alpha z)$. 2) What it will be if it propagates an optical wave with frequency of 10 THz.

Solution

1a) Since$\sigma/\omega\varepsilon' = 4/(2\pi \cdot 10^{8} Hz \cdot 81 \cdot 10^{-9}/36\pi) = 9 > 1$, we can consider seawater a good conductor. Then, using (2.43a), we get

$$\alpha \approx \sqrt{\frac{\sigma\omega\mu}{2}} = \sqrt{\frac{4 \cdot 2\pi \cdot 10^{8} \cdot 4\pi \cdot 10^{-7}}{2}} = 39.7 m^{-1}$$

1b) Then, attenuation

$$L_t = \frac{1}{z} 10\, log\, \tau = \frac{1}{z} 10 \alpha z\, log\, e = 4.34\alpha = 4.34 \cdot 39.7 = 172.5\ dB \cdot m^{-1}$$

2) Now, for *f*=10 THz, we have for seawater that $n" \approx 0.328$, and

$$\alpha \approx \frac{\omega n"}{c} = \frac{2\pi \cdot 10^{13} Hz \cdot 0.328}{3 \cdot 10^8 m/s} = 6.87 \cdot 10^4 m^{-1}$$

and

$$L_t = 4.34\alpha \approx 2.98 \cdot 10^5 \quad dB \cdot m^{-1}.$$

Exercise 6.

The absorption coefficient of glass at $\lambda = 10 \mu m$ is $\alpha = 1.8 cm^{-1}$.

Find: the imaginary part of refractive index *n*".

Solution

According to expression (2.43a):

$$\alpha \approx \frac{\omega n"}{c} = 1.8 cm^{-1} = \frac{2\pi}{\lambda} n"$$

from which we get

$$n" \approx \frac{\lambda}{2\pi} \alpha = \frac{10^{-5} m \cdot 1.8 \cdot 10^2 m^{-1}}{2\pi} = 2.9 \cdot 10^{-4}$$

Bibliography

[1] Jenkins, F. A., and H. E. White. 1953. *Fundamentals of Optics*. New York: McGraw-Hill.
[2] Born, M., and E. Wolf. 1964. *Principles in Optics*. New York: Pergamon Press.
[3] Fain, V. N., and Ya. N. Hanin. 1965. *Quantum Radiophysics*. Moscow: Sov. Radio (in Russian).
[4] Akhamov, S. A., and R. V. Khohlov.1965. *Problems of Nonlinear Optics*. Moscow: Fizmatgiz (in Russian).
[5] Lipson, S. G., and H. Lipson. 1969. *Optical Physics*. Cambridge: University Press.
[6] Akhamov, S. A., R. V. Khohlov, and A. P. Sukhorukov. 1972. *Laser Handbook*. North Holland: Elsevier.
[7] Marcuse, O. 1972. *Light Transmission Optics*. New York: Van Nostrand-Reinhold Publisher.
[8] Kapany, N. S., and J. J. Burke. 1972. *Optical Waveguides*. Chapter 3, New York: Academic Press.
[9] Fowles, G. R. 1975. *Introduction in Modern Optics*. New York: Holt, Rinehart, and Winston Publishers.
[10] Grant, I. S., and W. R. Phillips. 1975. *Electromagnetism*. New York: John Wiley & Sons.
[11] Plonus, M. A. 1978. *Applied Electromagnetics*. New York: McGraw-Hill.
[12] Kong, J. A. 1986. *Electromagnetic Wave Theory*. New York: John Wiley & Sons.
[13] Elliott, R. S. 1993. *Electromagnetics: History, Theory, and Applications*. New York: IEEE Press.
[14] Blaunstein, N., Sh. Arnon, A. Zilberman, and N. Kopeika. 2010. *Applied Aspects of Optical Communication and LIDAR*. New York: CRC Press, Taylor & Francis Group.
[15] Palais, J. C. 2006. *Optical Communications*, in Handbook: *Engineering Electromagnetics Applications*, edited by R. Bansal. New York: Taylor and Frances.
[16] Blaunstein, N., S. Engelberg, E. Krouk, and M. Sergeev. 2020. *Fiber Optic and Atmospheric Optical Communication*. Hoboken NJ: IEEE Press, Wiley.

CHAPTER 3

CORPUSCULAR NATURE OF LIGHT

3.1. Elements of Quantum Theory

Classical presentation of optical waves, as a part of electromagnetic waves with a narrow spectral band from 200 nm to 750 nm, along the whole electromagnetic spectrum, discussed in Chapters 1 and 2, during its performance from the middle of the nineteenth century to the beginning of the twentieth century, met in its practical applications several paradoxes that could not explain some experimentally observed phenomena, such as:

1. Spectral distribution of radiation excited by a heated body – radiation of the absolute black body.
2. Behavior of an optical wave as a flow of some "virtual" particles – pressure of light.
3. Photoelectric effect.
4. Construction of a stable atom.
5. Radiation and absorption of an atom – linear spectrum.
6. Equivalence and similarity of all atoms of the same elements.

First of all, we will consider the paradox of Maxwell's wave theory (see Chapter 2) application to black body radiation. Let us consider the heated body, as an absolute black body, which fully absorbs all wavelengths of the incident radiation. According to experiments carried out by Rayleigh, the density of the body radiation should increase in proportion to the square of frequency v (e.g., decrease of its intensity with wavelength $\lambda = c/v$), i.e., $\sim v^2$. This law is plotted in Fig. 3.1, shown by the yellow curve for a temperature of T = 5,250 K, where 0 K = -273 $^{\circ}C$.

As can be seen from Fig. 3.1, the density of radiation energy of the black body increases with an increase of radiation intensity and should be fully concentrated at the shortwave part of the spectrum, that is, increases with a decrease of wavelength λ from, say, 780 nm to 380 nm or an increase of frequency from 3.96 THz to 7.89 THz.

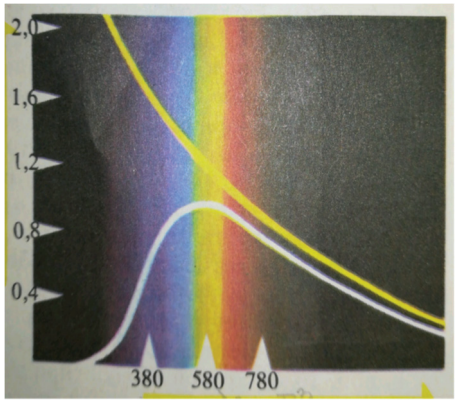


Figure 3.1. The intensity of hot body radiation vs. the wavelength in *nm*: according to classical electromagnetic theory (yellow curve) and to quantum theory (white curve).

To explain the intensity distribution actually obtained as shown by the white curve in Fig. 3.1, in the year 1900 Max Planck postulated that light energy can be transferred, not continuously, but in portions, called "quanta" or "photons". According to Planck, the energy excited by a black body depends only on the frequency (wavelength) of the excited radiation, but not on its intensity, and relates to it by the following mathematical formula:

$$E=hv \tag{3.1}$$

where $h = 6.625 \cdot 10^{-34}\, J \cdot s$ is the Planck constant.

According to Planck, the heat intensity distribution, shown in Fig. 3.1 by the white curve, for some maximum wavelengths (in our case it is 580 nm for T = 5,250 K), the intensity decreases according to an experimentally obtained heat intensity distribution but does not increase to infinity, as shown in Fig. 3.1 by the yellow curve. This phenomenon was postulated by Einstein in 1905, according to which high energy and low energy photons exist as quanta of light, the energy of which does not depend on the intensity of the light radiation, but on its frequency. So, the energy of each photon corresponds to its own frequency or the wavelength of the light (red, yellow, green, violet, and so on). And this is described by Planck's formula (3.1). But what is impossible – photons, as quanta of light, cannot be divided into two or more parts. These light particles are real and below we will present their mass with respect to the mass of an electron.

Moreover, Einstein formulated two principal laws:

- for energy of photon via its mass: $E=mc^2$, $c = 3\cdot10^8$ (m/s);
- for impulse of photon: $P=mc$ or $P=E/c$.

But as was shown in Chapters 1 and 2, light is an electromagnetic wave of specific spectral bands. So, photons also have wave properties. This dualism, called by Einstein the *wave-corpuscular dualism*, was proven experimentally. We will show its proof using a very simple experiment, shown in Fig. 3.2. As seen from the top panel, when one particle of light passes through the specially prepared slit, we obtain its position on the screen. But, when several light photons pass through the slit, they are concentrated sporadically on the screen (second panel from the top). When many photons pass through the slit, they are mostly concentrated at certain points on the screen, which correspond to the maximum of the interference picture, and less concentrated – at the minimum of the interference picture (third panel from the top). So, photons were distributed on the screen in the same manner, as the light as an electromagnetic wave was sent via the slit, as shown in the bottom panel.

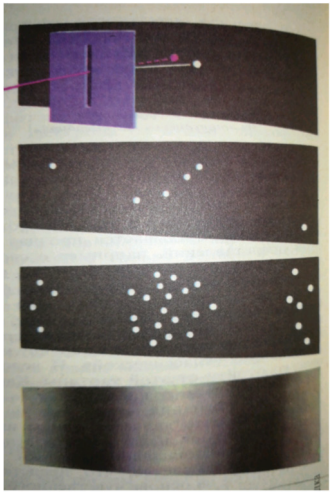


Figure 3.2. Experiment with quanta of light.

de Broglie introduced the relation between the impulse of the photon and the wavelength corresponding to it based on Einstein's law. Thus, if we account for (3.1) and take the relation $\lambda = c/\nu$, we will get: $\lambda =$

$ch/E = ch/mc^2 = h/mc,$ which yields:

$$\lambda = h/mc = h/P \tag{3.2}$$

Similar statements are correct for any corpuscular particle. So, the physical interpretation of *wave-corpuscular dualism* is the following:

The intensity of the wave corresponding to the desired particle at any given point is proportional to the probability of finding this particle at this point.

The de Broglie statement was also proven by Clinton Joseph Davisson and Lester Germer in 1927. They observed the diffraction of electrons as proof of the wave nature of electrons. They sent a beam of fast electrons onto a crystal and obtained a picture similar to that obtained in earlier tests on diffraction of roentgen beams by a monocrystal structure (see Fig. 3.3, top panel). The wavelength of the electron was defined by the use of the distance between the points of the diffraction picture and between the atoms in the crystal. The obtained results totally satisfied the de Broglie formula (3.2). The wavelength of the electron increased with a decrease of electron velocity v (but in the non-relativistic case, when the velocity v is less than the light speed ($v<c$)).
Later, Otto Stern performed similar experiments with beams of neutrons and protons, sent to atoms of Na crystal. Figure 3.3 (bottom panel) presents the diffraction of neutrons by the Na crystal.

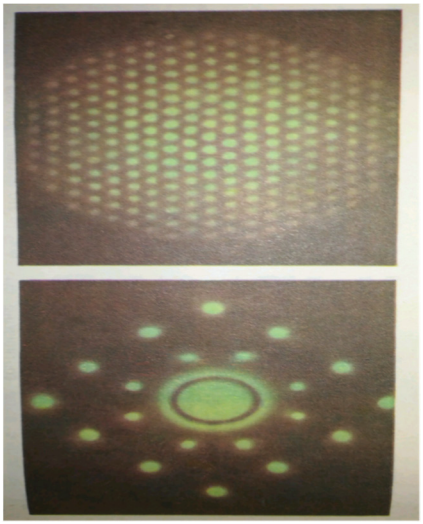


Figure 3.3. Picture of diffraction of roentgen beams by monocrystal (top panel) and neutrons by the Na crystal.

Using this wave-corpuscular dualism, physics met with an experimentally found difficulty, how exactly the impulse of the desired particle and its exact localization in space can be measured. This difficulty was solved by the use of Heisenberg's Principle. In 1925 Heisenberg postulated that for the same time, it is impossible to predict strict coordinates and velocity of any moving particle, such as light photons or electrons. In other words, if we know the exact position x of the particle, its wave function will be in the form of a wave packet with many λ (or $p = mv$ – with many velocities v). Conversely, if the impulse p (or velocity v) of the particle is known exactly, it corresponds to an infinite plane wave with a constant intensity over space. Then this particle cannot be found at any point in space. So, Heisenberg showed that unknowns of impulse Δp and a value of the region where a particle is localized Δx, must be related by the equation:

$$\Delta x \cdot \Delta p = h \tag{3.3}$$

3.2. Structure of the Atom

3.2.1 Wave – Corpuscular Dualism

The quantum theory obtained its final explanation after its usage for the definition of the nature and structure of the atom. In 1903 it was found experimentally that fast electrons pass through atoms. More precisely, this effect was found and explained by Ernest Rutherford in 1911 in experiments on alpha-particle scattering on the positively charged nucleus. He proved the corpuscular model of the atom in contradiction to the wave model of the atom proposed by de Broglie. At the same time, by use of the wave model proposed by de Broglie, if we put an electron in the bounded closed structure, as in a model of an atom, with the length L, the electron will have behavior similar to a wave and its wave function ψ is presented as a sinusoidal wave with maxima and minima at the special points depending on the number of waves N (see Fig. 3.4).

The wavelength of each wave with number n equals $\lambda_n = 2L/n$. n=1, 2,...,N. So, the electron has only a discrete sequence of impulses, $P_n = h/\lambda_n$ or $P_n = n \cdot h /2L$. The corresponding kinetic energy of the electron with number n equals:

$$E_{kn} = (1/2)m \cdot v_n^2 = (1/2)P_n^2 /m = h^2 \cdot n^2 /8m \cdot L^2 \qquad (3.4)$$

So, the energetic levels in the closed bounded structure (atom) are the following:

$$E_{kn} = (1/2)P_n^2 /m = h^2 \cdot n^2 /8m \cdot L^2 \qquad (3.5)$$

Here n is called the *non-zero quantum number*, and m is the mass of electron $m = 9.11 \cdot 10^{-31}$ kg.

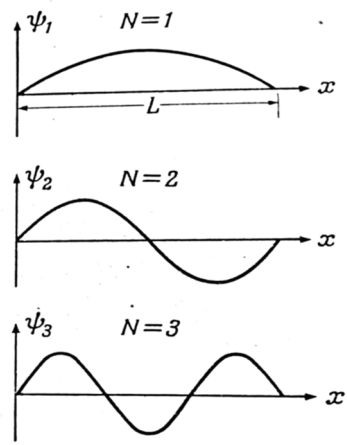


Figure 3.4. Electron wave functions ψ_n , n=1, 2,…, N for electron located inside the atom at the length L from the nucleus.

The corresponding lines that are called the *quantic transfers* or *spectral series*, as shown in Fig. 3.5, are the following:

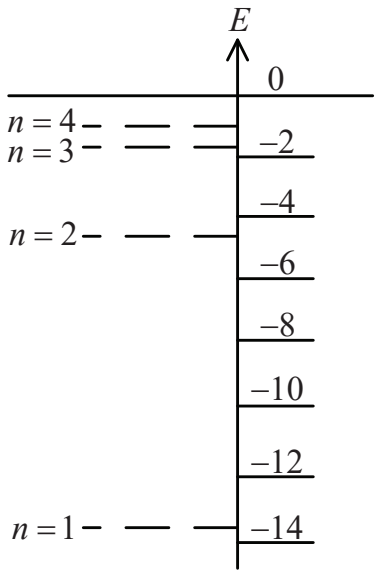


Figure 3.5. Spectral lines of an electron inside the hydrogen atom (H) vs. the values of discrete energy for an electron in the hydrogen atom.

According to the law of energy conservation:

$$h\nu = E_{n'} - E_n = [2\pi^2 \cdot m \cdot e^4/h^2]\,[(1/n'^2 - 1/n^2], \quad n' > n \tag{3.6}$$

or

$$h\nu = 13.6\,[(1/n'^2 - 1/n^2]\ \text{(eV)}, \tag{3.7}$$

where 13.6(eV) = $2\pi^2 \cdot m \cdot e^4/h^2$. If the hydrogen atom is in the stable state regime (n = 1), it obtains energy of -13.6 eV (see Fig. 3.5) which is enough to leave the atom.

The value of 13.6 eV, is, therefore, called the *ionized potential of hydrogen*. We notice that in Fig. 3.5, the energy for each energy level is presented from n = 1 to n = 3.

The minimal energy statement of the electron in the atom can be found for n = 1, as:

$$E_{k1} = (1/2)P_n^2/m = h^2/8m \cdot L^2 \tag{3.8}$$

This energy is called the *zero energy* of the electron. Finally, we can state:

In the closed bounded structure, as in the atom, the energy of the electron can obtain only discrete values.

Next, we will present in Fig. 3.6 the possible transitions of valence electrons for a hydrogen atom from one series of energetic levels to another.

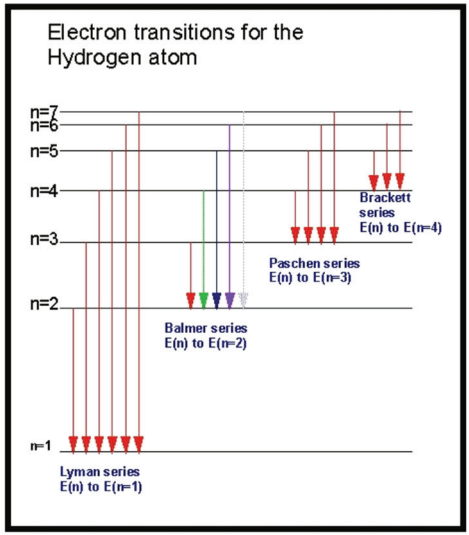


Figure 3.6. Possible transitions of valence electrons in a hydrogen atom.

Thus, transition from high energy levels $E(n)$, $n > 1$, to the "ground" level with $E(n = 1)$ is called the Lyman series; transitions from $E(n)$, $n > 2$ to $E(n = 2)$ are called the Balmer series, from $E(n)$, $n > 3$ to $E(n = 3)$ the Paschen series, and, finally, from $E(n)$, $n > 4$ to $E(n = 4)$ is called the Brackett series.

Finally, we can summarize for any atom the same concept as for the simplest hydrogen atom, that is, each valence electron has its own discrete energy level, as is shown in Fig. 3.7.

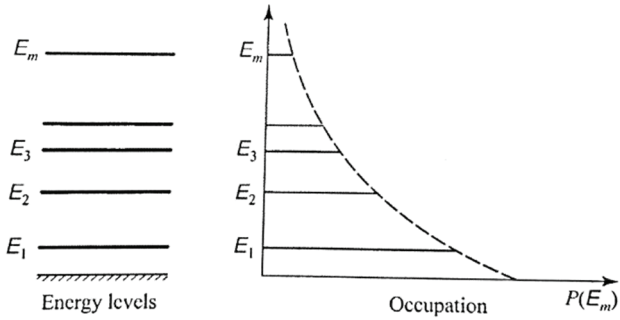


Figure 3.7. The left panel presents discrete energy levels of the valence electrons of each atom, the right panel presents the probability of occupancy of each electron at the corresponding energy level.

3.2.2 Bohr's Corpuscular Model of any Atom

Niels Bohr, in 1913, introduced the concept of the linear structure of atoms based on the photon postulates introduced by Planck and Einstein, which can be presented in the simplest form :

$$h\nu = 13.6\,[(1/n'^2 - 1/n^2)]\ \text{(eV)} \qquad n, n' = 1, 2, 3, \ldots, N. \tag{3.9}$$

From (3.9), Bohr found that the electronic levels of hydrogen equal: $-13.6[1/n^2)]$ (eV), which is clearly seen from Fig. 3.5.

At the same time, in his theory, Bohr also postulated that electrons move along circular orbits (see Fig. 3.8), as in classical physics, from which he obtained the momentum of movement around circular orbits equals $n/(h/2\pi)$. And, finally, in a hydrogen atom in the field of a positive proton, the electron has only discrete values of kinetic energy, described by number $n > 0$, that is:

$$E_n = [2\pi^2 m{\cdot}e^4 / h^2]/n^2 = h^2 n^2 / 8m{\cdot}L^2, \quad n = 1, 2, \ldots, N \tag{3.10}$$

Really, (3.10) describes the energetic levels of electrons in the hydrogen atom (simply called the *energy levels of hydrogen*). The model of atoms according to Bohr's presentation is seen in Fig. 3.8 for four quantum values, n=1, 2, 3, 4 and possible transfer from the higher levels (n=2 to n=4) to the lowers level (n = 1).

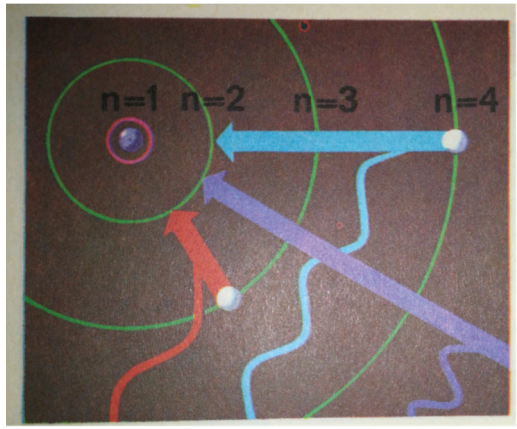


Figure 3.8. Possible transfer of electrons from levels with n>1 to the level of n=1, according to Bohr's model of the atom.

We should state that the linear energy levels of electrons in the hydrogen atom in Figs. 3.5 and 3.6 fully coincide with Bohr's model of the atom presented in Fig. 3.8. To find the orbit momentum of movement P_l of each electron in its own orbit, Schrödinger, analyzing his wave equation introduced for the atom energy states description, declared the quantum number l, a positive number including zero. So, any orbital momentum P_l along the vertical z-axis will have values of $m_l \cdot h/2\pi$. Here $P_l = 2r/\lambda \cdot l$, where r is the radius of the orbit. Vector P_l and the geometry of the problem are shown in Fig. 3.9, where $\cos\theta = m_l/l$, because the full momentum will equal $l \cdot h / 2\pi$.

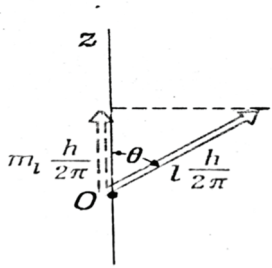


Figure 3.9. Geometry of local momentums of electrons in a hydrogen atom.

Relations between the main moment number n (related to the radius of the orbit r), orbital number l (related to the longitudinal angle θ), and m_l (related to the azimuthal angle φ) in a spherical system of coordinates (r, θ, φ) are the following:

$$\begin{aligned} n &= 1, 2 \\ l &= 0, 1, 2, \ldots, n\text{-}1 \\ m_l &= 0, \pm 1, \pm 2, \pm 3, \ldots, \pm l \end{aligned} \tag{3.11}$$

In 1925 Wolfgang Pauli found that the electronic structure of atoms can be explained if be postulated that at each electron orbit there can be only two electrons, but with opposite vectors of momentum of each electron movement. In other words, according to Pauli's Principle, at any orbit there cannot be more than one electron, but if two electrons exist, their momentum of rotation around the orbit (called *spin orbital momentum*) must be oriented opposite to each other. For these two electrons the spin number will be s = -1/2 for one, for which the momentum of rotation is left-hand oriented, and s = +1/2 for the second, for which the momentum of rotation is right-hand oriented.

For example, for $n = 1$, we have two wave functions corresponding to two electrons with opposite spin numbers:

$$\psi\ (n = 1, l = 0, m_l = 0, s = +1/2) \text{ and } \psi\ (n = 1, l = 0, m_l = 0, s = -1/2)$$

or in the compact form: $\psi_{100}(1/2)$ and $\psi_{100}(-1/2)$.

The information mentioned above allows us to summarize relations between quantum numbers n, l, m_l , and the spin values (usually denoted in the literature by $s/\hbar$ or m_s) in Table 3.1, for the main quantum number n=3.

Table 3.1. Distribution of orbits and orbital electrons depending on the meaning of shells and states.

n	l	m	$s/\hbar$	Allowable states in subshell	Allowable states in complete shell
1	0	0	$\pm\frac{1}{2}$	2	2
2	0	0	$\pm\frac{1}{2}$	2	8
	1	−1	$\pm\frac{1}{2}$	6	
		0	$\pm\frac{1}{2}$		
		1	$\pm\frac{1}{2}$		
3	0	0	$\pm\frac{1}{2}$	2	18
	1	−1	$\pm\frac{1}{2}$	6	
		0	$\pm\frac{1}{2}$		
		1	$\pm\frac{1}{2}$		
	2	−2	$\pm\frac{1}{2}$	10	
		−1	$\pm\frac{1}{2}$		
		0	$\pm\frac{1}{2}$		
		1	$\pm\frac{1}{2}$		
		2	$\pm\frac{1}{2}$		

Here for $l = 0$ (only one subshell), we have two allowable states and, therefore, two electrons at the same orbit with opposite spins; in total, 2 electrons with opposite spins. For $l = 0$ and $l = 1$ (two subshells), we have for each shell 2 electrons (for $l = 0$) and 6 electrons with opposite spins (for $l = 1$); in total – 8 electrons, and so forth.

To finish discussions on the corpuscular description of atom structures according to the quantum theory, we should notice that in the theory of semiconductors, which are usually the basic material of optical devices used in photonics and optical communication, wired and wireless, other notations of different electron states in atoms are used. For the definition of different states of the electron in the hydrogen atom the traditional (from spectroscopy) notations are usually used for each value of l. Thus, for $l = 0$, the following are used, symbol s, for $l = 1$ – symbol p, for $l = 2$ – symbol d, for $l = 3$ – symbol f, and for $l = 4$ – symbol g.

An example of relative displacement of energy levels with n=3 (3s, 3p and 3d levels) and n=4 (4s and 4p levels) in Na atoms is shown in Fig. 3.10, where the level (n=3, l=0) is written 3s; level (n=3, l=1) is written 3p, and the level (n=3, l=n-1=2) is written 3d. This means, according to (3.11), that for n=1, l=0, we have 1s status with $m_l = 0$ or giving only one energy level (i.e., orbit) and accounting for $s = -1/2$ and $s = +1/2$, finally giving, for this level, 2 electrons. When n=2, l=0, 1, we have 2s and 2p statuses with $m_l = 0, -1, +1$, i.e., 4 levels (orbits), and accounting for $s = -1/2$ and $s = +1/2$ for each level, yields finally 8 electrons. For n=3, l=0, 1, 2, we have 3s, 3p and 3d statuses with m_l =0 (l=0); m_l =0, -1, +1 (l=1); and m_l =0, -1, +1, -2, +2 (l=2), that is, 9 orbits. Accounting now for spin numbers $s = -1/2$ and $s = +1/2$, we find that 18 electrons fill these 9 levels (orbits).

The first generalization of the Bohr Theory was carried out by Arnold Sommerfield. He took several postulates from astronomy, the main one being that electrons can move along not only circular, but also elliptical orbits, as was postulated in astronomy by Kepler. So, to the main quantum number n correspond now n ellipses with different numbers of their centers. For $n = 1$ there exists only one orbit which is denoted by symbol 1s and its axis is one and the second is related as 1:1; that is, we have a circular orbit for 1s. We will now use index s, p, d, f, g, which is more convenient with respect to orbit quantum numbers l. All these states (orbits) are presented graphically as shown in Fig. 3.10a-c .

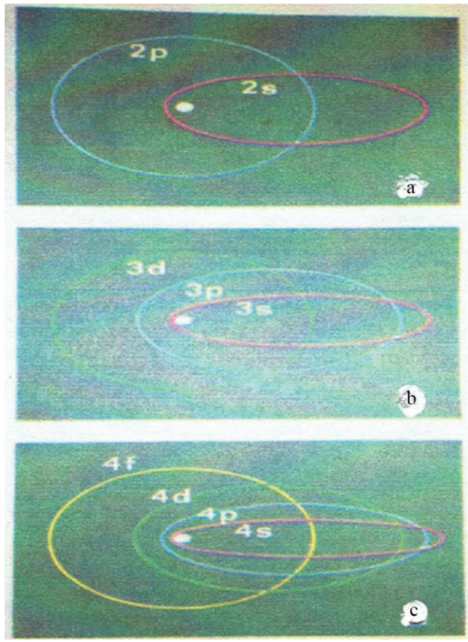


Figure 3.10a-c. Model of the atom as proposed by Sommerfield.

A general view of the relations between the symbols and the possible orientations of the spatial orbits (e.g., amount of orbital quantum numbers m_l) is presented in Table 3.2.

Table 3.2. The number of states and electron orbits' orientation in space around the nucleus of the atom.

Symbols	s	p	d	f	g
Quantum number *l*	0	1	2	3	4
Number of possible orientations of the orbit in space	1	3	5	7	9

Namely, a state 3*d* in the middle panel of Fig. 10 relates to n=3, with quantum number l=2, and has 5 different spatial orientations of orbital momentum of impulse that are placed normal to the plane of the orbit.

The rule of filling of energy levels discussed above can be broadened for all systems of chemical elements, which states that, according to the prohibiting law of Pauli, electrons first fill the lowest discrete levels and then the highest discrete level of any atom from the Mendeleev table of chemical elements. We present here only those stable atoms, among all 92 atoms, which are usually used as a basic material of gaseous, liquid, and solid optical devices (see Table 3.3).

Table 3.3. Part of Mendeleev's elements creating the main types of conductors and semiconductors and based on their use in optical devices.

Atomic number (Z)	Element	Number of electrons				Shorthand notation
		n = 1, **l** = 0	2: 0 1	3: 0 1 2	4: 0 1	
		1s	2s 2p	3s 3p 3d	4s 4p	
1	H	1				$1s^1$
2	He	2				$1s^2$
3	Li	helium core, 2 electrons	1			$1s^2\ 2s^1$
4	Be		2			$1s^2\ 2s^2$
5	B		2 1			$1s^2\ 2s^2\ 2p^1$
6	C		2 2			$1s^2\ 2s^2\ 2p^2$
7	N		2 3			$1s^2\ 2s^2\ 2p^3$
8	O		2 4			$1s^2\ 2s^2\ 2p^4$
9	F		2 5			$1s^2\ 2s^2\ 2p^5$
10	Ne		2 6			$1s^2\ 2s^2\ 2p^6$
11	Na	neon core, 10 electrons		1		[Ne] $3s^1$
12	Mg			2		$3s^2$
13	Al			2 1		$3s^2\ 3p^1$
14	Si			2 2		$3s^2\ 3p^2$
15	P			2 3		$3s^2\ 3p^3$
16	S			2 4		$3s^2\ 3p^4$
17	Cl			2 5		$3s^2\ 3p^5$
18	Ar			2 6		$3s^2\ 3p^6$
19	K	argon core, 18 electrons			1	[Ar] $4s^1$
20	Ca				2	$4s^2$
21	Sc			1	2	$3d^1\ 4s^2$
22	Ti			2	2	$3d^2\ 4s^2$
23	V			3	2	$3d^3\ 4s^2$
24	Cr			5	1	$3d^5\ 4s^1$
25	Mn			5	2	$3d^5\ 4s^2$
26	Fe			6	2	$3d^6\ 4s^2$
27	Co			7	2	$3d^7\ 4s^2$
28	Ni			8	2	$3d^8\ 4s^2$
29	Cu			10	1	$3d^{10}\ 4s^1$
30	Zn			10	2	$3d^{10}\ 4s^2$
31	Ga			10	2 1	$3d^{10}\ 4s^2\ 4p^1$
32	Ge			10	2 2	$3d^{10}\ 4s^2\ 4p^2$
33	As			10	2 3	$3d^{10}\ 4s^2\ 4p^3$
34	Se			10	2 4	$3d^{10}\ 4s^2\ 4p^4$
35	Br			10	2 5	$3d^{10}\ 4s^2\ 4p^5$
36	Kr			10	2 6	$3d^{10}\ 4s^2\ 4p^6$

The top of Table 3.3 presents the main quantum number n = 1, 2 …, 4. The corresponding charge number Z indicates for each element their relations to the groups of elements – from II, III, …, VI, to which the desired element is related. For each group of elements the second wide column presents the number of valence electrons for each desired element. Thus, the carbon (C) from group II has 6 (2+2+2) electrons redistributed at the 1s (2 electrons), 2s (2 electrons) and 2p (2 electrons). The oxygen (O) from the same group II has 8 valence electrons, (2+2+4) electrons redistributed at the 1s (2 electrons), 2s (2 electrons) and 2p (4 electrons), and so forth.

3.2.3 De Brogli's Wave – Corpuscular Dualism Concept

Accounting now for a corpuscular-wave description, and using for each atom the wave theory postulated by de Brogli, according to which and to the corresponding Schrödinger's equation, to each electron in various states corresponds its own wave function ψ, which describes some "replaced in space" electron, or the cloud of electrons around the nucleus of the atom. At the same time, according to Niels Bohr and then to Sommerfield's concept, this function ψ simply is the probability of finding any electron at its own orbit, corresponding to its discrete energy.

We will now put a question: How will an atom look if we can create a photo of the position of a single electron and many electrons fixed in various moments of time. We will present a very simple virtual experiment, shown in Fig. 3.11. We put photos – one, two, three, and many together for electrons in simple state 1s of hydrogen.

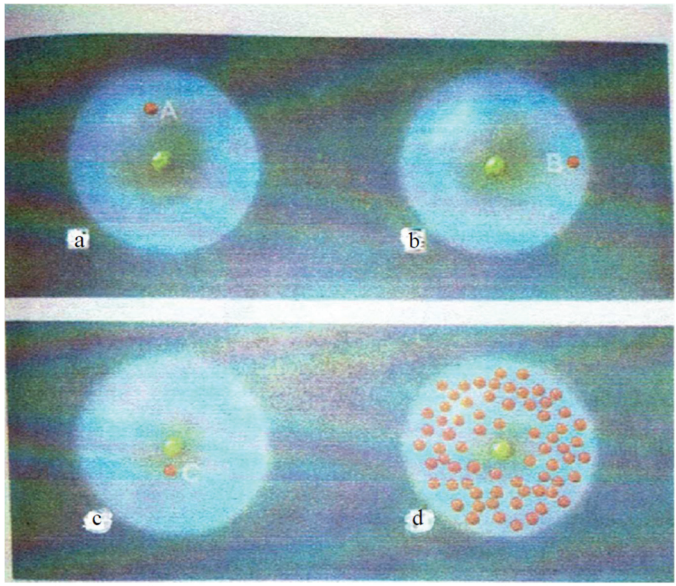


Figure 3.11. Presentation of randomly positioned electrons corresponding to 1s-conditions of hydrogen: a) electron A, b) electron B, c) electron C; d) many electrons together.

Firstly, we selected virtually electron A (the red sphere in Fig. 3.11a), then electron B (red sphere in Fig. 3.11b), and then electron C (red sphere in Fig. 3.11c) placed separately at various positions around the nucleus. If we now put many such photos of various electrons with their randomly distributed positions around the nucleus, we will see a stationary picture, which will correspond to the condition 1s of the hydrogen atom, as shown by Fig. 3.11d.

Figure 3.12 illustrates the electron cloud as a symbiosis of many close orbits of any multi-electron atom with the charge number Z and mass M, the discrete energy levels of which are generalized and presented above in Eqs. (3.5) and (3.10).

$$E_n = M \cdot Z^2 \cdot e^2 / [(4\pi\varepsilon)^2 \cdot 2h \cdot L^2 \cdot n^2], \; n=1, 2, \ldots, N \qquad (3.12)$$

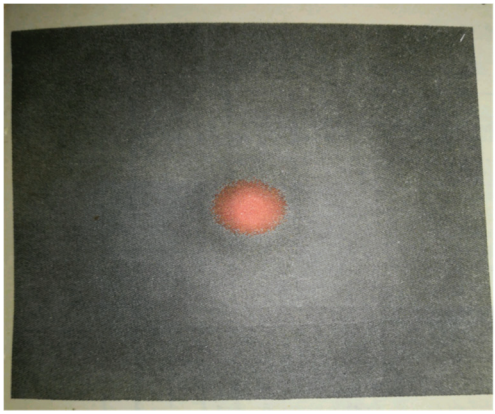


Figure 3.12. A view of a multi-electron atom with a cloud of electrons.

In Fig. 3.12, an electron cloud with $n = 1$ is shown by the dark color, and the outer electron with $n = 2$ by the lighter color. According to the statements above, $n = 2$, $l = 0$ and $n = 2$, $l = 1$ should have the same energy.

Let us give some examples:

1. Hydrogen (H) with a state $n = 1$, and Z = 1, consisting of one electron, has an ionization potential equaling $E_1 = 13.6$ eV to leave the atom (see above).
2. Helium (He) with Z = 2, according to formula (3.12), has ionization energy $Z^2 \cdot E_1 = 4 \cdot 13.6 = 54.5$ eV, which is in good agreement with experimental data.

3.2.4 Structure of Crystal Materials

First of all, we briefly describe the principal differences of solid crystal-like materials with respect to molecules and liquids. In crystal structures, the atoms are strongly localized at the corners of the crystal grid inside it. The relation between atoms occurs from electrons, which, according to the information presented in Fig. 3.13, are not exclusive to any one atom, but relate to all atoms inside the crystal grid because their wave functions spread through the lattice structure.

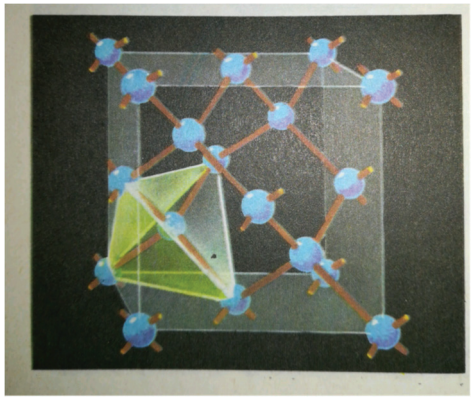


Figure 3.13. Grid structure of diamond. The blue spheres are atoms and the yellow small spheres surrounding them are valence electrons.

The existence of a crystal grid and oscillations of its composite elements is the matter of physics of solid bodies. We do not enter into deep discussions here, transferring the reader to the excellent books [1–3]. We will, however, notice that the properties of solid bodies are closely related to the types of relations between atoms: ionic and/or covalent (e.g., atomic or chemical). The structure of ionic grids, because they are constructed by pure electrostatic forces, has the same nature as molecules and liquids. As for covalent or atomic grid structures, it is more complicated to understand because their components "enjoy" each other. Fig. 3.13 presents an example of the crystal grid of diamond constructed on covalent (atomic) relations. Each atom is surrounded by four other atoms creating a quadratic-form geometrical structure. Simultaneously each of the atoms is the tip of such neighboring structures (see lower left of Fig. 3.13).

Many properties of solid substances, particularly their conductivity, can be fully explained by the use of the zonal model. Namely, the zone model can explain why dielectrics, semiconductors, and metals differ from each other. As was mentioned above, according to Pauli's Law, each level of an isolated atom can be occupied by not more than 2 electrons with opposite spins. Isolated atoms have thick lines, 1s, 2s, 2p, 3s, 3p,…, as shown in Fig. 3.14 (left panel).

In metals (Fig. 3.14, second panel) all levels are occupied, then in the energetic zone not even one electron is absent. They are seated in the so-called *valence zone* (shown by the black color). All electrons in this zone are not free and cannot move to create a current in the upper, *conductive zone*. In the latter zone in outer conditions (heating, pumping by an outer

light source or outer electric field), an electrical current can be created. In some elements, the highest level is not fully filled by electrons (such as metals, Na, Ca, and so on, having only one electron at their levels). So, for metals, from N separate levels, around half can be free to be filled. Thus, for metals, as shown in Fig. 3.14 (second panel), enough small voltage or light energy transmitted to electrons allows them to jump from lower levels to higher levels and, finally, pass via the prohibiting zone with energy E_g, as shown in Fig. 3.14.

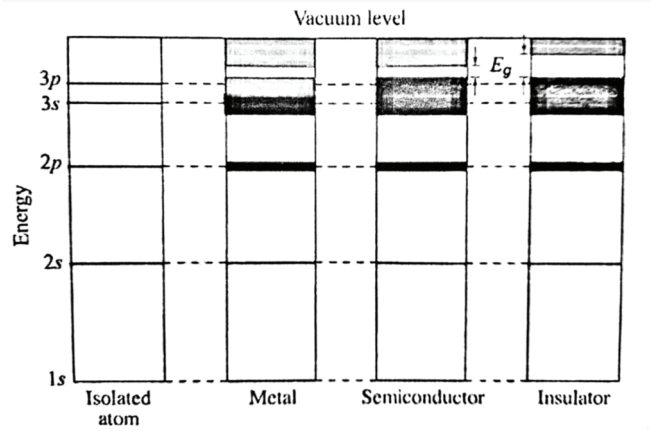


Fig. 3.14. Schematically presented zone structures of the separate atom, the metal, the semiconductor, and the insulator. (Pure dielectric.)

So, metals, such as Mg, Ca, and so on, having two electrons at each level, are very good conductors. Here due to the wide filled zone, it fully overlaps the prohibited regions of energy between zones and enters into a free zone. This effect is called overlapping of zones, and finally, electrons can move in space not occupied by other electrons. However, when the prohibiting zone is too wide (see Fig. 3.14, third and fourth panels), such overlapping is impossible. Moreover, it is impossible to convert electrons between levels, let's say, from 3s to 3p, from 4s to 3d (as it is fully occupied), 4p to 4d (it is also fully occupied), and so on. Such solids are called *semiconductors* and *dielectrics* (*isolators*) (see Fig. 3.14, last two panels).

3.3. Semiconductor Fundamentals

Below, we briefly, without entering into deep mathematical descriptions of the subject because it is out of the scope of this book, introduce the reader to semiconducting material fundamentals because they are the most applicable materials in *photonics*, dealing with photon flows, and *optoelectronics*, dealing with electron and hole flows. Both these fields are based on semiconducting materials, which absorb and emit photons by undergoing transitions among the desired energy levels of semiconductors as crystal materials. Indeed, the photons generate electrons and holes, and charged particles generate and control the flow of photons. We should, from the beginning, notice that:

- A semiconductor lattice cannot be viewed as a collection of non-interacting atoms, each with its own individual energy levels and probability (wave function). This is because in the proximity of each atom in the crystal lattice, the energy levels belong to the system as a whole and the wave functions ψ overlap each other (see Fig. 3.15).

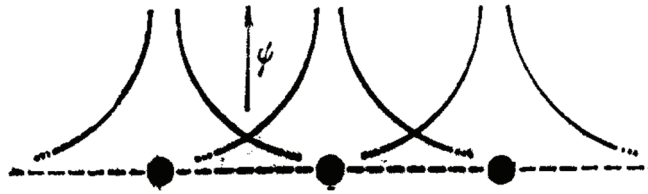


Figure 3.15. Wave functions of each atom in a semiconductor overlap each other according to wave-corpuscular dualism.

- Collections of close spaced energy levels form energy bands, which for $T = 0\ K$ or in the absence of an external excitation source, are either fully occupied or totally non-occupied. The higher non-occupying energy band is called the *conductive band*, whereas the lower fully occupied band is called the *valence band*. These two bands are separated by a forbidden band with the gap energy *Eg*, as shown in Fig. 3.14.

3.3.1. Zonal Structure of Semiconductors

To understand these main properties of semiconducting materials, we will return the reader to what was discussed from the beginning, i.e., to the zonal model of crystals. As was shown by Schrödinger and follows from his equation, for electron energy in a field of periodical potential that describes a collection of atoms in the lattice, splitting of the atomic energy levels and formation of energy bands results. Indeed, the crystal lattice potential associated with an infinite 1-D collection of atoms with lattice constant a, which is depicted schematically in Fig, 3.16a, can be approximated by a 1-D periodical rectangular potential introduced for the simplified Schrödinger's 1-D model, as illustrated by Fig. 3.16b. This model proves the results of the Schrödinger equation for such potential predicted energy bands with traveling-wave solutions, separated by prohibited bands with exponentially decaying solutions. The obtained results were also proved for the 3D case.

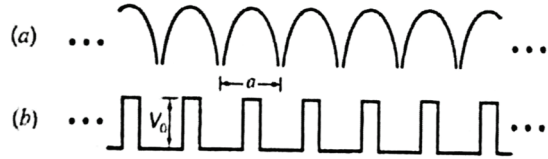


Figure 3.16. a) Solution of the Schrödinger's 1-D model, and b) its approximation.

Finally, the information mentioned above allows us to present for each type of semiconductor, pure or composite (doped), its own zonal structure. As an example, Fig. 3.17 presents the zonal structure for two semiconductors: Si (Silicon), as a pure semiconductor, and GaAs (Gallium-Arsenide), as a compound semiconductor. Here, each band contains a large number of densely packed discrete energy levels that can be approximated as a continuum (see Fig. 3.17). Thus, for Si E_g = 1.12 eV, while for GaAs E_g = 1.42 eV at a room temperature of 300 K.

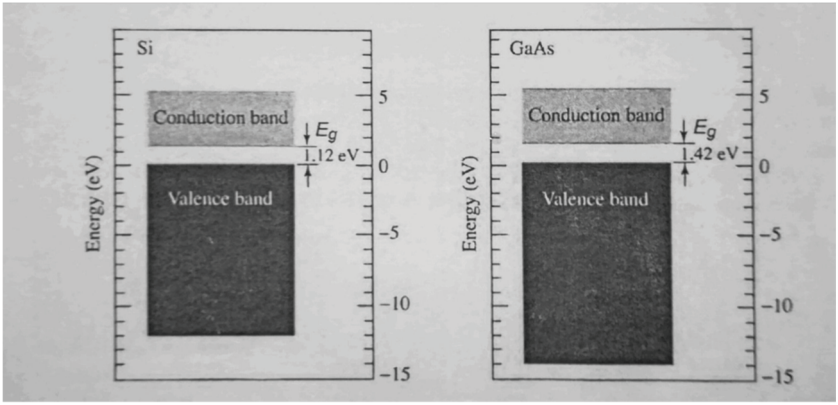


Figure 3.17. Zonal structure of Si (left panel) and GaAs (right panel).

As illustrated in Fig. 3.17, the valence and conductive bands are separated by an energy bandgap. These bands play a fundamental role in the definition of the electrical and optical properties of semiconductors, and not only them but also isolators and conductors.

3.3.2 Electrons and Holes

As was mentioned earlier, the wave functions $\psi(r)$ of electrons in semiconductors overlap and the *Pauli exclusion* and *prohibiting* principle applies and declares that no two electrons can occupy the same energy level, and if this does occur, the two electrons have opposite spin momentum number $s = -1/2$ and $s = +1/2$.

At a low temperature of $T = -273$ °C = 0 K, the energy levels inside the valence zone are fully occupied, while the conduction band is fully empty. With an increase of T, some electrons can be thermally excited from the valence zone into the empty conduction zone, where the occupied levels are abundant (see Fig. 3.18). If now an outer electric field is applied, these free (conductive) electrons can drift through the lattice as mobile carriers, creating the electric currents inside the semiconductor. Moreover, moving high energy electrons give room for electrons occupying lower energy levels inside the valence zone, to go upward to these liberated levels. The places of liberation are called *holes*. The movements of holes, therefore, are in the opposite direction to the movements of electrons. The hole therefore behaves as if it has a positive charge $+e$, opposite to the charge of the electron $-e$.

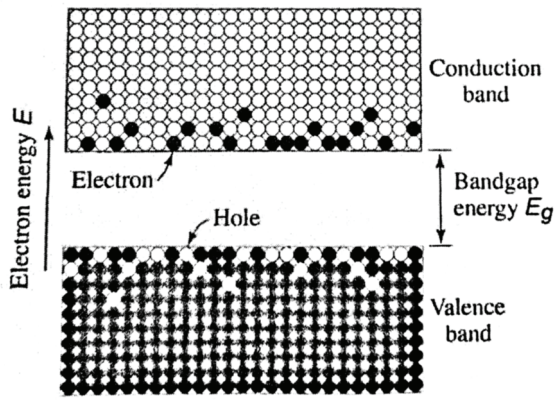


Figure 3.18. Locations of electrons and holes in the conductive and the valence zone, respectively.

To explain how electrons and holes move inside the crystal structure of a semiconductor, and for more evidence, we present in Fig. 3.19 the 2-D scheme of the 3-D model of crystal presented in Fig. 3.13, but for the specific case of a crystal of pure semiconductor Ge (germanium).

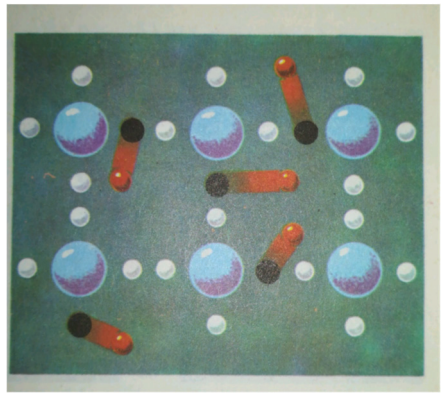


Figure 3.19. A fragment of 2-D structure of Ge: red spheres are electrons, black spheres are holes.

The red spheres are electrons that were transferred in the conductive zone. We notice that after leaving the corresponding atom, the electron creates an empty place called a *hole*. Because a crystal grid in normal conditions is electrically neutral, the existence of holes shows that at this place an electron is absent, which is equivalent to the existence in this place of a *positive charge* (+). Therefore, a pure semiconductor consists of an equal amount of negative charges (*electrons*) and positive charges (*holes*). In reality, holes react in the same manner as if they are positive charges.

Indeed, the similarity of holes with positive charges can be easily understood by introducing to the crystal a source of voltage. Then holes will move to the "–" electrode, but electrons will move to "+" electrode of the source. In reality, holes do not move but the same result can be obtained if the free electron from the neighboring pair enters into this vacant hole.

3.3.3 Joint Energy-Momentum Domain of Semiconductors

Before entering into the subject, let us notice the following. As was mentioned earlier, the energy of an electron in the free space with constant potential has the energy

$$E=p^2/2m_0 = h^2 \cdot k^2/2m_0 \tag{3.13}$$

where p is the absolute value of the momentum of the electron, k is the magnitude of the wave vector $\mathbf{k} = \mathbf{p}/h$, h is the Planck constant defined above, and m_0 is the electron mass, $m_0 = 9.11 \cdot 10^{-31}$ kg. As follows from Eq. (3.13), the E-k relation for a free electron is a simple parabola $E \sim k^2$.

In semiconductors, according to the Schrödinger 1-D model and its approximation made by Kronig-Penney (see Fig. 3.16b) due to the periodic potential generated by charges in the periodic crystal lattice, the E-k relations for electrons and holes in the conductive and valence zones have a form, as is presented for Si and GaAs in Fig. 3.20.

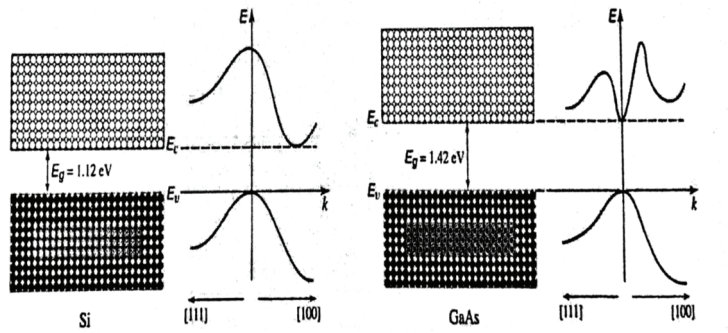


Figure 3.20. The *E-k* function for Si and GaAs along two crystal directions: toward the left and toward to right (rearranged from [1–3]).

The energy E is the periodic function of the components (k_1, k_2, k_3) of the vector **k**, with periodicities (π/a_1, π/a_2, π/a_3), where a_1, a_2, a_3 are the crystal lattice constants. In Figure 3.20, the particular directions of *E-k* cross section are presented: $k_1 = k_2 = k_3 = 1$ [1,1,1] (the left side in Fig. 3.20) and k_1=1, k_2=k_3 = 0 [1,0,0] (the right side in Fig. 3.20). It should be mentioned that the range of k values in the interval [-π/a. π/a] defines the first Brillouin zone [1–3]. Electrons not placed in the first Brillouin zone fill the second zone. They also fill the region between planes of the first zone and planes defined by conditions of diffraction from planes [110]. So, the energy of an electron in the conduction zone depends not only on the magnitude of its momentum but on the direction of its drifting in the semiconductor.

As follows from the illustration in Figure 3.21, near the bottom of the conduction band, the E-k relation can be described by the parabola:

$$E = E_c + h^2 \cdot k^2 / 2m_c \tag{3.14}$$

Here E_c is the energy at the bottom of the conduction band, m_c is the electron conduction band mass, and k is measured from the wave vector, where the minimum of energy occurs. The effective mass m_c, which differs from that in free space m_0, is the result of the influence of the ions of the lattice on the motion of a conduction band electron.

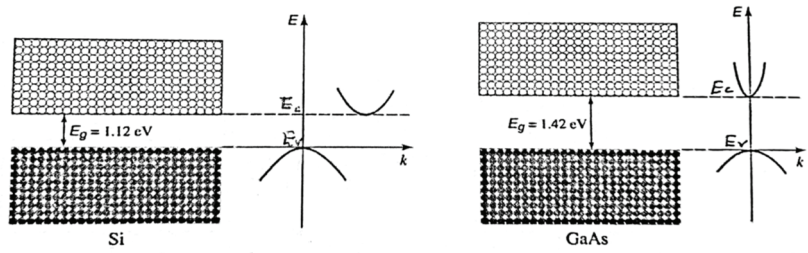


Figure 3.21. *E-k* diagram, as in Fig. 3.20, but approximated by parabolas for Si and GaAs semiconductors.

Similarly, near the top of the valence band, we get:

$$E = E_v + h^2 \cdot k^2 / 2m_v \tag{3.15}$$

where $E_v = E_c - E_g$ is the energy at the top of the valence band and m_v is hole valence band effective mass, which determines the effects of the lattice ions on the motion of the valence band hole. It depends on the crystal structure and direction of travel with respect to the lattice.

Approximating the *E-k* diagram for Si and GaAs by parabolas at the bottom of the conduction band and at the top of the valence band, as shown in Figure 3.21, we can explain such effective mass m_c for electrons and holes in the conductive and valence zones, respectively.

As an example, we present typical values of electron and hole masses in some selected semiconductors (with respect to the mass m_0 in free space). Thus, according to [1–3]:

for Si (indirect-bandgap)	$m_c/m_0 = 0.98$;	$m_v/m_0 = 0.49$
for GaAs (direct-bandgap)	$m_c/m_0 = 0.07$;	$m_v/m_0 = 0.50$
for GaN	$m_c/m_0 = 0.20$;	$m_v/m_0 = 0.80$

Semiconductors for which the conduction band and the valence band minimum energy correspond to the same k are *direct-bandgap* semiconductors. Otherwise, they are *indirect-bandgap* semiconductors.

3.3.4 P-Type and N-Type Semiconductors

Semiconducting material in which the amount of electrons prevail (with respect to the amount of holes) are called *n-type* semiconductors, and those where holes prevail are called *p-type* semiconductors. Combining these types of pure semiconducting materials, we finally obtain the *p-n* or *compound* semiconductors.

Let us use Sb, which is a fifth-valence substance, having one valence electron more than in Ge (see Table 3.3). But the atom of Sb is in a grid in the same manner as in the pure Ge (see Fig. 3.22).

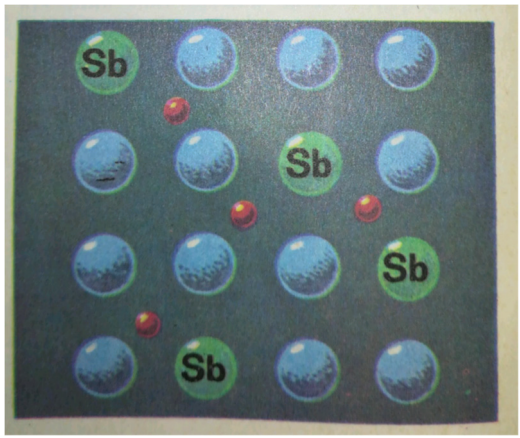


Figure 3.22. Compound GeSb *n*-type semiconductor.

The connection of outer electrons with other atoms in Sb is very low (~ 0.5 eV) and the thermal energy for $T = 290\ K$ is enough to liberate electrons to start to produce electrical current. As a result, we obtained the composite semiconductor on *n-type* with electrons prevailing – negative carriers of charge.

If now in the crystal grid, instead of atoms of Ge, In atoms are introduced In (see Table 3.3), having only a three-valence state, instead of a four-valence state of Ge, we will have an absence of an electron – or an additional hole, as shown in Fig. 3.23.

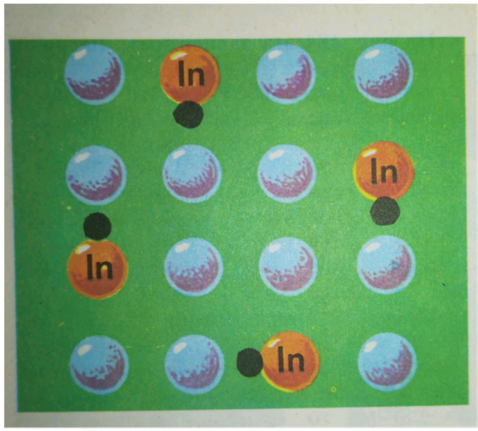


Figure 3.23. Compound InGe *p*-type semiconductor.

Such a semiconductor is called the composite InGe *p*-type semiconductor (with the absence of conductive electrons).

3.3.5 P-N Junction in Equilibrium

Overlapping differently pure regions of single semiconductor material are called *home-junctions*. The important example is a *p-n junction*, which occurs if in contact with both types of semiconductors, *n*-type and *p*-type, as shown in Fig. 3.24.

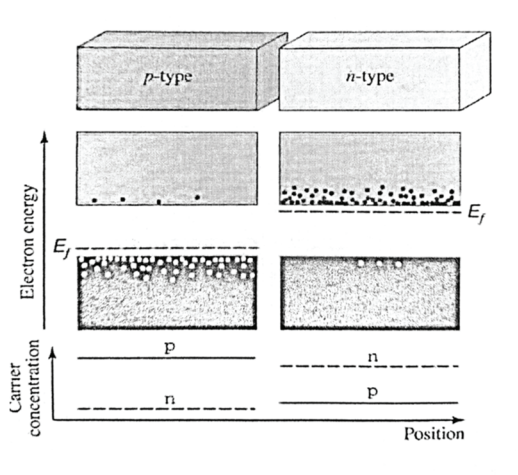


Figure 3.24. Schematically presented overlapping of two kinds of semiconductors, the *n*-type and the *p*-type.

A *p*-*n* junction consists of a *p*-type and an *n*-type section of the same semiconductor materials. The *p*-type region has many holes (majority carriers) and few mobile electrons (minority carriers) [see left-side two middle blocks in Fig 3.24]. The *n*-type region has many mobile electrons (majority carriers) and few holes (minority carriers) [see right-side two middle blocks in Fig. 3.24]. Both charge carriers are in conditions of random thermal motion in all directions inside their own materials.

The lower panel in Fig. 3.24 shows clearly the fact that the concentration of holes in *p*-type material is much higher than electrons (denoted by *n*), whereas the concentration of free electrons in *n*-type material is prevalent with respect to holes.

Fermi energy E_f, which defines the *minimum energy* needed to transfer an electron from the upper level of the valence zone to the lower level of the conductive zone, lies for both separate materials closer to each of the types: above the valence zone of the *p*-type semiconductor and below the conductive zone of the *n*-type semiconductor (shown in the two middle panels by dashed lines).

In this case, when $T > 0\ K$, electrons from the *n*-type semiconductor will penetrate to the *p*-type semiconductor through the junction created between them. In the same manner, the holes will penetrate from the *p*-type semiconductor to the *n*-type semiconductor via the junction. Finally, they will create a spatial electrical charge difference inside the junction, and therefore, the inner electric field, as shown in Fig. 3.25a.

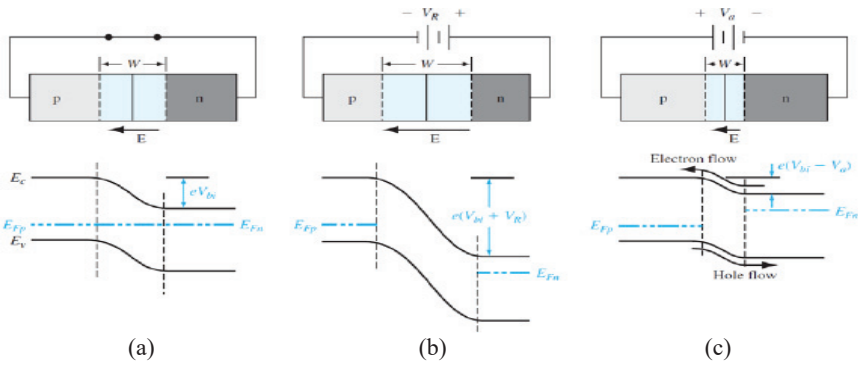


Figure 3.25. a) *p-n* semiconductor in equilibrium state (in absence of electric field); b) opposite to inner electric field; and (c) *p-n* junction under outer source – direct to inner electric field.

The inner electric field regulates the number of electron-hole pairs, which can be increased until that time, when this process is not compensated for by the inverse process of recombination of electron-hole pairs, as major carriers of charges. In this case, the condition of *dynamic equilibrium* is observed. Of course, the width of this junction is too thick – around a few micrometers [~10^{-3} mm].

When an outer electric source generates charge "–" in the *n*-region, and "+" in the *p*-region (see Fig. 3.25c), the outer and inner fields have the same direction, increasing the total current through the circuit. The *p-n* junction works as a *direct biased junction*.

When an outer electric source generates charge "–" in the *p*-region, and "+" in the *n*-region, the outer field has the opposite direction to the inner field and the absence of the current through the circuit is observed (see Fig. 3.25b). The *p-n* junction works as an *opposite biased junction*.

Exercises

Exercise 1.

1) Find the relation of photon energy in "eV" with the wavelength in "Ångstrom" (Å).
2) Find the wavelength (in μm, nm, and Å) of the proton with energy $E = 1$ MeV.

Solution

First of all we will find relations between the dimensions of the quantities. Thus:

$$1 \text{ eV} = 1.6\ 10^{-19} \text{ J} = 1.6\ 10^{-12} \text{ erg};$$
$$1 \quad \text{Ă} = 10^{-10}\ m = 10^{-4}\ \mu m = 0.1\ nm$$

(1) Now, we will find the relation between the impulse of a particle and its wavelength by use of de Broglie postulate and Planck's law, that is:

$$\lambda\ (\text{Ă}) = c/\nu = c \cdot h\ /\ \nu \cdot h = [6.62\ 10^{-34}\ (\text{J s}) \cdot 3 \cdot 10^{8}\ (\text{m/s})]\ /\ h\nu\ (\text{J}) =$$
$$= [6.61 \cdot 10^{-34}\ (\text{J s}) \cdot 3 \cdot 10^{8}\ (\text{m/s})]\ /\ h\nu\ (\text{eV}) \cdot 1.6 \cdot 10^{-19} =$$
$$= 12.390 \cdot 10^{-7}\ /\ h\nu\ (\text{eV})\ [\text{m}] = 12{,}390\ /\ h\nu\ (\text{eV})\ [\text{Ă}].$$
$$\text{Thus:} \quad \lambda\ (\text{Ă}) = 12{,}390\ /\ h\nu\ (\text{eV})\ [\text{Ă}] = 12.39/\ h\nu\ (\text{keV})\ [\text{Ă}]$$

Conclusion: If the wavelength corresponding to a particle decreases (frequency increases) then the energy of the particle increases, and vice versa.

2) For proton with energy E=1 Mev = $1 \cdot 10^{6}$ eV, we get:
$\lambda\ (\text{Ă}) = 12{,}390\ /\ E\ (\text{eV}) = 12{,}390\ /\ [E \cdot 10^{6}\ (\text{eV})] = 1.239 \cdot 10^{-2}\ [\text{Ă}]$

Exercise 2.

What is the relativistic mass of the photon? Present its expression via quantities "*h*", "λ" and "*c*".

Solution

1) According to Einstein's law, energy equals the product of the mass of the particle and the square of light velocity, i.e., $E = m \cdot c^2$, correspondingly, an impulse of the particle equals $P = m \cdot c$, that is, $E = P / m$.

2) The relativistic mass of a photon is a mass when its velocity equals the speed of light, i.e., $v = c$. If so, the relativistic mass of the photon equals:

$$m = E/ c^2 = h \cdot v / c^2 = h \cdot (c / \lambda) / c^2 = h / (c \cdot \lambda)$$

Exercise 3.

The energy of photon $E = 1$ eV. What is the value of this photon impulse and what is its wavelength in "Å" and in "μm"?

Solution

1) According to Einstein's law, pulse $P = E/c$, because $E = mc^2$ and $P = mc$. So:

$$P = 1 \text{ (eV)} / 3 \cdot 10^8 \text{ (m/c)} = 1.6\ 10^{-19} \text{ (J)} / 3 \cdot 10^8 \text{ (m/c)} = 5.3 \cdot 10^{-27} \text{ [kg m/s]}$$

2) λ (Å) = 12,390 / hv (eV) [Å] = 12,390 / 1 (eV) = 12390 (Å) = 1,2390 (μm)

Exercise 4.

A photon and an electron both have energy $E = 1$eV. Which of them has the longer wavelength?

Solution

1) For the photon (see Example 1):

$$\lambda \text{ (Å)} = 12{,}390 \text{ (Å)} = 12{,}390 \text{ μm}$$

2) For the electron, accounting for the de Broglie postulate, because in the non-relativistic case the velocity of an electron is less than the speed of light, i.e., v << c, and its impulse $p = mv << mc$, and its kinetic energy $E_k << m \cdot c^2$. In this case

$\lambda = h\,c\,/\,(E_k + m \cdot c^2) = h \cdot c\,/\,m \cdot c^2 = h\,/\,m \cdot c$, yields:

$\lambda = h\,/\,m \cdot c = 6.61 \cdot 10^{-34}$ (J·s) / [$9.11 \cdot 10^{-31}$ (kg) · $3 \cdot 10^8$ (m/s)] = $2.42 \cdot 10^{-12}$ (m) = $2.42 \cdot 10^{-6}$ (μm)

So, 1,239 (μm) >> $2.42\ 10^{-6}$ (μm).

Conclusion: The wavelength of the photon is longer (by about one million times) than that of the electron having the same energy of 1 eV.

Exercise 5.

The limit of the photon effect is characterized by the critical wavelength λ_{cr}, after which an electron cannot leave the material, that is, to pass the prohibited zone limited by an energy E_g.

Find: The critical length for metallic Cu, if its prohibited zone has a width of $E_g = 4.3$ eV. Notice that this "energy width" is exactly equal to the outwork of light, W_{out} to transfer the electron from the valence to the conductive zone giving the electron a kinetic energy E_k.

Solution

1) Energy of the photon needed to excite the valence electron transferring it from the valence to the conductive zone and obtaining the kinetic energy E_k can be found from the following relation:

$$\underline{h\nu} = E_k + W_{out}$$

2) Since in our case, this energy is to transfer the electron only to pass by the prohibited zone (e.g., having $E_k = 0$ in the conductive zone), we get:

$$\underline{h\nu} = W_{out} = E_g = 4.3 \text{ eV}$$

If so, we can rewrite this expression, accounting for $\nu = c\,/\,\lambda_{cr}$, as:

$$\lambda_{cr} = \underline{h \cdot c / E_g} = 6.62 \cdot 10^{-34}\ (\text{J}\cdot\text{s}) \cdot 3 \cdot 10^{8}\ (\text{m/s}) / 4.3 \cdot 1.6 \cdot 10^{-19}\ (\text{J})$$
$$= 2.87 \cdot 10^{-7}\ (\text{m}) = 0.287\ (\mu\text{m}) = 287\ (\text{nm})$$

Conclusion: Photon with λ_{cr} = 287 (nm), corresponding to the violet band of the visual light spectrum (occupying bandwidth from 200 nm to 750 nm, see Chapter 1), can be excited and it can be taken out from metallic Cu only one photoelectron.

Exercise 6.

The electron moves along a horizontal axis, and its movement is limited by the length of a box $L = 10^{-10}$ m, which models the simple atom.

Find: 1) The zero energy of electron corresponding to wave function with n=1, E_1.

2) Wavelength (in Å) of the photon excited after transfer of an electron from level n' = 2 to n=1.

Solution

1) Taking into account a general formula of electron wave functions inside the atom (as a closed box) $E_n = h^2 \cdot n / 8m \cdot L^2$, we get for n=1 the zero energy of the electron in the atom:

$$E_1 = h^2 / 8 \cdot m \cdot L^2 = [6.62 \cdot 10^{-34}\ (\text{J s})]^2 / [8 \cdot 9.11 \cdot 10^{-31}\ (\text{kg}) \cdot [10^{-10}\ (\text{m})]^2$$
$$= 6.02 \cdot 10^{-16}\ (\text{J}) = 37.5\ (\text{eV})$$

2). Transfer from level n' =2 to n=1 with excitation of the photon with energy hv can be found by the well-known formula:

$$hv = E_2 - E_1 = E_1 (n'^2 - n^2) = 37.5\ (4\text{-}1) = 112,5\ (\text{eV})$$

3). Then, according to Example 1:

$$\lambda\ (\text{Å}) = 12390 / \underline{hv}\ (\text{eV}) = 12390 / 112.5\ (\text{eV}) = 110\ (\text{Å})$$

Conclusion: Photon with λ = 110 (Å) = 11 (nm), does not lie in the frequency band of visual light; it lies in the ultraviolet bandwidth.

Exercise 7.

1) Find the maximum wavelength λ of light radiation for transfer of the electron from the "ground" level of an atom of hydrogen (H) defined by n_1 = 1 to the level defined by n_2 =2.
2) Find the maximum wavelength λ for electron transfer from the level with n_2 = 2 to the level n_3 = 3.

Solution

1) $E_1 = E_{min} = 2 \cdot \pi^2 \cdot m \cdot e^4 / h^2 = 13.6$ (eV)

2) Using the now well-known Bohr's formula

$$h\nu = -13.6 [(1/n_2^2) - (1/n_1^2)] = 13.6 [(1/n_1^2) - (1/n_2^2)]$$

we get for transfer from the level (orbit) n_1 = 1 to that with n_2 = 2:

$$h\nu = 13.6 (1 - 1/4) = 10.2 \text{ (eV)}$$

Then λ(Ă) = 12390 / $h\nu$ (eV) = 12390 / 10.2 (eV) = 1210 (Ă)

3) For n_2 = 2 to n_3 = 3 transfer we get:

$$h\nu = 13.6 [(1/n_2^2) - (1/n_3^2)] = 13.6 (1/4 - 1/9) = 19 \text{ (eV)}$$

Then λ(Ă) = 12390 / $\underline{h\nu}$ (eV) = 12390 / 19 (eV) = 6400 (Ă) = 640 (nm)

Conclusion: The photon during transfer from n_2 = 2 to n_3 = 3 has a wavelength, which lies inside the light spectrum, whereas the photon for the transfer from n_1=1 to n_2 = 2 is outside the light spectrum.

Exercise 8.

What is the relation between the spectral energy of helium (He) and hydrogen (H)? It is known that the charge number of He equals Z = 2, and for H it equals Z = 1.

Solution

1) We take into account the well-known Bohr's formula

$$E_n = 2\pi^2 \cdot m \cdot e^4 \cdot Z^2 / h^2 \cdot n^2 = 13.6\ (Z^2 / n^2)$$

If so, for H we have for Z= 1 and any energy level (orbit), we get

$$E_n = 13.6 / n^2$$

and for He with Z=2, we get

$$E_n = 13.6\ (Z^2 / n^2) = 4 \cdot 13.6 / n^2$$

So, for any transfer in an atom of H, we have the formula used in Example 7:

$$h\nu = 13.6\ [(1/ n^2) - (1/n'^2)]$$

Whereas for an atom of He, we get

$$h\nu = 4 \cdot 13.6\ [(1 / n^2) - (1 / n'^2)]$$

Conclusion: The energy spectrum of He is four times bigger than that for H.

Exercise 9.

Find the line (i.e., wavelength) of the He atom spectrum (with Z=2) similar to the line of the H atom spectrum and compare them.

Solution

1) Accounting for the energy of a photon excited from the H atom (with Z=1), according to knowledge obtained from Example 7, we get:

$$\nu_H = 13.6\ [(1 / n^2) - (1 / n'^2)] / h$$

from which

$$\lambda_H = 12390 / (h \cdot \nu_H)$$

2) Accounting for the energy of a photon excited from the He atom (with Z=2), according to knowledge obtained from Example 7, we get:

$$\nu_{He} = 4 \cdot 13.6 \cdot [(1 / n2) - (1 / n'^2)] / h$$

From which follows: $\nu_{He} = 4\nu_{H}$

Then

$$\lambda_{He} = 12390 / h \cdot \nu_{He} = 12390 / 4h \cdot \nu_H = (1/4) \cdot \lambda_H$$

Finally, we get: $\lambda_{He} = (1/4)\, \lambda_H$

3) Now, we introduce in the above formulas the corresponding quantities according to Example 6 for H for $n = 1$ and $n' = 2$ we get:

$$h\nu_H = 13.6 \cdot [(1 / n^2) - (1 / n'^2)] = 10.2 \text{ (eV)}$$

Then

$$\lambda_H = 12390 / (h \cdot \nu_H) = 12390 / 10.2 = 1216 \text{ (Å)}$$

At the same time for the He atom for the same transfer from $n = 1$ to $n' = 2$, accounting for the relations obtained above, we get:

$$h\nu_{He} = 13.6 \cdot [(1/ n^2) - (1/ n'^2)] = 4 \cdot 10.2 \text{ (eV)} = 40.8 \text{ (eV)}$$

Then

$$\lambda_{He} = (1/4)\, \lambda_H = 1216 / 4 = 304 \text{ (Å)}$$

Conclusion: for the same transfer of an electron from the ground level with n=1 to the level with n=2 for H and for He, we have found that for this transfer the excited photon frequency for H is four times bigger than that for He, whereas the corresponding line (wavelength) for He is four times less than for H.

Exercise 10.

How many electrons are in the atom sub-layers with the main quantum number n=2 and n=3?

Solution

1) For $n = 2$, the quantum number l, equaling n-1, is: $l = 0$ and $l = 1$.

for $l = 0$, we get $m_l = 0$ – *one orbit* (state or level)
for l=1, we get $m_l = 0, \pm 1$ – *three orbits* (states or levels)

In total we get *four states* timing on *two electrons* (according to the Pauli postulate), we finally get *8 electrons.*

2) For n=3, l = 0, 1, 2.

for l = 0, we get m_l = 0 – *one orbit* (state or level)
for l=1, we get m_l = 0, ±1 – *three orbits* (states or levels)
for l=2, we get m_l = 0, ±1, ±2 – *five orbits* (states or levels)

In total we get *nine states* timing on *two electrons* (according to the Pauli postulate), we finally get *18 electrons*.

Exercise 11.

How many electrons are in the valence sub-zone (subshell) 6*d*, and sub-zone (subshell) 6*f*?

Solution

1) Subshell 6*d* corresponds to n=6 and l=2, for which we have m_l = 0, ±1, ±2 for orbital quantum number. Finally, we have 5 oriented orbits or according to Pauli's prohibited rule – maximum of 2 electrons. So, finally subshell 6*d* has *10 electrons.*

2) Subshell 6*f* corresponds to n=6 and l=3, for which we have m_l = 0, ±1, ±2, ±3 for orbital quantum number. Finally, we have 7 oriented orbits, or according to Pauli's prohibited rule – a maximum of 2 electrons. So, finally sub-zone 6*f* has *14 electrons.*

Bibliography

[1] Jenkins, F. A., and H. E. White. 1953. *Fundamentals of Optics*. New York: McGraw-Hill.
[2] Born, M., and E. Wolf. 1964. *Principles in Optics*. New York: Pergamon Press.
[3] Charles, A. W., and R. M. Thomson 1964. *Physics of Solids*. New York: McCraw-Hill Book Co.
[4] Fain, V. N., Ya. N. Hanin. 1965. *Quantum Radiophysics*. Moscow: Sov. Radio (in Russian).
[5] Lindner, H. 1975. *Das Bild der Modern Physik* (*View on the Modern Physics*). Berlin: Urania – Verlag, Germany (in German).
[6] Kireev, P. S. 1975. *Physics of Semiconductors*. Moscow: High School Publisher (in Russian).
[7] Lipson, S. G., and H. Lipson. 1969. *Optical Physics*. Cambridge: University Press.
[8] Akhamov, S. A., R. V. Khohlov, and A. P. Sukhorukov. 1972. *Laser Handbook*. North Holland: Elsevier.
[9] Marcuse, O. 1972. *Light Transmission Optics*. New York: Van Nostrand-Reinhold Publisher.
[10] Fowles, G. R. 1975. *Introduction in Modern Optics*. New York: Holt, Rinehart, and Winston Publishers.
[11] Yariv, A. 1976. *Introduction in Optical Electronics*, Chapter 5. New York: Holt, Rinehart, and Winston.
[12] Hecht, E. 1987. *Optics*. MA: Addison-Wesley, Reading.
[13] J. Dakin and B. Culshaw, eds. 1988. *Optical Fiber Sensors: Principles and Components*. Artech House, Boston-London.
[14] Kopeika, N. S. 1998. *A System Engineering Approach to Imaging*. Washington: SPIE Optical Engineering Press.
[15] Bansal, R., ed. 2006. Handbook: *Engineering Electromagnetics Applications*. New York: Taylor and Frances.

CHAPTER 4

BASIC PRINCIPLES OF PHOTONICS AND LASER OPERATION

4.1. Boltzmann Distribution

As follows from the discussions in Chapter 3 based on wave-corpuscular dualism, in any gas of atoms, or any materials consisting of atoms, each atom can obtain its allowed energy level from a set of E_1, E_2,…, E_m, as shown by Fig. 4.1 (left panel). In thermal equilibrium at temperature T their motions reach their steady-state regime with the probability $P(E_m)$ that the arbitrary atom is in the energy level E_m given by the Boltzmann distribution:

$$P(E_m) \sim \exp\{-E_m / k_B T\} \tag{4.1}$$

where k_B is the Boltzmann constant equal to $k_B = 1.38 \cdot 10^{-23} J \cdot K^{-1}$. The coefficient of proportionality is chosen such that the total cumulative probability equals unit. The occupation probability is an exponential function as displayed in Fig. 4.1 (right panel).

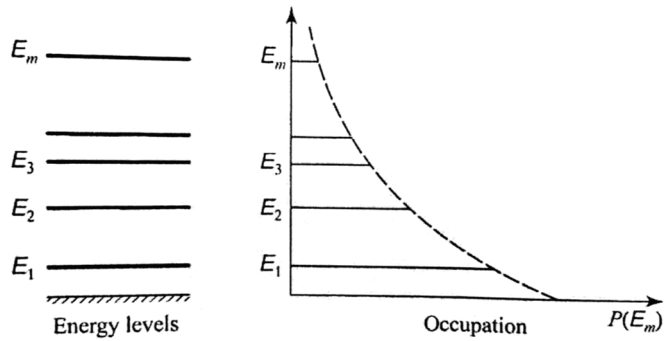


Figure 4.1. Discrete energy levels of atoms (left panel) and the probability of their energy distribution (dashed curve in the right panel).

Considering the Boltzmann distribution for a large number of atoms N, and assuming that an arbitrary number of atoms N_m accompany energy level E_m, then the fraction N_m/N is proportional to $P(E_m)$. If N_1 atoms occupy level 1 and N_2 atoms occupy a higher level 2, the population ratio is, on average:

$$N_2/N_1 \sim \exp\{-(E_2-E_1)/(k_B T)\} \tag{4.2}$$

This quantity depends on temperature. For $T = 0$ K = -273 °C, all atoms are at the lowest levels (called the *ground state*). With an increase in temperature, a higher energy level can have a greater population than a lower energy level. This non-equilibrium case is known as a *population inversion*, providing the laser actions discussed further in Chapter 5. The same approach can also be taken for electrons filling metals, semiconductors, or dielectrics. Thus, according to wave-corpuscular dualism discussed in the previous chapter, in metal electrons filled the lowest energetic levels (states) to create the so-called *Electron-Fermi gas*. This peculiarity was mentioned by Enrico Fermi, according to which all energetic states – from the lowest to a state with kinetic energy $(E_k)_0$, called the *Fermi boundary*, are filled by N electrons, every two of which, according to Pauli's principle (see Chapter 3), fill each quantum level (state). This energy equals:

$$(E_k)_0 = h^2/(8m_e)\cdot(3N/\pi)^{2/3} \tag{4.3}$$

where again $h = 6.625\cdot10^{-34}$ $J\cdot$s is the Planck constant, $N = N / V$ – number of free electrons in 1 cm^2. This result does not depend on the shape of the volume of metal V, which was presented as a box with a length L (see the previous chapter), but only on the density of free electrons in the metal. On this boundary, the energy spectrum decreases sharply (see Fig. 4.2 below). As will be shown below, a Fermi spectrum of energy distribution differs from Boltzmann or Maxwell's statistics in gases.

4.2. Fermi-Dirac Energy Distribution

According to quantum theory, briefly discussed in Chapter 3, each discrete system with overlapping wave functions, such as a multi-electron atom, metal, or semiconductor, is subject to the *Pauli Prohibiting Principle*. According to this principle, each energy level can be occupied by not more than 2 electrons with opposite spins, but most of them are occupied by at least one electron. If so, the number of electrons N_m in state m can be either 0 or 1.

We notice that the word "spin" was introduced by George Uhlenbeck and Samuel Goudsmit from the USA during their investigations of atomic quantum structure. For a wave mechanical view, the concept of "spin" was introduced by Paul Dirac investigating not only electrons, but also other elementary particles, such as protons, having the same spin as electrons.

The probability of occupancy of a state of energy E can be described by the *Fermi-Dirac distribution* (called also *Fermi function*):

$$f(E) = \{\exp[(E-E_f) / (k_B T)] + 1\}^{-1} \tag{4.4}$$

Here E_f is the Fermi energy introduced in Chapter 3 as a boundary energy of an inner electron to leave any solid crystal or multi-electron atom. $f(E) = 1$ indicates that the state of energy E is definitely occupied. As shown in Fig. 4.2, it lies along the horizontal axis between 0 and 1. This function decreases monotonically with increasing E and equals ½ when Fermi energy equals E_f.

We should notice that $f(E)$ is neither a probability density function nor a probability distribution function, but rather a distribution of probabilities for different values of E, each of which lies between 0 and 1.

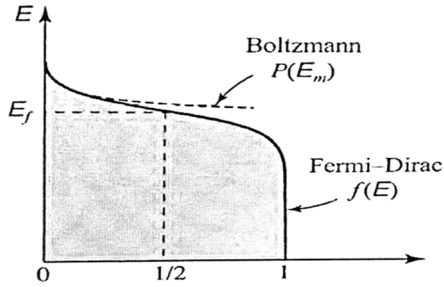


Figure 4.2. Comparison of Fermi function (continuous curve) and Boltzmann probability distribution (dashed curve).

When $E >> E_f$ and $E >> k_B T$ the Fermi function behaves like the Boltzmann probability distribution $P(E_m) \sim \exp(E_m / k_B T)$, since in general for atomic electrons in outer subshells, energy levels involving optical transitions are often characterized by Boltzmann distribution.

It should also be noted from the beginning that Fermi wanted to use for electron gas the Bose distribution, which was usually used for photons, as quanta of light (called *bosons*), with their spin momentum $+1(h/2\pi)$ and $-1(h/2\pi)$. But then, according to Pauli's law, Fermi used his own

statistics for electrons with their spins +1/2(h/2π) and -1/2(h/2π). Therefore, particles, including electrons, which follow the concept of Pauli and Fermi statistics (according to Fig. 4.2), are called *fermions* in the literature [1–7]. Hence, following the Fermi distribution for electrons, as a gas, and according to Pauli's prohibiting law, only two electrons with different spins can fill each discrete energy level.

4.3. Interaction of Photons with Atoms

4.3.1 Thermal Emission – Spontaneous, Stimulated, and Absorption of Photons

To characterize the interaction of any atom and radiated light photons, the so-called *lineshape function* and the *transition cross section* are usually introduced in photonics [1–7, 9–11].

The transition cross section, $\sigma(v)$, can be determined via its area S as an integral of $\sigma(v)$ over the spectrum of frequencies v and has dimensions cm^2 / Hz. Usually, it is called the *transition strength* or *oscillator strength* and presents a strength of interaction. Additionally in the literature, a normalized function $g(v) = \sigma(v) / S$ is introduced and called the *lineshape* function (or *profile* function), which has dimensions Hz^{-1} and unity area (e.g., integral of $g(v)dv = 1$). The transition cross section can be written in terms of its strength and profile function as

$$\sigma(v) = S \cdot g(v) \tag{4.5}$$

The *profile function* $g(v)$ is centered around the resonance frequency v_0, where $\sigma(v)$ is maximal (see Fig. 4.3). So, the transition for photons occurs at $v = v_0$,

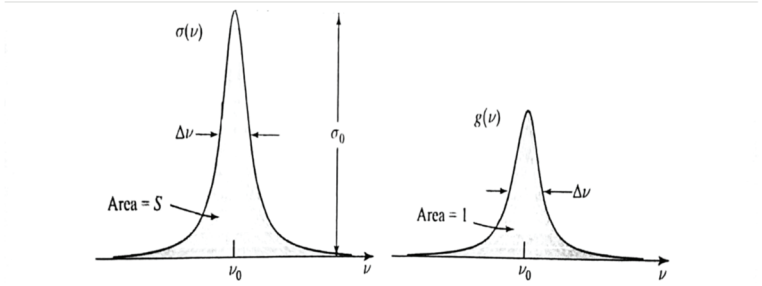


Figure 4.3. The transition function (left-side) and the lineshape (profile) function (right-side).

The width of $g(\nu)$ is known as the *transition linewidth,* usually determined as a width $\Delta\nu$, on which $g(\nu)$ is half its maximum value. Since the area of $g(\nu)$ is unity, its width is inversely proportional to its central value, i.e.:

$$\Delta\nu \sim 1/g(\nu_0) \tag{4.6}$$

It is useful to define a peak transition cross section at the resonance frequency $\sigma_0 = \sigma(\nu_0)$. The function $\sigma(\nu_0)$ is fully characterized by its height σ_0, width $\Delta\nu$, area S, and profile $g(\nu)$, as clearly illustrated in Figure 4.3.

Spontaneous Emission. According to Refs. [1–8], we define t_{sp} as a *spontaneous lifetime* of transition from mode 2 to mode 1, as

$$PDF_{sp} = M(\nu_0)\cdot c\cdot <S> \sim 1/t_{sp} \tag{4.7}$$

where $M(\nu_0) = 8\pi\cdot\nu_0^2/c^3$ [s/m^3] is the modal density through which one can transform PDF_{sp} of spontaneous emission over all modes into each mode using the weighted modal density $M(\nu)$. From Eq. (4.7) it follows that PDF_{sp} is a non-dimensional function. Because the shape of the average (over a full spectrum of frequencies) cross section $<\sigma(\nu)>$ is narrow, but $M(\nu)$ is wide and constant around $M(\nu_0)$, we can simplify general formulas for PDF_{sp} obtained in Refs. [1–7] and present it by Eq. (4.7). This equation determines spontaneous emission of one photon to any mode, which is independent of the cavity volume. This gives us the possibility to find the average area of a 2-D cavity, <S>, consisting of electrons, as

$$<S> = \lambda^2 / 8\pi\cdot t_{sp} \tag{4.8}$$

The transition strength can be determined from experimental measurements of the spontaneous lifetime. Thus, for atomic hydrogen $t_{sp} = 10^{-8}$s for atomic transition from the first exciting state.

Now we can find the relations between these two specific functions, $<\sigma(\nu)>$ and $g(\nu)$. Using Eq. (4.8), and relation $<\sigma(\nu)> = g(\nu)\cdot <S>$, we finally have the relation between the spontaneous transition function and the profile shape function, which is called the *average transition cross section*:

$$<\sigma(\nu)> = \lambda^2\cdot g(\nu) / 8\pi\cdot t_{sp} \tag{4.9}$$

The same characteristic, but for central frequency will equal:

$$<\sigma_0> = <\sigma(\nu_0)> = \lambda^2\cdot g(\nu_0) / 8\pi\cdot t_{sp} \tag{4.10}$$

So, because $g(\nu_0)$ is inversely proportional to linewidth $\Delta\nu$ (according to Eq. (4.6)), $\langle\sigma(\nu_0)\rangle$ is also inversely proportional to the $\Delta\nu$ for a given t_{sp}.

Stimulated Emission and Absorption. If, due to outer radiation, the atom is transferred from the ground (lower) energy level to the high energy level and the corresponding mode (according to the *wave-corpuscular dualism*) contains the photon, the atom can be induced to emit another photon into the same mode. Such emission of photons is called *stimulated emission*.

The PDF of an emission taking place from t to $t+\Delta t$, depends on frequency ν and on transition cross section $\sigma(\nu)$ centered on the atomic resonance frequency $\nu = \nu_0$:

$$PDF_{st} = c \cdot \sigma(\nu) / V \tag{4.11}$$

If there are n photons in the light mode, the PDF that the atom is stimulated to emit an additional photon, as in a case of absorption, equals:

$$PDF_{st} = c \cdot n \cdot \sigma(\nu) / V \tag{4.12}$$

There is a possibility to present simply spontaneous and stimulated emission (see Fig. 4.4).

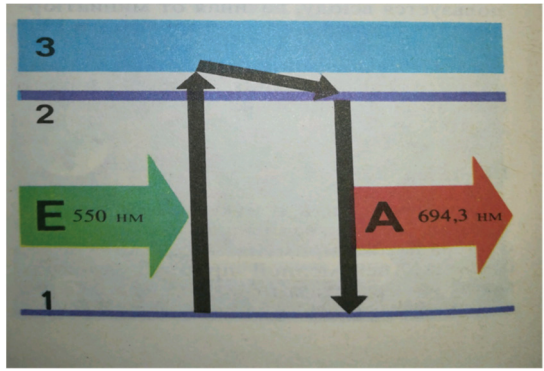


Figure 4.4. Sketched simple presentation of the stimulated (from level 1 to level 3) and the spontaneous (from level 2 to level 1) emission.

As is clearly seen, the atom, after interaction with the photon, absorbs it and jumps to the higher level of energy E_3, which corresponds to the photon with frequency ν_{31} (that corresponds to a wavelength $\lambda_{31} = c / \nu_{31}$ = 550 nm, or green light). Due to the instability of the level with energy, the

atom falls into the unstable level with energy E_2 (called *metastable* level), and after spontaneous emission it falls to the ground level with energy E_1. Such a transaction corresponds to frequency v_{21} which corresponds to the wavelength of λ_{21} = 694 nm, or red light. This process takes a longer time and is called a slow process of *spontaneous emission*. Such double-cascade transitions can be stimulated by photons with energy hv_{21}, as shown in Figure 4.4.

Let us describe now this process mathematically accounting for two kinds of transition: a) stimulated by monochromatic (*single-mode*) light, and b) stimulated by *broadband light*.

a) We consider a single-mode light and its interaction with the atom when a stream of photons impinges on it. Let monochromatic light of frequency v and intensity I and the mean photon flux density

$$\phi = I / hv \text{ [photon/cm}^2\text{]} \tag{4.13}$$

interact with the atom whose resonant frequency is v_0.

We also suppose that the probabilities of stimulated emission and absorption are similar, that is, $PDF_{st} = PDF_{abs}$ in such a consideration. If the interacting volume has the form of a cylinder with volume V, height h, and base area A, and assuming that n photons are involved in the interactional process, we get $V=h\cdot A$. A flux of photons crossing area A is $\Phi=A\cdot\phi$ [photons/s]. Because photons move with the speed of light c, all of them cross the base of the cylinder within one second. If so, in any time the cylinder contains n photons, where

$$n = \phi \cdot A = \phi \cdot V / d \tag{4.14}$$

which yields

$$\phi = n \cdot d / V \tag{4.15}$$

Absorption of the photon can be viewed, according to Figure 4.4, as the transition of the atom from the lower energy level E_1 to the higher energy level E_3. This process is induced by a photon with the probability density function (*PDF*):

$$PDF_{abs} = c \cdot \sigma(v) / V \tag{4.16}$$

If there are n photons in the light mode, the PDF that the atom absorbs one photon is n-times greater since the events are mutually exclusive, i.e.,

$$PDF_{abs} = c \cdot n \cdot \sigma(\nu) / V \tag{4.17}$$

Accounting now for Eq. (4.15) and Eq. (4.17), yields

$$PDF_{abs} = PDF_{st} = \phi \cdot \sigma(\nu) \tag{4.18}$$

Formula (4.18) determines the photon flux captured by the atom for the purpose of stimulated emission or absorption.

b) In the case of stimulation by broadband light, let us consider an atom in a cavity of volume V containing multimode polychromatic light of spectral energy density $\varkappa(\nu)$ (energy per unit bandwidth per unit volume), which is much broader than the linewidth. The average number of photons in the frequency range of $[\nu, \nu+d\nu]$ is $\varkappa(\nu) \cdot V d\nu / h\nu$.

So, the overall probability of absorption or stimulated emission can be found via the integral, which accounting for a slow varied $\varkappa(\nu)$ with respect to a sharp $\sigma(\nu)$, can be simplified as:

$$PDF_{abs} = PDF_{st} = \varkappa(\nu_0) \cdot d \cdot \langle S \rangle / h\nu_0 \tag{4.19}$$

or accounting Eq. (4.8), we get:

$$PDF_{abs} = PDF_{st} = \lambda^3 \cdot \varkappa(\nu_0) / 8\pi \cdot h \cdot t_{sp} \tag{4.20}$$

So, because $g(\nu_0)$ is inversely proportional to linewidth $\Delta\nu$, according to Eq. (4.6), $\langle \sigma(\nu_0) \rangle$ is inversely proportional to the $\Delta\nu$ for a given t_{sp}. Accounting now for relations between the wavelength and the central photon frequency, $\lambda = c/\nu_0$ and the mean number of photons per mode [3–8], $\langle n \rangle = \lambda^3 \cdot \varkappa(\nu_0) / 8\pi \cdot h$, we get:

$$PDF_{abs} = PDF_{st} = \langle n \rangle / t_{sp} \tag{4.21}$$

The mean photon number $\langle n \rangle$ has physical meaning. Indeed, the quantity $\varkappa(\nu_0)/h\nu_0$ represents the mean number of photons per unit volume in the vicinity of frequency ν_0 and $M(\nu_0)$ is the number of modes per unit volume $\Delta\nu_0$ in the frequency domain. The two PDFs, for stimulated emission and absorption, are thus the factor of the event when the mean photon number, $\langle n \rangle$, is greater than that for spontaneous emission, since each mode contains

an average of $<n>$ photons.

In most literature related to photonics descriptions, useful parameters are usually introduced to describe the processes of spontaneous and stimulated emission and absorption. These coefficients are called *Einstein's coefficients*. It was postulated by Einstein as follows:

The atom interacts with broadband radiation of spectral energy density $\varkappa(\nu_0)$ *under conditions of thermal equilibrium*.

According to this postulate, an expression for the probability densities of spontaneous and stimulated transitions was evaluated by introducing the so-called Einstein's coefficients:

$$PDF_{sp} = A, \quad PDF_{st} = B \cdot \varkappa(\nu_0) \tag{4.22}$$

which are associated with spontaneous and stimulated transitions (or absorption, see above), respectively. Their ratio yields:

$$B / A = \lambda^3 / (8\pi \cdot h) \tag{4.23}$$

4.3.2 Thermal Equilibrium Between Atoms and Photons

Despite the fact that in Chapter 3 we briefly described conditions of thermal equilibrium to explain the Max Planck law regarding *black body* light absorption, let us return to this subject from the mathematical point of view and describe this phenomenon occurring during interactions of photons with atoms, considering a thermal light, as a universal form of radiation under conditions of thermal equilibrium in the absence of outer energy sources. Such light, as was mentioned in the previous chapter, is emitted by *black bodies* that absorb all light energy incident on them.

A macroscopic approach that balances spontaneous emission, stimulated emission, and absorption under conditions of *thermal equilibrium* leads to the spectral intensity of thermal light. Equation (4.7) or (4.8) is a point of our analysis. We consider a cavity with unit volume whose walls have a large number of atoms with two energy levels E_1 and E_2, separated by energy $h\nu$. Levels 1 and 2 consist of $N_1(t)$ and $N_2(t)$ atoms, respectively.

We first consider the *spontaneous emission* alone. The probability a single atom in the upper-level 2 undergoes spontaneous emission into any of the modes within the time duration from t to $t + dt$, is $P_{sp}dt = dt / t_{sp}$. So, the average number of photons within dt is $n_2 = N_2(t) \cdot dt / t_{sp}$. Hence, the negative rate of increase of $N_2(t)$ arising from spontaneous emission can be found from the differential equation:

$$dN_2(t)/dt = - N_2/t_{sp} \tag{4.24}$$

The solution of Eq. (4.24) can be easily found as

$$N_2(t)=N_2(0)\cdot\exp(-t/t_{sp}) \tag{4.25}$$

which is presented in Figure 4.5, where for $t = t_{sp}$, the decay of the upper-level population, $N_2(0)$, caused by spontaneously emitted photons, is up to e^{-1}. So, this process takes time around $t = t_{sp}$.

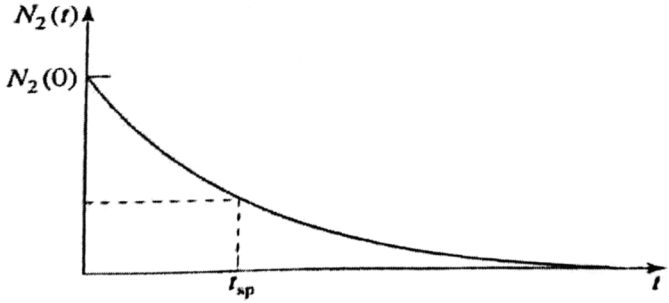


Figure 4.5. Decay of the upper-level population N_2 (t=0) by e^{-1} factor for $t = t_{sp}$ according to time dependence described by Eq. (4.25).

If we now also incorporate the *absorption* and account for the capability of N_1 atoms to absorb, we will get the rate of increase of the population of the atoms at the upper energy level, arising from absorption, which can be found by the use of (4.21) as:

$$dN_2(t)/dt = PDF_{sp}\cdot N_1 = \langle n\rangle\cdot N_1 / t_{sp} \tag{4.26}$$

Similarly, *stimulated emission* gives rise to a negative rate of increase of atoms in the upper state 2, expressed as

$$dN_2(t)/dt = - PDF_{st}\cdot N_2 = - \langle n\rangle\cdot N_2 / t_{sp} \tag{4.27}$$

As is clearly seen from Eq. (4.26) and Eq. (4.27), the rate of $N_2(t)$ arising from absorption and stimulated emission are proportional to $\langle n\rangle$. All three processes, described by Eqs. (4.25) to (4.27), yield the final equation:

$$dN_2(t)/dt = - N_2/t_{sp} + \langle n\rangle\cdot N_1 / t_{sp} - \langle n\rangle\cdot N_2 / t_{sp} \tag{4.28}$$

In the absence of any outer source of light radiation, the steady-state regime, when $dN_2(t)/dt = 0$, gives:

$$N_2/N_1 = <n>/(<n> +1) \tag{4.29}$$

From relation (4.29) follows that $N_2 < N_1$.

According to a thermal equilibrium condition, we can, as above, use the Boltzmann's distribution (4.2), rewriting it as:

$$N_2/N_1 = \exp\{-(h\nu)/(k_B T)\} \tag{4.30}$$

Substituting Eq. (4.30) in Eq. (4.29), one can find the average number of photons per modes near frequency *ν:*

$$<n> = 1/[\exp(h\nu)/(k_B T)-1] \tag{4.31}$$

Equation (4.31) allows us to find a mean number of photons in a mode of thermal light for which the occupation of modal energy level follows the photon distribution

$$PDF(n) \sim \exp[-(h\nu)/(k_B T)] \tag{4.32}$$

This relation argues the self-consistency of the analysis carried out above and its correction for the description of any real situation in a solid cavity consisting at the walls of many atoms of arbitrary mode-states, existing in thermal equilibrium. Photons interacting with atoms in thermal equilibrium of temperature *T*, are also in thermal equilibrium at the same temperature. Therefore, we can call them a *photon gas*.

Accounting that the average energy of photon gas is $<E> = <n>\cdot(h\nu)$ and accounting for Eq. (4.32) yields:

$$<E> = h\nu/[\exp(h\nu)/(k_B T) - 1] \tag{4.33}$$

Multiplying the average energy per mode, $<E>$, by the modal density, $M(\nu)$, defined as $M(\nu_0) = 8\pi\cdot\nu^2/c^3$, we can obtain the spectral energy density $\rho(\nu) = M(\nu)\cdot<E>$, measured in energy (in Joules) per unit frequency bandwidth (in Hz) per unit cavity volume (in m^{-3}), can be presented in the following form:

$$\rho(\nu) = [8\pi\cdot h\nu^3]/c^3\cdot[\exp(h\nu)/(k_B T) - 1] \tag{4.34}$$

The spectral energy density function described by Eq. (4.34) is known as the *black body radiation spectrum*. It is a function of the frequency of light, mode of frequency v and temperature T. Below we present it as a function of frequency and temperature following the Max Planck validation of Eq. (4.33).

As was mentioned in Chapter 3, Max Planck in 1900 had found the theoretical proof of formula (4.33), which was agreed with by experiments. His calculation led to the expression for the black body spectrum via <E> by quantizing the energy of each mode. At the same time, as it is known from classical physics and statistical mechanics, the average energy per one mode $<E> = k_B T = Constant$ and independent of the modal frequency. Such a formulation was postulated by *Rayleigh–Jeans*, which for black body radiation gives:

$$\rho(v) = 8\pi \cdot v^2 \cdot k_B \cdot T / c^3 \tag{4.35}$$

At the same time, from Eqs. (4.33) and (4.34), obtained also by Max Planck, it follows that for $hv << k_B T$, when $\exp(hv/ k_B T) \sim 1 + (hv) / (k_B T)$, these formulas are deduced to classical, $<E> = k_B T$, [from Eq. (4.33)], and to Eq. (4.35) from Eq. (4.34).

4.4. Electron and Hole Concentration in Semiconducting Materials

According to the discussion above and in Chapter 3, the quantum theory postulates the state of an electron and a hole in any semiconductor through their energy E, their vector **k**, and their spin s. Their concentration, as a function of energy E requires knowledge of two features:

1) the density of states or energy layers in each semiconductor,
2) the probability that some of these levels are occupied.

As for an electron inside the conductive zone, it can be approximately described by its effective mass m_e, hidden into a 3-D box of dimension d with perfectly reflecting walls, with infinite rectangular potential inside, as was discussed in Chapter 3. The standing-wave solutions require the discrete components of the wave vector **k** with coordinates

$$\mathbf{k} = \{k_x = q_1 \cdot \pi/d,\ k_y = q_2 \cdot \pi/d,\ k_z = q_3 \cdot \pi/d\} \tag{4.36}$$

and with the respective mode positive numbers (q_1, q_2, q_3). The tip of the vector **k** can lie at the point of a lattice whose unit cell has dimension π/d and volume $(\pi/d)^3$.

If so, there are $(\pi/d)^3$ points per unit volume in k-space. The number of states for which k lies between 0 and k is determined by the number of points lying within the positive octant of a sphere of radius k with volume $V = (1/8)\cdot(4\pi/3)\cdot k^3 = (\pi k^3) / 6$. According to the Pauli statement, there are two states of electrons that are possible at each state (i.e., orbit) with the corresponding spins $s = -1/2$ and $s = +1/2$. If so, we finally get a number of points in volume $V=2\cdot[(\pi k^3) / 6]/(\pi/d)^3 = k^3\cdot d^3/3\pi^2$.

So, in a unit volume of a cube we have a number of $V/d^3 = k^3/3\pi^2$ points. Finally, the number of states with electron wavenumber ranged between k and $k+dk$ per unit volume is:

$$\rho(k)\cdot dk = [d(k^3/3\pi^2)/dk]\cdot dk = (k^2/\pi^2)\cdot dk \tag{4.37}$$

and the density of states is:

$$\rho(k) = k^2/\pi^2 \tag{4.38}$$

The results above allow us to point out that despite the fact that the result can be obtained from classical electrodynamics for an electromagnetic resonator, with ν-k relation $\nu=c\cdot k /2\pi$, in semiconductor physics the allowed solutions for k are converted to discrete levels of energy via quadratic E-k relations, given by Eq. (3.13) in Chapter 3, which are near the conduction and valence bands, respectively.

If we now go to energy presentation, accounting for relations between energy E and impulse p as a function of k, we can represent now a number of conduction band energy levels per unit volume via the octave of a spherical surface shown in Figure 4.6.

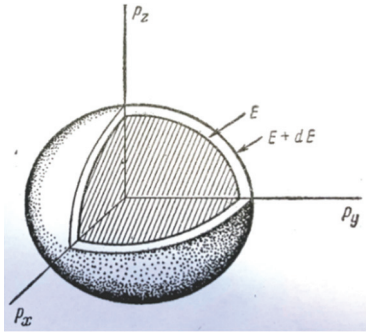


Figure 4.6. The octave of spherical surface where distribution of energy is lying between E and $E + dE$ in the coordinate system $\{\rho_x, \rho_y, \rho_z\}$.

If $\rho_c(E)dE$ represents the number of conduction band energy levels per unit volume lying between E and $E+dE$, then, because the direct correspondence between E and k exists, the densities $\rho_c(E)$ and $\rho(k)$ are related as $\rho(k)dk = \rho_c(E)dE$. So $\rho_c(E) = \rho(k)dk/dE$. The same will be found in the valence zone with bandgap energy levels, and $\rho_v(E) = \rho(k)dk/dE$. Taking now into account that mentioned above, we get for densities of states, respectively:

$$\rho_c(E) = (2m_c)^{3/2} \cdot (E-E_c)^{1/2} / [2(\pi^2 \cdot h^3)], \qquad E > E_c \tag{4.39}$$

$$\rho_v(E) = (2m_v)^{3/2} \cdot (E-E_v)^{1/2} / [2(\pi^2 \cdot h^3)], \qquad E < E_v \tag{4.40}$$

The relations $\rho_c(E)$ and $\rho_v(E)$ versus $(E-E_c)^{1/2}$ and $(E-E_v)^{1/2}$, respectively, the result of the quadratic E-k formulas (3.14) and (3.15) [see Chapter 3] for electrons in the conduction zone and holes in the valence zone near the band edges, respectively, as seen from Figure 4.7a.

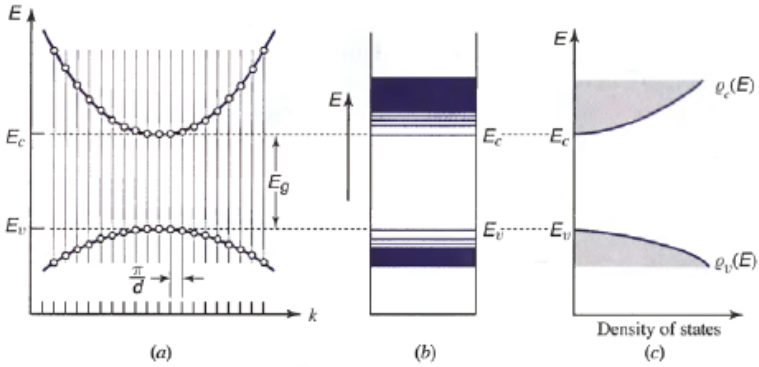


Figure 4.7. The plot of: a) the E-k quadratic dependence, b) the energy levels discrete distribution, and c) the $\rho(E) - E$ dependence for electrons and holes in conductive and valence zones, respectively.

As follows from Figure 4.7a, the E-k quadratic dependence is zero at the band edge and increases away from it with a rate which depends on the effective masses of electrons and holes. Figure 4.7b presents energy levels at the range of wavenumbers k. Figure 4.7c presents densities of states of electrons in the conduction energy band and holes in the valence energy band according to (4.39) and (4.40), where m_c and m_v are average numbers presented above.

The probability of occupancy of states in the valence zone by holes and in the conductive zone by electrons can be found by use of the following assumptions. As was mentioned in Chapter 3, at T = -273 °C = 0 K (or absence of outer sources of thermal excitation), the energy levels inside the valence zone are fully occupied and there are no holes, while the conduction band is completely empty (it contains no free electrons). With increase of temperature, excitations raise some electrons from the valence zone to the conduction zone, creating empty states in the valence zone, or holes. *Fermi function*, defined in Chapter 3, following the principle of statistical mechanics, will determine the probability that in conditions of thermal equilibrium with temperature T, an electron occupied a state with energy E,

$$f(E) = \{\exp [(E-E_f) / (hv) / (k_B T)] +1\}^{-1} \quad (4.41)$$

At T = 300 K, $k_B T$ = 0.026 eV, k_B is the Boltzmann's constant, introduced in the previous chapter. A new energy characteristic, E_f, is the Fermi energy or Fermi level. The probability function $f(E)$, is also called the Fermi-Dirac distribution.

According to this distribution, each energy level E is either occupied with probability $f(E)$ or is empty with probability 1- $f(E)$. In other words:

$f(E)$ is a probability of occupancy by an electron in a conductive band, and 1- $f(E)$ is a probability of occupancy by a hole in a valence band.

For T=0 K, $f(E)$ =1 for $E < E_f$, and $f(E)$ = 0 for $E > E_f$ (as follows from Fig. 4.8, right-side panel). For T>0 K, $(E-E_f) >> k_B T$, $f(E) \sim \exp [(E-E_f) / k_B T]$ (plotted in the middle panel). So, the high-energy tail of the Fermi function in the conduction band (see coincidence of the left side and middle panels in Fig. 4.8) decreases exponentially with increasing energy E.

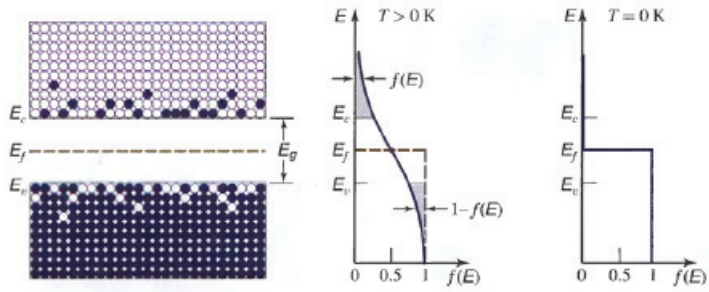


Figure 4.8. Schematic distribution of electrons (dark spheres) and holes (white squares) with Fermi energy at the middle of the prohibited zone (left side panel), distribution of energy of electrons and holes for T > 0 K (middle panel), and energy of electrons and holes for T > 0 K.

The Fermi distribution is thus proportional to the Boltzmann distribution. According to symmetry, when $E < E_f$, and $(E\text{-}E_f) << k_BT$, we get $1\text{-}f(E) \sim \exp[(E\text{-}Ef) / k_BT]$, and below the Fermi level the probability of occupancy by holes in the valence band also decreases exponentially (see coincidence of the left side and middle panels in Fig. 4.8).

Let us now consider the concentration of carriers, electrons, and holes in the thermal equilibrium statement. For this purpose, we consider that $n(E)\cdot\Delta E$ and $p(E)\cdot\Delta E$ are the numbers of electrons and holes per unit volume, respectively, with energy lying between E and $E + \Delta E$. The densities $n(E)$ and $p(E)$ can be obtained via densities of states at the energy level E and probabilities of occupancy of the level by electrons and holes, i.e.,

$$n(E)= \rho_c(E)\cdot f(E) \tag{4.42a}$$

$$p(E)= \rho_v(E)\cdot[1\text{-}f(E)] \tag{4.42b}$$

The full concentration can be found via integrals of expressions (4.42a) and (4.42b). In this case, the Fermi energy (for $n=p$) lies in the middle of the bandgap (see left side panel in Fig. 4.9). Moreover, in materials with $m_c = m_v$, $n(E)$ and $p(E)$ are symmetric (see the right-side panel in Fig. 4.9).

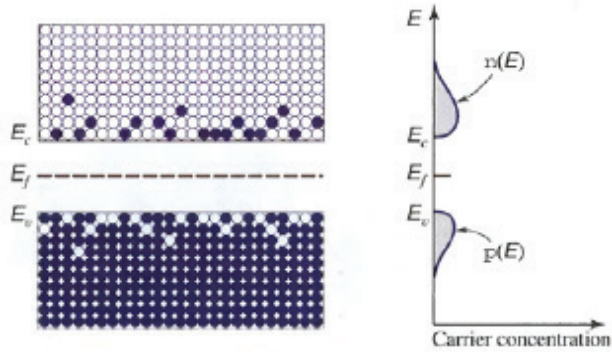


Figure 4.9. Left panel presents distribution of electrons (black spheres) and holes (white squares) in the conductive and valence zones, respectively; right panel presents symmetrical distribution of electron and hole energy via their concentration.

In most pure semiconductors, however, the Fermi level is not at the middle of the bandgap. Thus, the energy band diagrams, Fermi function, and equilibrium concentration of electrons and holes for *n*-type and *p*-type semiconductors are shown in Figs. 4.10a and 4.10b, respectively. Donor electrons occupy an energy E_D slightly below the conduction band edge (see Fig. 4.10a). For k_BT = 0.026 eV (T = 300 K) and E_D =0.01 eV, most donor electrons will be excited into the conduction band. We see increase of $n(E)$ compared to $p(E)$. The Fermi level will be above the middle of the bandgap. For the *p*-type semiconductor with acceptor level energy E_A slightly above the valence zone, (see Fig, 4.10b), conversely, $p(E) > n(E)$.

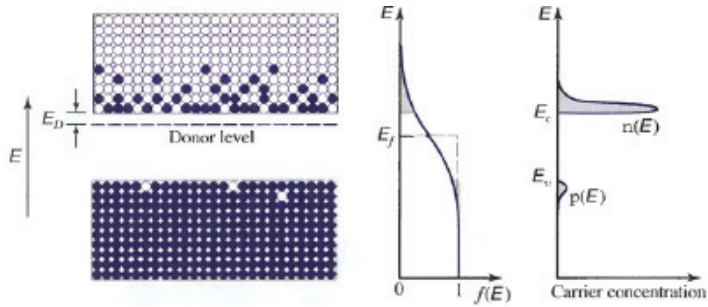


Figure 4.10a. Sketched scenario, as in Fig. 4.9, but for *n*-type semiconductor when $n(E) > p(E)$.

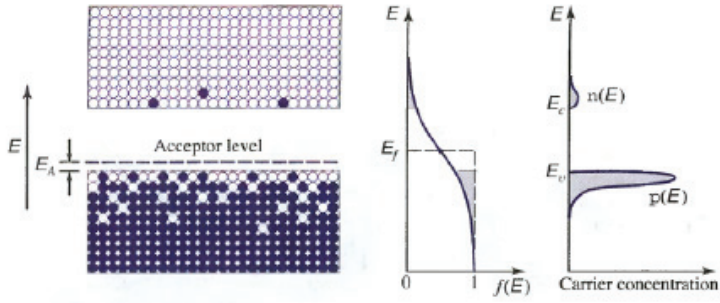


Figure 4.10b. Sketched scenario, as in Fig. 4.10a, but for *p*-type semiconductor when $p(E) > n(E)$.

In doped structures, according to electrical neutrality, $n + N_A = p + N_D$ (for fixed donors and acceptors). We notice that all the above discussions were regarding semiconductors in thermal equilibrium.

But if thermal equilibrium is destorbed, let us say, by an external electric current or a photon flux induces band-to-band transitions under too high a rate for inter-band equilibrium to be achieved, despite the fact that the conduction band electrons and valence band holes are in their own equilibrium. This situation is known as quasi-equilibrium, which arises when relaxation (decay) times for transitions within each of the bands are much shorter than the relaxation time between the two bands. Thus, the inter-band relaxation time is less than 10^{-12}seconds, whereas the radiative electron-hole recombination time is 10^{-9}seconds. Under these conditions, two separate Fermi functions are used for two bands: $f_c(E)$ and $f_v(E)$, as well as two energy levels, denoted by E_{fc} and E_{fv}. They are called *quasi-Fermi levels* (see Fig. 4.11). When they lie inside the conduction and valence bands, respectively, the concentration of both electrons $n(E)$ and hole $p(E)$ can be quite large.

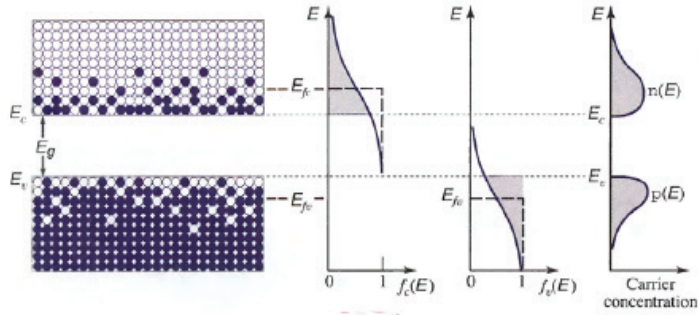


Figure 4.11. Sketched scenario, as in Figs. 4.10a,b, but for the case of the absence of thermal equilibrium, where Fermi energy levels are presented as two energy levels denoted by E_{fc} and E_{fv}.

The quasi-Fermi levels are:

$$E_{fc} = E_c + (3\pi^2)^{2/3} \cdot h^2 \cdot n^{2/3} / 2m_c \tag{4.43a}$$

$$E_{fv} = E_v + (3\pi^2)^{2/3} \cdot h^2 \cdot n^{2/3} / 2m_v \tag{4.43b}$$

For arbitrary T, if the amounts of n and p are sufficiently large, so that $E_{fc} - E_c >> k_B T$ and $E_{fv} - E_v >> k_B T$, the quasi-Fermi levels lie deep within the conduction and valence bands (see right panels in Fig. 4.11).

4.5. Law of Mass Action

Before entering into the subject, let us consider approximations of the Fermi function. Indeed, it was shown that when $(E-E_f) >> k_B T$, $f(E) \sim \exp[(E-E_f) / k_B T]$, i.e., the Fermi function is an exponential function. Similarly, when $(E-E_f) << k_B T$ [or $(E_f - E) >> k_B T$], $1 - f(E) \sim \exp[(E-E_f) / k_B T]$, i.e., it also is an exponential function. These conditions apply when the Fermi level lies within the bandgap, but away from its edges by an energy level of at least several times $k_B T$. Thus, at T=300 K, $k_B T$ = 0.026 eV, whereas in Si E_g = 1.12 eV and in GaAs E_g = 1.42 eV, that is for both semiconductors, the bandgap energy is more than the thermal energy $k_B T$.

So, we can use the above approximation of the Fermi function for an integral presentation of electron-hole pairs both in pure (intrinsic) and composite (doped) semiconductors, we get:

$$n = N_c \cdot \exp\left[(E_c - E_f) / k_B T\right], \tag{4.44a}$$

$$p = N_v \cdot \exp\left[(E_f - E_v) / k_B T\right], \tag{4.44b}$$

$$n \cdot p = N_c \cdot N_v \cdot \exp\left[-(E_g / k_B T)\right]. \tag{4.44c}$$

where

$$N_c = 2\left[(2\pi \cdot m_c \cdot k_B T) / h^2\right]^{3/2} \tag{4.45a}$$

$$N_v = 2\left[(2\pi \cdot m_v \cdot k_B T) / h^2\right]^{3/2} \tag{4.45b}$$

For $m_c = m_v$, if E_f is closer to the conduction zone, then $n > p$, whereas if E_f is closer to the valence zone, then $p > n$. In thermal equilibrium, the product $n \cdot p$ is independent of the location of Fermi energy E_f. Indeed,

$$n \cdot p = 4\left[2\pi \cdot k_B T/h\right]^3 \cdot (m_c \cdot m_v)^{3/2} \cdot \exp\left[-(E_g / k_B T)\right] \tag{4.46}$$

The constancy of the concentration (population per unit volume) product is called the *law of mass action*. It is useful both for pure and doped semiconductors, for which $n = p = n_i$. From Eqs. (4.44c) and (4.46) we get:

$$n_i = (N_c \cdot N_v)^{1/2} \cdot \exp\left[-(E_g / 2\, k_B T)\right] \tag{4.47}$$

which is called the *intrinsic carrier concentration*, leading to a new presentation of *mass action*:

$$n \cdot p = n_i^2 \tag{4.48}$$

Thus, for $T = 300$ K in Si: $n_i = 1.5 \cdot 10^{16}$ [m^{-3}]; in GaAs. $n_i = 1.8 \cdot 10^{12}$ [m^{-3}]; in GaN: $n_i = 1.9 \cdot 10^{-4}$ [m^{-3}]. As for doped semiconductors of *n*-type, $n = N_D$ and $p = n_i^2 / N_D$. This is only if $E_f >> kT$ and lies within the bandgap. If it lies inside the conduction or valence zone, we deal with a *degenerate semiconductor*, and approximations (4.44) to (4.46) cannot be used. So, in the equations $n \cdot p > n_i^2$ or $n \cdot p < n_i^2$ is valid.

4.6. Generation and Recombination of Electrons and Holes in Thermal Equilibrium

The thermal excitation of electrons from the valence band into the conduction band results in *electron-hole generation* (see Fig. 4.12). Thermal

equilibrium requires that this generation process be accompanied by a simultaneous reverse process of de-excitation, which is called *electron-hole recombination*. This process occurs when an electron decays from the conduction zone to fill a hole in the valence zone (see Fig. 4.12). The energy released by the electron may take the form of an emitted photon. In this case the process is called *radiative recombination*. Non-radiative recombination can occur via a number of independent competing processes, including transfer of energy to lattice vibrations creating one or more photons or to another free electron.

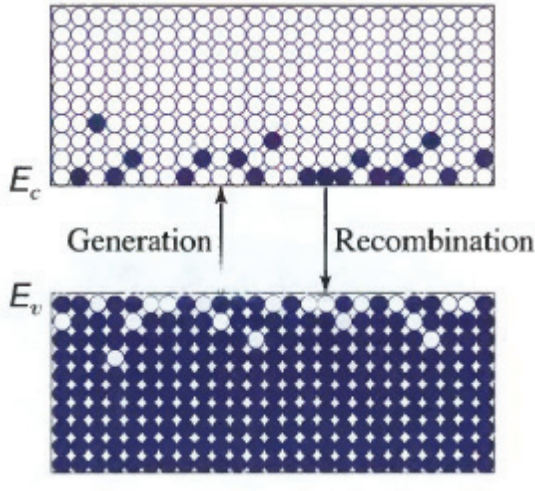


Figure 4.12. Schematic presentation of effects of the generation of electron and recombination electron with hole.

Non-radiative recombination also takes place at surfaces and indirectly via *traps* or *defect centers*, which are associated with impurities or defects of the lattice that lie within the forbidden zone. These impurities or defect states can act as a recombination center if it is capable of trapping both an electron and a hole, increasing their probability of recombination (see Fig. 4.13).

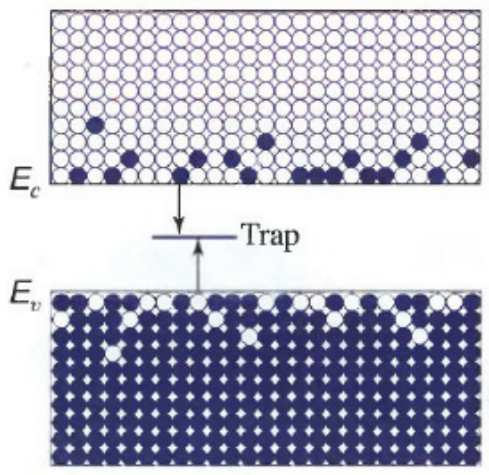


Figure 4.13. Schematic presentation of effects of generation of electron and recombination electron with hole in existence of traps inside the forbidden zone.

Impurity-assisted recombination can be radiative and non-radiative. Taking both an electron and a hole for recombination, it determines the rate of recombination, as the product of the concentration n and p. Thus:

$$rate\ of\ recombination = r \cdot n \cdot p \tag{4.49}$$

Here the recombination coefficient r (in cm^3/s) depends on the characteristics of the material, including its composition and defect density, temperature and doping level.

Electron-Hole Injection. When generation and recombination rates are in balance, usually called the *steady-state regime*, we deal with equilibrium concentration of electrons n_0 and holes p_0. If G_0 is the rate of thermal electron-hole generation at a given temperature, and r is the rate of pair recombination, then in thermal equilibrium

$$G_0 = r \cdot n_0 \cdot p_0 \tag{4.50}$$

The product is approximately the same whether the material is n-type, p-type or intrinsic (e.g., pure). If now some external (non-thermal) injection mechanism, such as light falling on the material, occurs, additional pairs of electron-holes will be generated at a steady rate R (pairs per unit volume per

unit time). A new steady-state will be reached in which concentrations are: $n = n_0 + \Delta n$ and $p = p_0 + \Delta p$. It is clear that $\Delta n = \Delta p$, since electrons and holes are created in pairs. Now a new rate of generation and recombination can be summed up to give:

$$G_0 + R = r \cdot n \cdot p \tag{4.51}$$

Accounting for Eq. (4.50), after straightforward computations we get:

$$R = r(n \cdot p - n_0 \cdot p_0) = r \cdot \Delta n \cdot (n_0 + p_0 + \Delta n) = \Delta n / \tau \tag{4.52}$$

With

$$\tau = [r \cdot (n_0 + p_0 + \Delta n)]^{-1} \tag{4.53}$$

For the case of $(n_0 + p_0) >> \Delta n$ (the case of insufficient injection) (4.53) yields

$$\tau = [r \cdot (n_0 + p_0)]^{-1} \tag{4.54}$$

which is called the *excess–carrier recombination lifetime* [5–14].

For *n*-type material $n_0 >> p_0$ and $\tau \sim 1 / (r \cdot n_0)$, whereas for *p*-type material $n_0 << p_0$ and $\tau \sim 1 / (r \cdot p_0)$. However, these formulas are not correct in the presence of traps that play an important role in the process [1–3].

Now we can describe the physical meaning of *electron-hole recombination lifetime***.** For this purpose, we will introduce the rate equation for injected-carrier concentration written in such a manner [3–5, 12–15]:

$$d(\Delta n)/dt = R - \Delta n / \tau \tag{4.55}$$

In a steady-state regime $d(\Delta n)/dt = 0$, deducing to Eq. (4.52). Now, if the source of injection is removed at the time t_0, i.e., $R=0$, Δn decays exponentially with time τ according to law:

$$\Delta n(t) = \Delta n(t) \cdot \exp\{-(t - t_0)/\tau\} \tag{4.56}$$

In another limiting case of the presence of a strong injection, as follows from Eq. (4.53), the lifetime of electron-hole recombination τ *itself* is a function of Δn.

On the other hand, in steady-state, if the rate R is known, the steady-state injected concentration can be determined by:

$$\Delta n = R \cdot \tau \tag{4.57}$$

from which one can obtain the total concentration of electrons, $n=n_0 + \Delta n$, and holes, $p=p_0 + \Delta p$, accounting for $\Delta n = \Delta p$.

We now introduce the *internal quantum efficiency* of semiconductor materials η_i, which is defined as a ratio of the radiative electron-hole recombination r_r to the total electron-hole recombination coefficient r, which is a sum of the radiative r_r and non-radiative r_{nr} recombination coefficients, i.e.

$$\eta_i = r_r / r = r_r / (r_r + r_{nr}) \tag{4.58}$$

Equation (4.58) can be written via the recombination lifetimes, τ_r and τ_{nr}:

$$1/\tau = 1/\tau_r + 1/\tau_{nr} \tag{4.59}$$

So, the total internal quantum efficiency can be easily found as

$$\eta_i = r_r / r = (1/\tau_r) / (1/\tau) = (1/\tau_r) / (1/\tau_r + 1/\tau_{nr}) = \tau_{nr} / (\tau_r + \tau_{nr}) \tag{4.60}$$

The radiative recombination lifetime determines the rate of photon absorption and emission, as was explained above, and it depends on the carrier (electron and photon) concentration and the material parameter r_r. For low to moderate injection rates

$$\tau_r = [r_r (n_0 + p_0)]^{-1} \tag{4.61}$$

which is in accordance with Eq. (4.54).

The non-radiative recombination lifetime is described by the similar equation:

$$\tau_{nr} = [r_{nr} (n_0 + p_0)]^{-1} \tag{4.62}$$

However, this parameter is more sensitive to the centers of defects existing in the forbidden (depletion) zone, than for the concentration of electrons and holes in the conduction and valence zones, because non-radiative recombination takes place via defect centers in the forbidden zone. Typical values of recombination coefficients and lifetimes are presented in Table 4.1.

Table 4.1. Two types of recombination and their lifetimes.

Material	r_r (cm^3/s)	τ_r	τ_{nr}	τ	η_i
Si	10^{-15}	10 ms	100 ns	100 ns	10^{-5}
GaAs	10^{-10}	100 ns	100 ns	50 ns	$5\cdot10^{-1}$
GaN	10^{-8}	20 ns	0.1 ns	0.1 ns	$5\cdot10^{-3}$

The radiative lifetime for Si is orders of magnitude longer than its overall lifetime because of its indirect *bandgap*. This results in a small internal quantum efficiency. For GaAs and GaN, having a *direct bandgap,* they show larger internal quantum efficiency ($5\cdot10^{-1}$ and $5\cdot10^{-3}$, respectively).

4.7. Photon Interactions with Semiconducting Materials

Before entering into a description of laser physical aspects, let us consider briefly the process of emission and absorption that can occur in semiconducting materials based on the discussions briefly described above considering the interaction of photons with atoms (see also Refs. [12–17]).

4.7.1 Processes of Emission and Absorption of Light in Semiconductor Materials

Figure 4.14 illustrates three main processes occurring in laser semiconductor diodes: a) spontaneous emission, b) absorption, and c) stimulated emission.

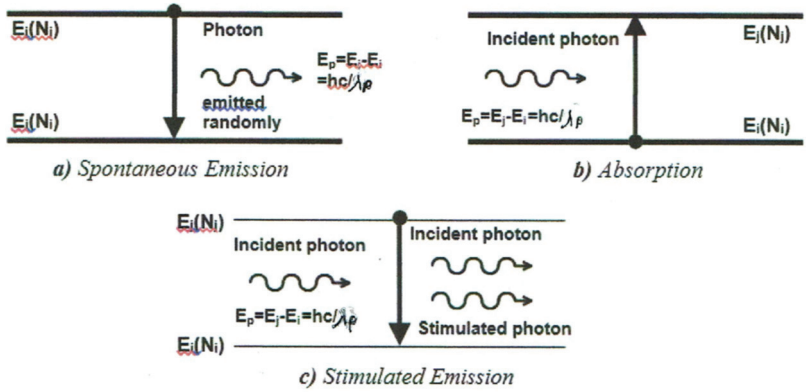


Figure 4.14. Schematically presented a) spontaneous emission, b) absorption, and c) stimulated emission.

Spontaneous Emission. An electron occupying an orbit of energy level E_j within an atom may randomly make the transition to another orbit of energy E_i (Figure 4.14a) by giving up a photon of energy $E_p = E_j - E_i = h\upsilon = hc/\lambda$. Each photon produced results from one atom making the transition from state E_j to E_i. Hence, the *rate of photon emission* is equal to $-dN_j/dt$ which is directly proportional to the density of atoms, N_j (atoms per m^3), in the energy level E_j, and we can write:

$$dN_j/dt = - A_{ji}N_j \tag{4.63}$$

Equation (4.63) represents the instantaneous rate of decline of the population N_j of energy level E_j given no other influences. Hence, if we pump a large number of atoms, N_j, from the ground state up into energy level E_j and then sharply switch off the pumping source, the population N_j and the spontaneous emission intensity will decay according to Eq. (4.63). We can readily solve this equation to give N_j as a function of time as follows:

$$N_j = \Delta N_j \exp(-A_{ji}\, t) = \Delta N_j \exp(- t / \tau_{ji}) \tag{4.64}$$

Equation (4.64) indicates that a population perturbation, in excess of thermal equilibrium, decays exponentially with a time constant $\tau_{ji} = 1/A_{ji}$, which is referred to as the *spontaneous emission lifetime* of state j. We now notice that Eq. (4.64) is correct only for transitions between two specific energy levels. More generally, spontaneous emission may occur to a number of energy levels below state j and we have for the rate of spontaneous emission

the following solution:

$$N_j = \Delta N_j \exp(-A_j t) = \Delta N_j \exp(-t/\tau) \quad (4.65)$$

where $A_j = \Sigma A_{jn}$ and $\tau_j = 1/A_j$ are the cumulative spontaneous emission rate constants for transitions between states j and n, and τ_j are the cumulative spontaneous emission time constants for these n transitions, respectively. We do not consider here additional non-radiative processes occurring in the lattice due to vibration or collisions.

Absorption. In Fig. 4.14b, photons of energy $E_p = E_j - E_i = h \cdot \upsilon = h \cdot c/\lambda$ may be absorbed by atoms in energy level E_i, which make the transition to energy level E_j as their electrons move between the corresponding orbits. Since each transition involves the absorption of one photon into one atom, the rate of absorption is equal to $-dN_i/dt$, and proportional to the population density of atoms, N_i, in energy level E_i and to the photon energy density at frequency υ. Hence, we can write:

$$dN_j/dt = B_{ij} N_i \rho(\upsilon) \quad (4.66)$$

where $\rho(\upsilon)$ is the photon energy density per unit frequency interval at frequency υ within a broadband radiation field (i.e., $\rho(\upsilon)d\upsilon$ is the photon energy density in the frequency interval υ to $\upsilon + d\upsilon$) and B_{ij} is the proportionality constant for absorption.

Stimulated Emission. In this process, a photon of energy $E_j - E_i = h \cdot c/\lambda$, incident on an atom which has an electron in an orbit of energy level E_j, stimulates that electron to make the transition to energy level E_i giving up an additional (to spontaneous emission) photon in the process (Figure 4.14c). Intuitively, the rate of stimulated emission is proportional to the population, N_j, of the E_j energy level and to the incident photon energy density per unit frequency interval, $\rho(\upsilon)$, giving

$$dN_j/dt = B_{ji} N_i \rho(\upsilon) \quad (4.67)$$

where B_{ji} is the proportionality constant for stimulated emission. The photons resulting from stimulated emission have the same energy (wavelength and frequency), direction, phase and polarization as the incident stimulating photons. This process results in the creation of new photons, which are added to the incident beam. Hence, *if stimulated emission dominates over absorption the incident beam is amplified* (the aspect of amplification in lasers will be discussed in Chapter 6).

The proportionality constants A_{ji}, B_{ij} and B_{ji}, in the rate expressions for spontaneous emission, absorption, and stimulated emission, are referred to as the *Einstein Coefficients*. We will not enter into his theory, based on the Boltzmann law and on Planck's theory of radiation from a black body as discussed in Chapter 3, but we will present again the relations between these coefficients, rewriting them as elements of matrices, that is,

$$B_{ij}g_i = B_{ji}g_j \tag{4.68}$$

$$A_{ji} = [8\pi n^3 h\upsilon^3 / c^3]\, B_{ji} \tag{4.69}$$

In Eq. (4.68), the functions $g_i(\upsilon)$ and $g_j(\upsilon)$ are referred to as the *lineshape* of the transition and $g(\upsilon)d\upsilon$ is the relative probability that light is absorbed or emitted by the *i-j* or *j-i* transitions, respectively, in the frequency range υ to $\upsilon + d\upsilon$. We do not enter deeper into this subject because it is out of the scope of this book; we only will notice that over the full range of possible frequencies of optical radiation of any material, the probability of emission or absorption must be unit. This condition, then, allows finding the distribution of the lineshape function $g(\upsilon)$ in the frequency domain both for absorption, and stimulated and spontaneous emission of light occurring in any solid material, namely, in semiconductors.

According to that mentioned above, we can present the emission and absorption via the steady-state nature of each atom of a specific substance – to be in discrete energy levels that can be listed in order of ascending discrete values of energy E_{ij}: E_1, E_2, E_3,…, E_n. This means that each atom of any material or substance has a characteristic set of energy, called *steady-state conditions* of the atoms and free electrons inside the material, as an atomic system, crystal-like or liquid. Under conditions of thermal equilibrium for temperature $T > 0$ *K,* the number of atoms having energy E_i is related to the number of atoms having energy E_j by the Boltzmann relation [1–9]

$$E_i / E_j = \exp[(E_j - E_i) / k_B T] \tag{4.70}$$

where the Boltzmann's constant equals: $k_B = 1.38 \cdot 10^{-23} J \cdot K^{-1}$. Here, the energy of transfer of the atom (or corresponding valence electron) from lower energy level i to higher energy level j ($j > i$) according to quantum theory, can be written as :

$$h\nu_{ji} = E_j - E_i, \qquad j > i \tag{4.71}$$

Similarly, relation (4.71) states that one quantum of light – a photon, with energy hv_{ji} can be absorbed by the atom, which in consequence has increased in energy from one of its steady states of the atomic system with energy E_i to another steady-state of the atomic system with energy E_j. Consequently, a photon will be emitted when a downward transition occurs from E_j to E_i, and this photon will have the same frequency v_{ji}.

In this context, considering a flux of q photons across unit area per unit time, we can write an intensity of light radiation by use of the "wave-corpuscular" dualism and present the light intensity as a stream of photons [3–8, 16, 17], i.e.:

$$I = q \cdot h \cdot v \tag{4.72}$$

Similarly, any other quantity defined within the wave context also has its counterpart in the corpuscular context. So, in our further explanation of matter, we will use both the wave and the corpuscular (i.e., particle) representation. If so, Eq. (4.70) to (4.72) state that the light frequencies emitted in the form of photons or absorbed from photons by atoms fully characterize each material or substance, crystal-like or liquid under consideration. When an excited system returns to its lowest state, some return pathways are more probable than others, and these probabilities, described by the corresponding statistics, Fermi (for electrons inside atoms) or Boltzmann (for photons), are also characteristic of the specific atoms or materials under consideration.

In other words, the light wave, as a continuous electromagnetic wave, can be regarded as a probability function whose intensity at any point in space (or within an atom) defines the probability of finding a photon (or an electron) there. According to this wave-particle dualism, the emission and/or absorption spectrum of any material can be used for its identification and to determine the quantity present. These ideas form the substance of the subjects known as photonics and spectroscopy, which are very extensive and powerful tools in materials analysis, but outside the scope of our book.

4.8. Physical Principles of Laser Operation

The laser is a very special source or detection-based element, the discovery of which in 1960 by Maiman [8] gave a push to optical fiber and wireless optical communication. The word *laser* is an acronym for *Light Amplification by Stimulated Emission of Radiation*, and we will briefly describe the processes on which it depends. Further deep analysis of laser characteristics was carried out by Russian researchers during the sixties of

the last century and their results are fully described in Ref. [2, 4].

As was mentioned in the previous paragraph, a photon could cause an atomic system to change from one of its steady states to another according to the process described by (4.71), that is, the change of the atomic system from a lower to a higher energy state. However, if the system was already in the higher of the two states when the photon acted, then this action would cause a transition down to the lower state, still in accordance with (4.71), but now by changing j on i and i on j (here becomes $j < i$). This process is called *stimulated emission* since the effect is to cause the system to emit a photon with energy hv_{ji} corresponding to the energy lost by the atomic system. Finally, we have two kinds of photons – the *acting photon* (as an element of outer light radiation) and the *emitted photon* (as an element of light excited by the material as an atomic system) [9–17].

Let us explain stimulated emission by the use of a very simple scenario illustrated qualitatively in Fig. 4.15. Here the emitted light photon (denoted by "E") has a wavelength of 550 nm that corresponds to the "green" visual light spectrum.

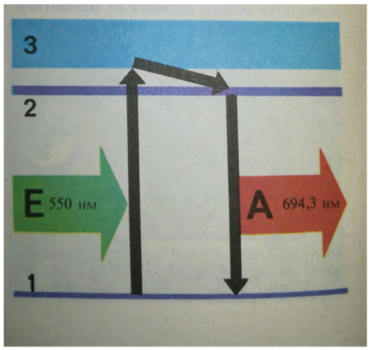


Figure 4.15a. Sketch on the process of stimulated emission of a laser.

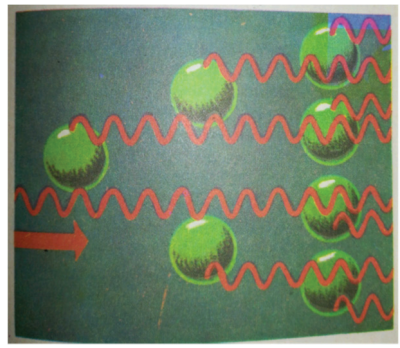


Figure 4.15b. The avalanche process of photon radiation in lasers.

The emitted photon transfers its energy to the 1st level (of lower energy) atom or its valence electron, giving rise to its jump into the 3rd level (of higher energy), which is nonstable [lifetime is $t_1 \sim 10^{-8}$ s]. Then, the electron fast falls onto the metastable 2nd level with the lifetime of the electron of about $t_2 \sim 10^{-2}$ s. This metastable level 2 accumulates many such atoms (e.g., valence electrons) because $t_2 >> t_1$ (Fig. 4.15a). So, during the longer period after the fall from metastable level 2 to ground level 1 with respect to that from the unstable 3rd level to the metastable 2nd level, a lot

of "red" photons with a wavelength of ~695 nm will be created (from one to two, from 2 to 4, from 4 to 8, and so on, see Fig. 4.15b).

This process was called *stimulated radiation* by Einstein and it is the main process of radiation accompanying the operation of avalanche laser sources and diodes, based on the avalanche (exponential) growth of photons stimulated by the laser itself caused by stimulated emission. Finally, we have *laser light*.

We must also mention that another process exists when an atomic system is not in its lowest energy state and is not in a stable equilibrium condition. If it has not had any interaction with the outer background but is embedded into a hot environment (even with a room temperature of 290 K = 17 °C), it will eventually fall to its lower state. Thus, an atomic system with state E_j will fall spontaneously to the lower state E_i even without the stimulus of photon energy hv_{ji} in a time which depends on the exact nature of the equilibrium conditions. The emitted photon that results from this type of transition is thus said to be due to *spontaneous emission* [9–15].

To understand quantitatively how a laser works in these two regimes, spontaneous and stimulated, let us consider a simple two-level atomic system with energies E_0 and E_1, respectively, as shown in Fig. 4.16a. We also suppose that this two-level system is illuminated by light radiation at a frequency:

$$\nu_{10} = (E_1 - E_0) / h \tag{4.73}$$

Initially, if the system is in thermal equilibrium at temperature T, the relative numbers of atoms (or valence electrons) will be, according to [12–17],

$$E_0/E_1 = \exp [(E_1 - E_0) / k_B T] \tag{4.74}$$

As follows from (4.74), for $E_1 > E_0$, we obtain $N_1 < N_0$. This means, if we assume the probability of transition is the same for two transitions, more atoms will be raised from the lower to the higher state than vice versa since, according to (4.66), there are more atoms in the lower state. As the intensity of radiation is increased (i.e., radiation at frequency ν_{10} is steadily increased from zero), the number of downward spontaneous transitions will increase as the occupancy of the upper state rises, tending toward the saturation condition where the occupancies of the two states and the rates of transition in the two directions are equal. Now we will consider two variants of photon emission: *stimulated* and *spontaneous*, as shown in Fig. 4.14 on the left and right sides, respectively.

This process is called the two-level atomic process. In the case of the spontaneous emission regime, the desired electron from the conductive zone, the lowest level of the conductive-wedge level of energy E_1 spontaneously falls into the valence zone, filling the lowest to the valence-wedge free level of energy E_0 (called the process of electron-hole recombination). This process is accompanied by the emitting of a photon with energy $h\nu_{10}$. When, conversely, a photon with energy enters into the semiconductor with energy $h\nu_{10}$, it stimulates an electron to enter from the valence zone into the conductive zone, and after (due to changes of temperature or other conditions) falls back down, emitting the photon with the same energy $h\nu_{10}$ (see Fig. 4.16a). But these are very primitive stages of spontaneous and stimulated emissions. Often, there are more complicated stages of stimulated-spontaneous mechanisms of emission observed in semiconducting lasers.

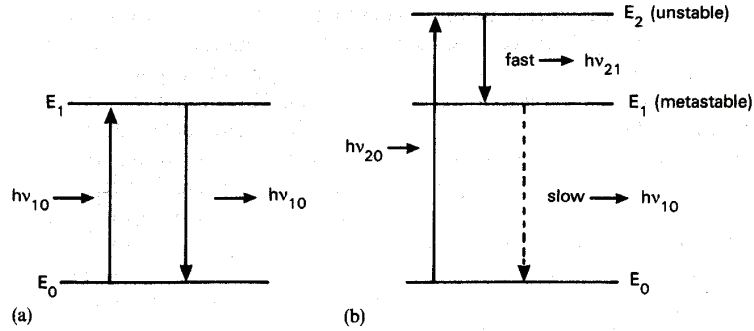


Figure 4.16. a) two-level atomic system for *spontaneous* emission description, and b) three-level atomic system for *stimulated* emission description (according to [17]).

Indeed, considering now a three-level atomic system shown in Fig, 4.16b, we have a lowest level E_0, a metastable level E_1, and an unstable level E_2, a simple sketch of which was presented in Fig. 4.15a and explained qualitatively.

If the three-level system, being initially in thermal equilibrium, is irradiated with light frequency

$$\nu_{20} = (E_2 - E_0) / h \tag{4.75}$$

the effect is to raise a large number of atoms (or valence electrons) from the level with energy E_0 to the level with energy E_2. These particles then decay quickly [because the lifetime at this level ~ 10^{-8} s] to the state E_1 by *spontaneous* emission only (since the input light frequency ν_{20} does not correspond to this transition with frequency ν_{21}), and subsequently only slowly from this metastable long-lived [with t~10^{-2} s, see above] return back to the ground state E_0. Owing to this process, a larger number of atoms can be in state E_1 than in state E_0. Since this process does not correspond to Boltzmann distribution (4.70), it is known as an *inverted population* [2]. Because the process of transition from the ground level to the unstable level is about a million times shorter than that from unstable to metastable, a lot of excited atoms (i.e., electrons) are accumulated into the metastable level. After the incidence of the second beam of light on this inverse population at frequency the effect is described by (4.73)-(4.74).

It produces a downward movement by stimulated emission as it can excite atoms from E_0 to E_1. Thus, a lot more stimulated photons are produced than are absorbed by excitations (see also qualitative explanation shown in Fig. 4.15b). We call this the beam receiving *gain* from the atomic system (material or substance). In other words, the light beam is *amplified*. The system is said to be pumped by the first beam to provide gain for the second beam (as shown by Fig. 4.16b). We have the effect of the *light amplification by stimulated emission of radiation*, that is, we have obtained the *laser* [3–12].

Now putting the desired material system between two parallel mirrors, we can not only amplify the stimulated photons, but also produce an *oscillator*, because as follows from Fig. 4.17a and 4.17b, such oscillations cover a wide spectrum of frequencies radiated by the laser. Moreover, there is a difference between laser radiation and visual light. Thus, the process of radiation by laser occurs with high accuracy and the phase of all radiated photons fully coincide with each other. Finally, the resulting laser beam will oscillate strongly (Fig. 4.17b). Such radiation of the laser is called *coherent*, and differs from that of visual light, generated let us say, by a lamp, which consists of a lot of partial short wavelength oscillations (see Fig. 4.17a).

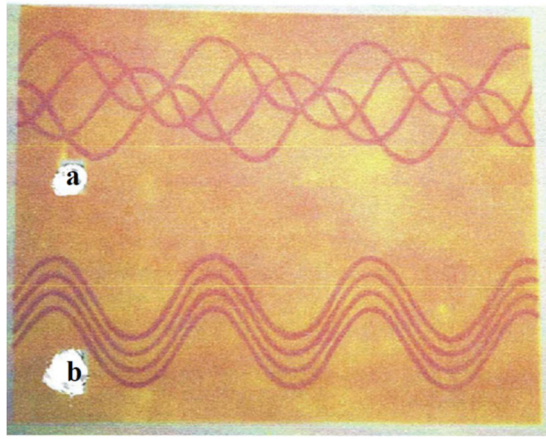


Figure 4.17. a) Radiation of natural light, and b) coherent radiation of laser.

Finally, we have obtained monochromatic (with narrow frequency band) or polychromatic (with wide frequency band) coherent (with well-defined phase), and well-collimated light: we have *laser light*. The features and operation properties of the laser have been described and are the most commonly used light sources and detectors usually used in optical communication. Because, as was outlined in [9–12], most laser sources are currently constructed by use of semiconducting materials, or crystal-like materials, we briefly described the physical principles of semiconductor operation based on the zonal structure of materials described in Chapter 3.

Bibliography

[1] Born, M., and E. Wolf. 1964. *Principles in Optics*. New York: Pergamon Press.
[2] Fain, V. N., and Ya. N. Hanin. 1965. *Quantum Radiophysics*. Moscow: Sov. Radio (in Russian).
[3] Lipson, S. G., and H. Lipson. 1969. *Optical Physics*. Cambridge: University Press.
[4] Akhamov, S. A., R. V. Khohlov, and A. P. Sukhorukov. 1972. *Laser Handbook*. North Holland: Elsevier.
[5] Marcuse, O. 1972. *Light Transmission Optics*. New York: Van Nostrand-Reinhold Publisher.
[6] Siegman, A. E. 1976. *Lasers*. University Science Books: Mill Valley, CA.
[7] Yariv, A. 1976. *Introduction in Optical Electronics*, Chapter 5. New York:

Holt, Rinehart, and Winston.
[8] Maiman, T. H. 1960. “Stimulated optical radiation in ruby materials.” *Nature* 187:493–503.
[9] Kressel, H., and J. K. Butler. 1977. *Semiconductor Lasers and Heterojunction LEDs*. New York: Academic Press.
[10] Sze, S. M. 1985. *Semiconductor Devices: Physics and Technology*. New York: Wiley.
[11] Kressel, H., ed. 1980. *Semiconductor Devices for Optical Communications*. New York: Springer-Verlag.
[12] Agrawal, G. P., and N. K. Dutta, *Long-wavelength Semiconductor Lasers*, New York: Van Nostrand Reinhold, 1986.
[13] Dakin, J., and B. Culshaw, eds. 1988. *Optical Fiber Sensors: Principles and Components*, vol. 1. Boston-London: Artech House.
[14] Dakin, J., and B. Culshaw, eds. 1988. *Optical Fiber Sensors: Principles and Components*, vol. 2. Boston-London: Artech House.
[14] Coldren, L. A., and C. W. Corzine. 1995. *Diode Lasers and Photonic Integrated Circuits*. New York: Wiley.
[15] Morthier, G, and P. Vankwikelberge. 1997. *Handbook of Distributed Feedback Laser Diodes*. Norwood, Ma: Artech House.
[16] Palais, J. C. 2006. “Optical communications.” In Handbook: *Engineering Electromagnetics Applications*, edited by R. Bansal. New York: Taylor and Frances.
[17] Blaunstein, N., S. Engelberg, E. Krouk, and M, Sergeev. 2019. *Fiber Optic and Atmospheric Optical Communication*. Hoboken, New Jersey: Wiley.

CHAPTER 5

FUNDAMENTALS OF LIGHT EMITTERS, OPTICAL DIODES AND DETECTORS

As was mentioned in Chapters 3 and 4, the most commonly used light sources in optical communication, wired and wireless, as well as in LIDAR applications, are those based on semiconducting solid materials. Among them, most attractive for practical applications, are the light emitting diode (LED), the laser diode (LD), the photodiodes of *p*-*n* and *p*-*i*-*n* types, and the avalanche photodiode (APD) [1–11]. All of them act as emitted sources (e.g., lasers, see Chapter 4) or receiving detectors, which have found importance in electronics, rectifiers, logic gates, voltage regulators, or tuners, and in optoelectronic diodes, as well as in solar cells. Now we will start to describe the operational parameters and characteristics of optical sources and detectors, as both sides, the beginning and the final, terminals of any wired or wireless link, any optical network, and of LIDAR (see Chapter 1).

5.1. P-N Junction Operation Mode in Semiconductor Devices

In Chapters 3 and 4, it was shown that the pure semiconductor, where the amount of electrons prevails with respect to holes, has properties of the *n*-type semiconductor, and that, in which the amount of holes prevail, has properties of the *p*-type semiconductor. We put a question: *what will happen if we contact both types of pure semiconductor*, as shown in Figure 5.1. In this case, when $T > 0$ K, electrons from the *n*-type semiconductor will penetrate to the *p*-type semiconductor through the junction created between them. In the same manner, the holes will penetrate from the *p*-type semiconductor to the *n*-type semiconductor via the junction [1–6]. In other words, the *p*-*n junction* is a *home-junction* between *p*-type and *n*-type semiconductors. Finally, they will create a spatial electrical charge difference inside a home-junction and, therefore, the inner electric field, as shown in Figure 5.1.

The inner electric field regulates the number of electron-hole pairs, which increases with time, when this process is not compensated by the inverse process of recombination of electron-hole pairs, as major carriers of charges. In this case, as was mentioned in Chapters 3 and 4, the condition of *dynamic equilibrium* is observed. Of course, the width of this junction is too thick – around a few micrometers, and in Figure 5.1 it is represented as wider for a clear understanding of the process. The Fermi energy level E_f, determined in the previous two chapters, as *the maximum energy obtained by valence electron/hole to pass by the valence zone*, and is depicted by the dashed line in the middle panel of Figure 5.1 and by the dashed line inside the forbidden zone.

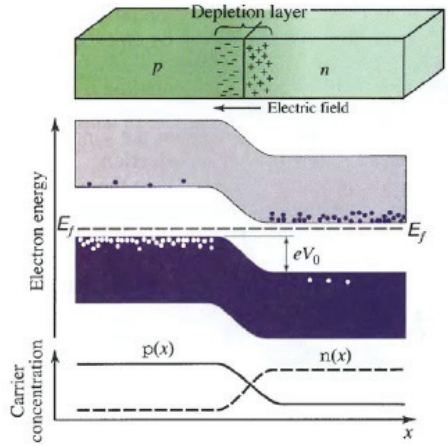


Figure 5.1. P-N junction – schematically presented principle of operation.

Figure 5.2 shows the difference between situations when an outer source is absent (outer source voltage $V = 0$) and when it exists with an outer voltage $V > 0$, called in literature the *biased p-n junction*.

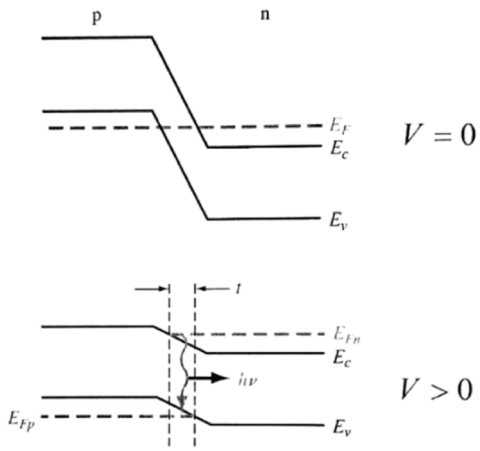


Figure 5.2. Difference between p-n semiconductor states without ($V = 0$) and with ($V > 0$) outer source.

As is clearly seen from Figure 5.2, for $V = 0$ there is a unified Fermi energy level E_f for the *p*-semiconductor and the *n* semiconductor inside the *p*-*n* junction (denoted in the upper panel of Figure 5.2 by the dashed line). When the *p*-*n* junction is electrically biased with an outer source of voltage $V>0$, splitting of the corresponding Fermi energy level in two is observed within the valence zone (with energy E_{Fp}) and the conductive zone (with energy E_{Fn}) denoted by dashed curves in the bottom panel of Figure 5.2.

We will now summarize all features regarding *p*-*n* junctions mentioned in Chapter 3 and shown in Figure 5.1 to Figure 5.3. When two semiconductors, *p*-type and *n*-type, are arranged to be in contact, as shown in Figures. 5.1 to 5.3, the following effects take place:

1. In the absence of an outer source and in thermal equilibrium, electrons and holes diffuse from areas with high concentration towards areas of low concentration (left panel of Figure 5.3). Electrons diffuse from the *n*-region to the *p*-region leaving behind their movements positively charged ionized donors. In the *p*-region the electrons recombine with existing holes, and only near the boundary of the junction are the rest of the electrons (denoted by signs "-" in the top-left panel of Figure 5.3). Holes diffuse from the *p*-region to the *n*-region leaving behind their movements negatively charged ionized acceptors. In the *n*-region the holes recombine with existing electrons, and only near the boundary of

the junction are the rest of the holes (denoted by signs "+" in the top-left panel of Figure 5.3).

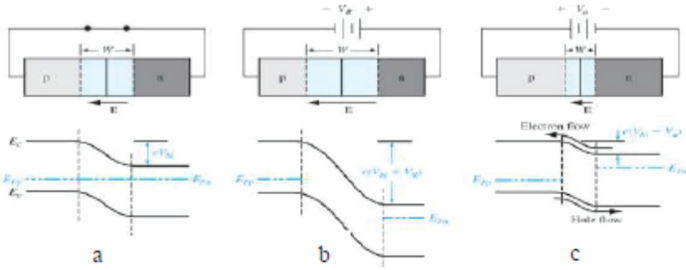


Figure 5.3. a) P-N junction in stationary regime; P-N-junction under outer source which is: direct with (c) and opposite to (b) the inner electric field.

2. As a result, a narrow region on both sides of the junction is nearly depleted, and mobile charged carriers are developed called the *depletion layer* or *p-n junction*. It contains *fixed* charges (positive (+e) in *n*-type semiconductors and negative (-e) in *p*-type semiconductors), creating an electric field in this layer directed from n-type to p-type, as shown at the left-top panel of Figure 5.3.

3. In thermal equilibrium and in the absence of an outer source, there is only a single Fermi function for the entire structure, which describes the transition of electrons or holes from the valence zone to the conductive zone and vice versa. So the Fermi energy levels in the *p*-band and *n*-band are the same (see dashed line in the top panel of Figure 5.2, and the left-top panel of Figure 5.3 denoted by E_F).

All the above features of the carrier (electrons and holes) movements are correct in the absence of an outer source.

4. When the *p-n* junction is *biased by the outer source*, its potential difference V provides a lower potential energy on the *n*-side relative to the *p*-side (two right panels in Figure 5.3). This field with the potential difference V_0 (called *built-in* or *intrinsic*) obstructs the diffusion of further mobile carriers through the junction region.

5. In the case of $V>0$ net (cumulative) current flows across the *p-n* junction and the connecting electric circuit, because these currents

are associated with drift and partly with the diffusion of both carriers, electrons and ions.

6. When we put charge "+" in the *p*-region, and "-" in the *n*-region, the outer field has the *opposite direction* to the inner field (i.e., the width of p-n junction will decrease) and the *increase of a current* through the circuit is observed (see Figure 5.3c). The P-N junction works as a *direct-biased (forward-biased) junction*.

7. Conversely, when we put "+" in the *n*-region, and "-" in the *p*-region (Figure 5.3b), the outer and inner fields have the *same direction* (i.e., the width of p-n junction will increase), *decreasing total current* through the circuit. The P-N junction works as an *opposite-biased (or reverse/inverse –biased) junction.*

To illustrate this process by the use of numerical examples let us consider that the concentration of electrons and holes under a temperature of $T \sim 290$ K is $2.5 \cdot 10^{13}$ particles / cm^3, then their product, $p \cdot n = 6.25 \cdot 10^{26}$. Let us consider that after some diffusion of electrons, we obtained in the junction $\sim 10^{16}$ electrons / cm^3.

As for holes, their number will be $\sim 6.25 \cdot 10^{26} / 10^{16} = 6.25 \cdot 10^{10}$ holes / cm^3 due to the law of working masses [1–4, 11] (see Chapter 4). Due to the transfer of electrons from the *n*-region to the *p*-region the concentration of electrons is decreased by 10^6 times [from the beginning there were 10^{16} electrons in the *n*-region]. The same process of decreasing hole concentration will be observed during the transfer of holes in the opposite direction – from the *p*-region to the *n*-region. Continuous decrease of concentration of carriers, electrons and holes, at the proximity of the boundary layer (i.e., *junction*) is shown by the two curves in the bottom panel of Figure 5.1. According to this figure, at the middle (where the curves cross each other) both charges are presented in equal concentration, which according to the law of working masses equal $2.5 \cdot 10^{13}$ particles / cm^3.

Thus, the total concentration of particles is $5 \cdot 10^{13}$ particles / cm^3, that is, $(10^{16} / 5 \cdot 10^{13})$ 200 times less in the junction with respect to each of them in the *n*-region and *p*-region separately (we remember that initially it was $\sim 10^{16}$ particles / cm^3, see above). So, the *p-n* junction is a *depletion zone* of charges. But, here the process is completely different when compared with diffusion in gases. Here, entering in the *p*-region, electrons, leave in the *n*-region positive holes. At the same time, holes diffused from the *p*-region to the *n*-region, leave in the *p*-region electrons with negative spatial charge. Finally, between these two spatial charges, the inner electric field is created, as shown in the left panel of Figure 5.3. The more particles

that diffuse, the stronger this inner electric field. Such diffused carriers in both directions are called the *majority electrons* and *majority holes*.

Let us now consider a more practical case, shown in Figure 5.4, where such a *p*-*n* semiconductor was connected to the electric battery in such a manner that the "+V" of the electrical source was in contact with the *p*-type semiconductor, and its "0" (called the *ground voltage*) was connected to the *n*-type semiconductor. In this case, the outer electric field will occur and the direction of which (according to the rule of electrostatics) will be directed from "+V" to "0", that is, will be directed opposite to the inner electric field (presented also in Figure 5.1 and Figure 5.3).

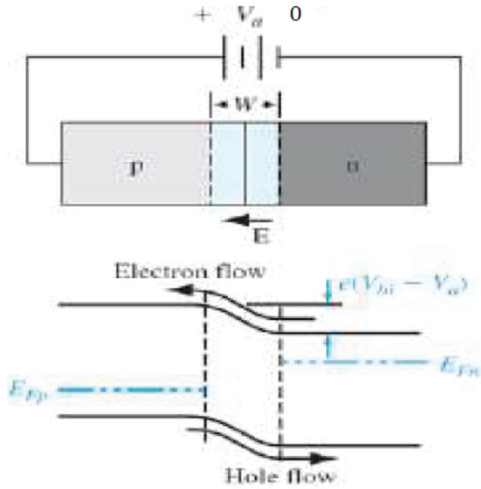


Figure 5.4. Processes occurring in P-N-junction introduced in an electrical circuit, called the biased junction.

As seen from the middle panel of Figure 5.4, the energy of the prohibited (depletion) zone between the valence and conductive energy band decreases from $e \cdot Ve$ to $e \cdot (Ve - V0)$, helping electrons more easily leave the valence band and fill the conductive band, finally, increasing part of the drift current (compared with inner diffusion) generated along the biased *p*-*n* junction.

Now, if "+V" of the electric source will be put to the *n*-type semiconductor and "0" to the *p*-type semiconductor, both fields, inner and outer, having the same direction, will increase the width of the junction region (see also Figure 5.3b, middle panel), leading to *decrease* of the total

current through the outer circuit to the minimum. Such a thick inner junction was called the *inverse-biased junction*, and the potential energy in the *n*-region is increased compared with the *p*-region.

If, conversely, "+V" of the electric source will be put to the *p*-type semiconductor and "0" to the *n*-type semiconductor (as shown in Figure 5.3c right panel), the outer field has the *opposite direction* to the inner field, will decrease the width of the junction region, *increasing* the possibility of the major carriers (electrons from the *n*-region and holes from the *p*-region) passing through the junction, and therefore, and finally, leads to *increase* of the total current through the outer circuit to the maximum. Such a thick inner junction was called the *forward-biased junction*.

So, the *p-n* junction works as a *diode* generating *minority* carriers inside the junction regulated by an outer electric field, which drift through the junction: electrons in the direction towards "+" and holes in the direction towards "-" of the biased electrical circuit. In a thin *p-n* junction minor carriers exist, electrons $\Delta n << n$, and holes $\Delta p << p$ (see the bottom panel in Figure 5.4), which give impact in their diffusion through the junction, called *minority carriers diffusion*, as was mentioned above with the help of Figure 5.1. In this case, one can find analytically close relations between the major and minor carriers (electrons (n) and holes (p)) playing the main role in drift and diffusion processes, respectively.

As was shown in Chapter 4, in the case of equilibrium, when $\Delta n = \Delta p$, (see also bottom lines in Figure 5.4) with an increase of temperature the amount of charged particles, electrons and holes, taking part in through-diode current creation, are

$$n = n_0 + \Delta n \text{ and } p = p_0 + \Delta p \tag{5.1}$$

It should be noticed that usually $\Delta n = \Delta p << n_0 + p_0$. As mentioned above, this allows us to define the *p-n* diode as

P-N diode works as a device guiding current only in one direction, called a forward-biased device.

Hence, a *p-n* junction, operating as a *diode*, has the following current-voltage (or *volt-ampere*) characteristic, called in the literature the *Shockley equation* [1–4, 9–11]:

$$i = i_s\{\exp(e \cdot V / k_B \cdot T) - 1\} \tag{5.2}$$

So, in the forward-biased *p-n* diode the current of *majority carriers* gives an increase of the total (cumulative = drift + diffusion) current by the exponential factor $\exp(eV/ k_BT)$. In formula (5.2) i_s is a minimum current

that is created by diffusion of *minor carriers* located inside the depletion junction in the inverse-biased device (when $V<0$). It can be expressed via the area of *p-n* junction, A, and their charge, $q = Z_e$ [for electrons and holes $Z=1$], number of injected charge carriers, n_i, number of minor diffused chargers (electrons) from *n*-type, N_n, number of minor diffused chargers (holes) from *p*-type, N_p, their lifetime before mutual recombination in *p-n* junction, τ_n and τ_p, respectively, as:

$$i_s = q \cdot A \cdot n_i{}^2 \cdot [(D_n/\tau_n)^{1/2} / N_n + (D_p/\tau_p)^{1/2} / N_p] \tag{5.3}$$

In Eq. (5.3) D_n and D_p are the coefficients of diffusion of minor electrons (from *n*-type) and minor holes (from *p*-type), which are functions of mobilities of these minor carriers, μ_n and μ_n, and temperature of the environment, T (in Kelvin):

$$D_n = \mu_n \cdot k_B \cdot T/q \text{ and } D_p = \mu_p \cdot k_B \cdot T/q \tag{5.4a}$$

To find N_n and N_p, we need to give relations between the conductivity of electrons, σ_n, and holes, σ_p, and their partial resistivity, ρ_n, and ρ_p, via their mobilities, μ_n, and μ_p, that is,

$$\sigma_n = 1/\rho_n = q \cdot \mu_n \cdot N_n$$

$$\sigma_p = 1/\rho_p = q \cdot \mu_p \cdot N_p$$

from which we get:

$$N_n = 1/(q \cdot \mu_n \cdot \rho_n) \text{ and } N_p = 1/(q \cdot \mu_p \cdot \rho_p) \tag{5.4b}$$

A sketched view of such a *p-n* diode (a), its electrical scheme (b), and volt-ampere characteristic (c), are illustrated by Figure 5.5.

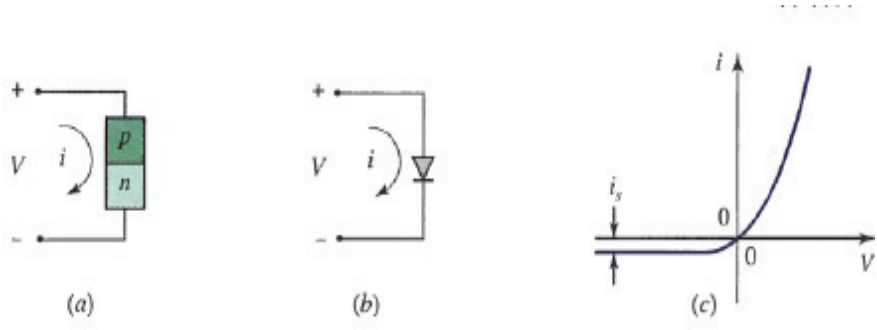


Figure 5.5. Sketched view of: a) p-n diode, b) electrical scheme, and c) its volt-ampere characteristics.

The excess carriers - electrons Δn entering the *p*-region and holes Δp entering the *n*-region, become the majority carriers, which then recombine with the local majority carriers in the *n*-region and *p*-region, respectively and, finally, concentration decreases in both regions. This process is known as *majority carrier injection*. An inverse process of decay of electron-hole pairs is known as *majority carrier extraction.*

5.2. Laser Diodes

5.2.1 Light-Emitted Diode (LED)

As follows from the information presented above the simultaneous availability of electrons and holes enhances the flux of emitted photons from a semiconductor – a light beam.

Definition of LED: The light-emitted diode (LED) is a *forward-biased p-n* junction semiconductor with a large radiative recombination rate arising from injected minority carriers. The semiconductor material is usually direct-bandgap (see definition in Chapter 4).

The difference between laser diodes (LDs) and LEDs is the following: in an LED the current density is low, and the light is generated by spontaneous emission. In laser diodes large current densities supply large numbers of electrons into the active region conduction band creating a high electron population and a large number of holes into the valence band creating empty electron energy levels.

The forward-biased diode, as an electronic device, is characterized by the photo-generated current, the major carriers of which are forward-directed through the junction, that is, are usually direct-bandgap. As follows

from the information presented above the simultaneous availability of electrons and holes enhances the flux of emitted photons from a semiconductor. Electrons are the major charge in *n*-type material, and holes are the major charge in *p*-type material, but the generation of copious amounts of light requires that both electrons and holes would be localized in the same region of space. This condition can be really achieved in the junction region of a forward-biased *p-n* diode described previously in paragraph 5.1.1.

As we mentioned in Chapter 3, *p-n* junction electron-hole pairs strongly recombine with each other emitting light as a stream of photons. In a steady-state regime, in the absence of outer forces (electric field or outer light), the velocity of generated electron-hole pairs and combined electron-hole pairs will be in equilibrium conditions. The process of strong electron-hole recombination together with spontaneous emission of electrons falling from the conductive to valence zones, together create a flux of photons, as shown in Figure 5.6, called in literature *electroluminescence*.

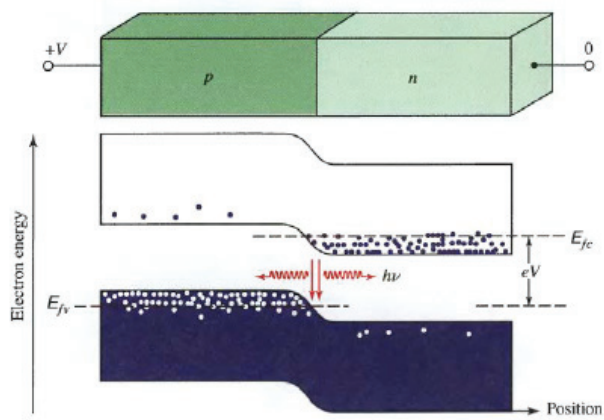


Figure 5.6. Energy band diagram of *p-n* junction that is strongly forward-biased by an applied voltage *V*. Dashed lines represent the Fermi levels, which are separated as a result of the bias.

Let us return to processes occurring in LED and put a question: how is light emitted from the *p-n* junction? As now shown in Figure 5.6, a forward-directed circuit with input voltage *V* causes holes from the *p*-region and electrons from the *n*-region to be forced and drift into the common *p-n* junction region by the process of minority carrier injection, where they

recombine (therefore this region is called the *depletion region*) and emit photons. So, the simultaneous abundance of electrons and holes within the junction region results in an injection of light radiation as a stream of photons due to strong electron-hole recombination [7–11].

A rate of injection of current carriers with charge $q = e$, i, can be found as:

$$R = i / q \cdot V \quad [\text{cm}^{-3} \cdot \text{s}^{-1}] \tag{5.5}$$

At the same time, under outer light radiation (see Fig. 5.6) *p-n* junction of the semiconductor can increase the current inside the outer electric scheme. So, we can state that LEDs can be used both in photonics and in photo electronics. Knowledge of quantum efficiency, η_i, as a ratio of the number of photons emitted by carriers to the number of carriers, and a total current according to Eq. (5.2) allows us to find the flow of light emitted by an LED, i.e.,

$$\Phi = \eta_i \cdot i / q = \eta_i \cdot V \cdot R \tag{5.6a}$$

Accounting for conditions $\Delta n = \Delta p$ discussed in paragraph 5.1 and introducing the time the process of carriers injection and the time of mutual recombination, $\tau = \eta_i \cdot \tau_r$, gives another expression of the photon flux emitted by LED:

$$\Phi = \eta_i \cdot \Delta n \cdot V / \tau = \Delta n \cdot V / \tau_r \tag{5.6b}$$

The output power of a LED increases linearly from zero with the applied drive current (see Figure 5.7).

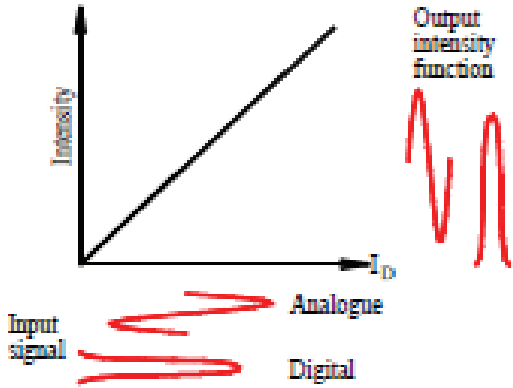


Figure 5.7. The optical intensity vs. the drive current for two kinds of signals: analog and digital.

In a *digital modulation* scheme (see definitions in Chapter 1) the device is biased at zero drive current and current pulses representing digital ones are applied to generate the optical pulses (see Figure 5.7).

For *analog modulation* (see definitions in Chapter 1), the device is biased mid-way between zero and the maximum drive. The power/current characteristic and the simple modulation schemes imply that the biasing and drive circuitry are straightforward. In addition, the current density in LEDs is lower than that in lasers.

Mathematically, the linear dependence of optical power of emitted light by an LED versus the amplitude a_i of the light emitted by an LED and the drive current can be simply presented by the following relation:

$$P = a_i i \tag{5.7}$$

We should also notice that light is generated in semiconductor sources /lasers as electrons fall from the bottom of the conduction band to the top of the valence band producing photons with a minimum energy equal to the bandgap, E_g. The electrons occupy a small range of energy levels (states) at the bottom end of the conduction band and fall to any level in a range of empty states (holes) at the top end of the valence band. There is thus a small distribution of photon energies and corresponding wavelengths in the emitted light.

The distribution of wavelengths of light generated depends on the distribution of electron energies in the conduction band and the distribution of empty states as in the valence band. The number of available states as a function of energy is defined by the *density of states function* (Figure 5.8a) and the distribution of electrons in these states is defined by the Fermi-Dirac function (Figure 5.8b).

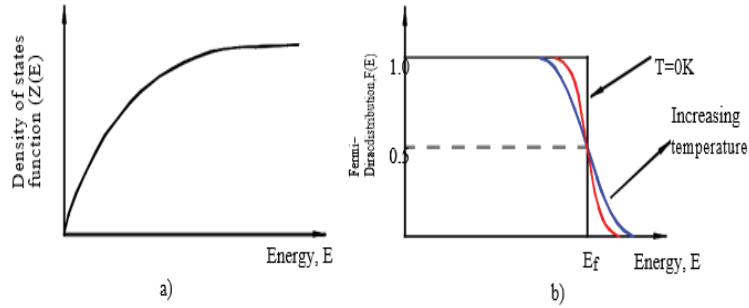


Figure 5.8. a) The density of states function and b) the Fermi-Dirac function**.**

Due to their energy distribution along the narrow optical radiated bandwidth LEDs are more reliable than laser diodes with less chance of current-induced heating at defects and eventual destruction. The output power of LEDs for communications applications ranges from about 25 μW (-16 dBm) to about 1 mW (0 dBm) with powers around 100μW (-10 dBm) being typical.

In Chapter 10 it will be shown that the spatial and coherence properties of the LED emission are poor, and it is a difficult optical problem to collect the light and focus it down to a small spot size. Hence, for single mode fiber with its core diameter less than 8 μm, the power launching from a LED is highly inefficient, incurring losses of 20–25 dB. For this reason, LEDs are only used with multimode fiber applications

5.2.2. Laser Diode (LD)

Laser diodes (LD), have a *forward-biased p-n* semiconductor junction with two parallel surfaces that act as reflectors, and therefore, works as an optical amplifier. Amplification is achieved by use of the returning ray paths provided by mirrors for optical feedback.

These mirrors are usually implemented by cleaving the semiconductor material along its crystal planes, as shown in Figure 5.9

according to [9–11]. The sharp refractive index difference between the crystal and the surrounding air causes the cleaved surfaces to act as reflectors. The semiconductor crystal acts both as a gain medium and as an optical resonator, as illustrated in Figure 5.9, achieving a sufficiently large gain coefficient. At the same time, the feedback converts the optical amplifier into an optical oscillator, i.e., into a polychromatic coherent laser.

The laser diode (LD) has many features similar to the light-emitted diode (LED) because in both devices the source of energy is an electric current injected into a *p-n* junction. However, the light emitted from an LED is generated by *spontaneous* emission, whereas the light from an LD arises from *stimulated* emission (see definitions above).

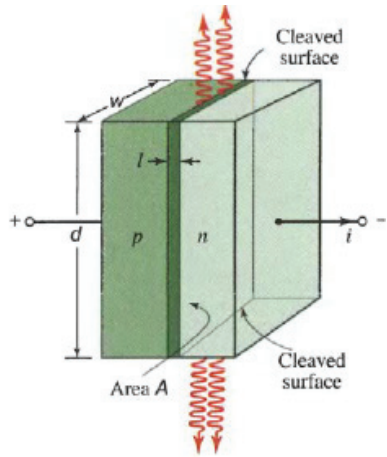


Figure 5.9. A sketched view of LD as a forward-directed *p-n* semiconductor junction with two parallel cleaved surfaces that act as reflectors (according to [11]).

There is a principal difference between LED and LD detectors. If in Figure 5.7 for LED the drive current starts from zero, in an LD detector the drive current starts from some threshold i_{TH}, as shown in Figure 5.10 (rearranged from [9, 10]), after which the same linear dependence of light power versus the drive current takes place according to the same formula (5.7), but for $i > i_{TH}$. The process is as follows. As the drive current to a laser diode increases from zero, light is generated by spontaneous emission until the current is high enough to achieve a population inversion which provides sufficient stimulated emission and gain to exceed the loss and tune-on laser oscillation. The precise point at which the gain exceeds the loss and laser operation begins is called the *threshold* and the drive current at this point is

known as the threshold current, i_{TH} (see Figure 5.10). Below the threshold the laser operates like an LED and the power increases slowly with increasing current. Above the threshold the power increases linearly rapidly with increasing drive current in the region of full laser operation (see Figure 5.10).

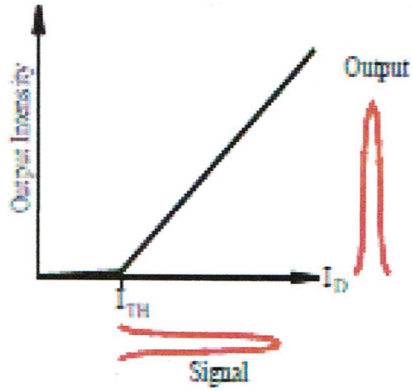


Figure 5.10. Output intensity of optical pulse signal vs. drive current.

Current pulses representing the digital ones are then applied to generate the optical pulses (see Figure 5.10). The threshold feature causes an operational problem because the threshold current increases with temperature.

Physical processes of light emission in LDs depend on the refractive properties of two-sided mirrors with respect to that of the semiconducting *p-n* junction surface. Therefore, the coefficient of reflection $\acute{R}$ from two-sided mirrors around the *p-n* junction depends on the refractive index of the semiconducting surface, n_{sc}, as:

$$\acute{R} = [(n_{sc} - 1) / (n_{sc} + 1)]^2 \tag{5.8}$$

It is clear that the reflection coefficient is less than unit because losses exist due to refraction both in the mirrors, α_m, and in the semiconductor α_{sc}, respectively. Moreover, for simplification of the subject under discussion, we assume that for both mirrors the same refraction occurs and therefore, the same coefficients of losses can be taken into consideration, i.e., $\alpha_{m1} = \alpha_{m2}$, and $\acute{R}_1 = \acute{R}_2 = \acute{R}$. In such assumptions, the attenuation parameters of the mirrors depend on the reflection coefficient defined by (5.8) and the height of the LD surface, *d* (see Figure 5.9):

$$\alpha_m = \alpha_{m1} = \alpha_{m2} = \ln(1/\acute{R}) \cdot d^{-1} \tag{5.9}$$

In this case, laser diodes (LD), working as resonators, emit light with intensity proportional to:

$$I \sim \acute{R}^2 \cdot \exp(-2\alpha_r \cdot d) \tag{5.10}$$

where the total attenuation coefficient, α_r, is a sum of attenuation on the *p-n* junction surface and on both mirror surfaces, that is,

$$\alpha_r = \alpha_{sc} + \alpha_{m1} + \alpha_{m2} = \alpha_{sc} + \ln(1/\acute{R}) \cdot d^{-1} \tag{5.11}$$

To emit enough light intensity the coefficient of light emission γ_e (called *amplification*) must exceed the total attenuation coefficient, i.e., $\gamma_e > \alpha_r$. In the case of $\gamma_e = \alpha_r$ the total current density, accounting for losses in LD, equals the total density of current in LD with an absence of losses: $J_T = J$. Otherwise, for $\gamma_e > \alpha_r$, LD works as a laser amplifier (see details in Chapter 7).

Laser diodes (LD) have a number of advantages with respect to other types of lasers, such as easy pumping by electric current injection and easy modulation by electric current injection. However, to maintain the bias at the threshold and maintain constant pulse amplitude throughout the operating life of the laser, active temperature control is used. Therefore, the biasing and drive circuitry for lasers is considerably more complex and therefore more expensive than for LEDs. Moreover, the broader bandwidths and lower coherence of LDs (with respect to LEDs) limit their usage in some applications in optical communication and LIDAR.

5.3. Photodiodes

5.3.1 The *p-n* Photodiode

As above for photodetectors, photodiode detectors use photo-generated charge carriers for their operation. A photodiode is a *p-n* junction whose reverse-biased current through the junction increases when it absorbs photons. Such photodiodes are faster than photoconductors, they do not exhibit gain. The reverse-biased *p-n* junction under light irradiation is shown in Figure 5.11 (according to [4, 9, 11]). We will notice again that the inverse-biased diode, as an electronic device, is characterized by the photo-generated current, the major carriers of which are inverse directed through

the junction, that is, are usually inverse-bandgap.

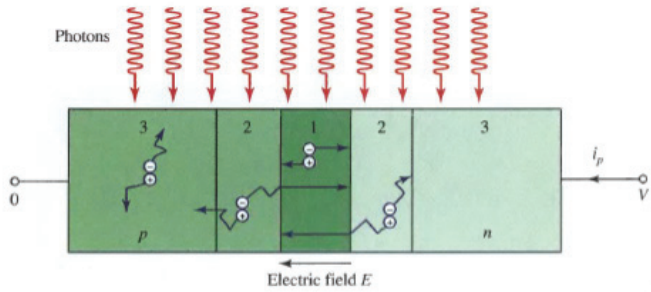


Figure 5.11. An ideal reverse-directed *p-n* photodiode irradiated by photons. The drift and diffusion regions are indicated by 1 and 2, respectively.

Photons are absorbed everywhere inside the *p-n* semiconductor. The intensity of light created by the photon stream, within the material at a depth of say, x can be described by the following exponential function

$$\Phi(x) = \Phi(0) \exp\{-\alpha x\} \tag{5.12}$$

where α is the coefficient of photons' absorbance inside the semiconductor material. It is a measure of the thickness of material required to absorb the optical radiation. For example, if $x = 2/\alpha$, 86% absorption is achieved, and if $x = 3/\alpha$, this rises to 95% [4–10].

The absorption of photons generates electron-hole pairs. Only in the presence of an ambient electric field, denoted in Figure 5.12 by the arrow, can the charge carriers be transported in the desired direction. Since only in the *depletion layer* is the electric field supported, this is a region in which photocarriers are generated: electrons and holes.

There are three possible locations of electron-hole pairs:

1. In the depletion region (1) , where electrons and holes drift in opposite directions under the influence of the electric field (see Figure 5.11), electrons move on the *n*-side, and holes on the *p*-side denoted by dashed lines in the corresponding right and left regions in Figure 5.12. The resulting reverse current occurs (directed from the *n*-region to the *p*-region). Each carrier creates an electric current pulse with gain $G = 1$ in the outer circuit.

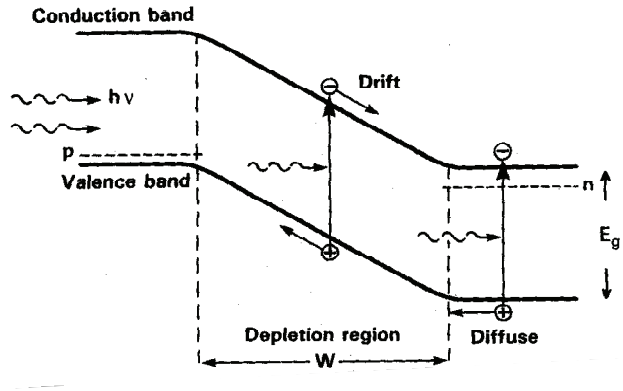


Figure 5.12. Scheme of processes occurring in *p*-*n*- photodiode: diffusion of electrons and holes inside *n*-region and *p*-region, respectively, denoted in Figure 5.11 as region 3, and their drift through the depletion region.

2. Electrons and holes in region 3 (away from region 1) cannot be transported because of the absence of an electric field (see Figure 5.11). They diffuse randomly until they recombine in the depletion region 1.

3. Outside the depletion region, but in its vicinity, in region 2 (see Figure 5.11) electrons and holes have a chance to enter the depletion layer by random diffusion. An electron from the *p*-region and a hole from the *n*-region are transported across the junction, contributing charges in the outer circuit. But the process of diffusion is slower than that for drift. In this case, the additional carrier diffusion current in the depletion region acts to enhance quantum efficiency η, which is defined by the following expression [9, 10]:

$$\eta = j_{ph} / e \cdot \Phi \tag{5.13a}$$

where j_{ph} is the density of photocurrent passing through the outer circuit, e the electron charge equaling (e = -1.6 10^{-19} C), Φ is the flux of photons entering into the diode working surface. Here we take into account that a hole has +e charge. In [9, 10], formula (5.13a) is given by the following expression

$$\eta = i_{\text{ph}} \cdot h\upsilon / e \cdot P_r \tag{5.13b}$$

Here P_r is the optical power incident on the diode surface of area S; $i_{\text{ph}} = j_{\text{ph}} \cdot S$ is the photocurrent. This leads to the definition of the photodetector *responsivity*

$$R = \text{e} \cdot \eta / h\upsilon \tag{5.14}$$

measured in units of Ampere per Watt (A/W).

Accounting for the physical processes carried out in a photodetector of width d, the quantum efficiency can be written as

$$\eta = (1\text{-}R)\zeta\,[1\text{- }\exp(\text{-}\alpha d)] \tag{5.15}$$

where, more precisely, R is the optical responsivity of the source, ζ is the fraction of electron-hole pairs that *successfully contribute to the detector current*, α is the absorption coefficient of the material [in $(\text{cm})^{-1}$], and d is the photodetector depth.

Finally, the photocurrent in the *p-n* photodetector can be determined via the responsivity of the diode as:

$$i_{\text{ph}} = R \cdot P_r \tag{5.16}$$

Response Time of Photodiodes. As was discussed above, two times play a role in the response time of photodiode detectors:

1) The transit time of carriers across the depletion layer ($t_{te} = w_d / v_e$ for electrons and $t_{th} = w_d / v_h$ for holes, where w_d is *the* width of junction, v_e and v_h are velocities of electrons and holes).

2) *RC* time response.

In photodiodes, there is an additional contribution to the response time which can occur from diffusion from region 2 to the depletion layer (see Figures 5.11 and 5.12). But this process is slower than drift.

The maximum times for this process are the carrier lifetime (τ_p for electrons in the *p*-region and τ_n for the holes in the *n*-region). The effect of diffusion time can be decreased by use of *p-i-n* diodes, which we will discuss later. In any way, photodiodes are faster than photoconductive detectors, since a strong electric field in the depletion layer causes a large velocity for the generated carriers.

Finally, we will notice that *p-n* photodetectors can be fabricated from many pure semiconductor materials [1–7, 11], as well as from compound or composite semiconductors, such as SiCr, InGaAs, GaAsP, etc.

They are constructed in such a manner that optical light falls normally to the *p-n* junction region, instead of parallel to it, as in LEDs or LDs.

5.3.2 The *p-i-n* Photodiode

A *p-i-n* diode is a *p-n* junction with an intrinsic lightly doped layer sandwiched between the *p*-region and *n*-region (see Figure 5.13). It can operate under various conditions, direct-biased (or forward-biased) and inverse-biased electronic devices with two kinds of bandgap arrangement.

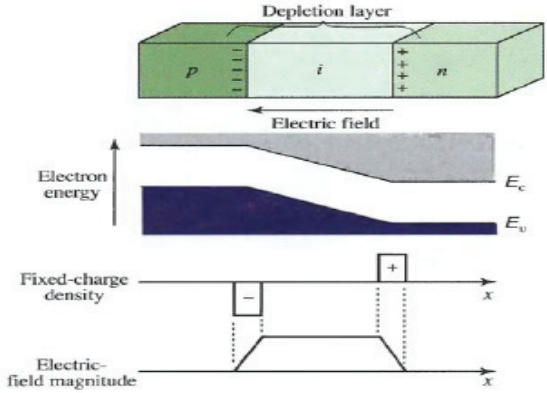


Figure 5.13. Schematic presentation of a *p-i-n* photodetector (top panel), energy distribution in valence and conductive bands with their own Fermi energies E_v and E_c, (middle panel), carriers density distribution at the wedges of two zones (middle panel), and electric field distribution inside the thick depletion layer (bottom panel).

Because the depletion layer extends into each side of a junction by a distance inversely proportional to the doping concentration, the depletion layer of the *p-i* junction penetrates deeply into the *p*-region. Similarly, the depletion layer of the *i-n* junction extends well into the *n*-region. As a result, the *p-i-n* diode can behave like a *p-n* junction diode, but with a depletion layer that encompasses the entire intrinsic region, as shown by Figure 5.13. The electron energy, density of fixed charges, and the electric field in a *p-i-n* junction diode in thermal equilibrium, as illustrated in Figure 5.14, differ with respect to those shown for *p-n* photodiodes, arising additional charges, "+" for *n*-side and "-" for *p*-side, at the wedges of the valence and conductive regions, and, therefore, resisting penetration of minor charges,

into these sides, respectively, for holes and electrons. In this case, a full current limits to zero at both wedges of the *n*-side and *p*-side of a *p-i-n* photodiode.

The simple sketch of this process occurring in a *p-i-n* photodiode and the electric field distribution versus distance along a *p-i-n* photodiode are illustrated in Figure 5.14 (according to [9, 10]).

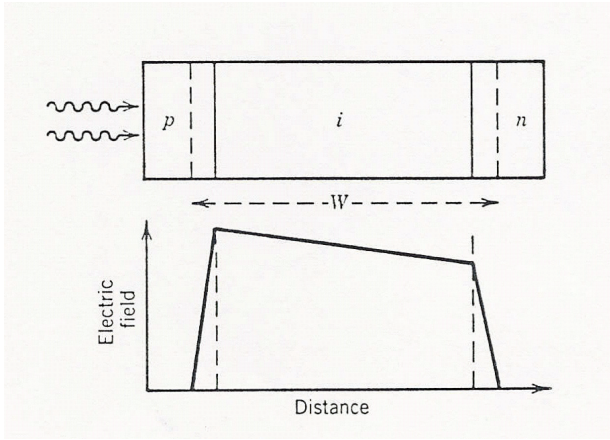


Figure 5.14. A scheme of a *p-i-n* photodiode with the width of the depletion region *W* (top panel) and electric field distribution along the *p-i-n* detector (bottom panel).

This structure serves to extend the width of the region supporting an electric field, in effect widening the depletion layer. As a detector, the *p-i-n* photodiode has a number of advantages over the *p-n* photodiode:

- Increasing the width of the depletion layer with width *W* where the generated carriers can be transported by drift, increases the area available for capturing light.

- Increasing the width of the depletion layer increases the response time of photodiode detectors, which depend on the *transit time* of carriers across the depletion layer (W_j/v_e for electrons and W_j/v_h for holes, W_j is the width of the junction, v_e and v_h *are* velocities of electrons and holes) and on *RC* time response inside the outer electrical circuit.

- Increasing the width of the depletion region reduces the junction capacitance, which determines the electrical and noise parameters of the optical detector, and thereby the *RC*

time response inside the outer circuit. But the whole transit time of carriers (electrons and holes) increases with the width W of the depletion region.

- Reducing the ratio between the diffusion length and the drift length of the diode results in a greater proportion of the generated current being carried by the faster drift process.

In Figure 5.15 (rearranged from [11]), the responsivity R (in A/W) according to (5.14) of the ideal Si photodiode (with quantum efficiency η =1) and the typical available Si photodiode are compared. The maximum responsibility is at a wavelength that is shorter than the bandgap wavelength λ_g (or frequency $v_g = c/\lambda_g$).

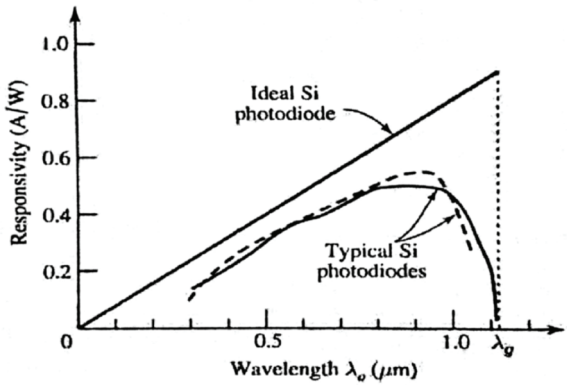


Figure 5.15. Responsivity R dependence vs. the wavelength of the *p-i-n* diode (according to [11]).

This occurs because Si is the indirect-bandgap material. The photon absorption transitions therefore take place from the valence zone to conduction-zone states that typically lie above the conduction band edge (see Figure 5.11, where the process is general for both types of photodiodes). The *p-n* and *p-i-n* detectors have many similar properties and operational parameters. Namely, a high value of external quantum efficiency in such kinds of photodetectors, described by (5.13a) or (5.13b), depends on the following [10]:

- Reducing reflections from the detector surface, achieved by the use of an antireflection coating.

- Maximizing absorption within the depletion region, which depends on device design and requires the width of the depletion region to range from W ~ 2/α to W ~ 3/α.

- Avoiding the major carrier pairs (electrons and holes) recombination, achieved through device design based on minimization of photon absorption outside the depletion region.

We also should notice that for a *p-i-n* diode, detector reverse bias is normally applied so that a wide depletion region is created, and carrier generation predominantly takes place there. Carriers are swept through by the drift field with little or no recombination. The generation of electron-hole pairs outside the depletion region relies upon the process of diffusion to drive carriers toward the junction and hence contribute to the photocurrent. In the event that electrons are generated by light radiation (e.g., by a stream of photons) and holes recombine before reaching the junction, they do not contribute to the process of total photocurrent. Hence, carrier generation outside the depletion region can lead to recombination losses and additional effects on the rise and fall of the operational time of the diode, influencing the speed and bandwidth of the *p-i-n* detector.

As was mentioned above, the width of junction W ultimately limits the transition time for electrons and holes to drift with velocities v_e and v_h across the depletion layer. Therefore, the frequency band of the detection that is inversely proportional to the response time can be estimated according to [4] as follows. For a mean transit time $<\tau>$, this frequency band equals at the 3dB detector level: $f|_{3dB} \sim 1/<\tau>$.

Moreover, as was mentioned in [10], the role of a junction that determines the depth of the depletion layer can be characterized by the detector capacitance C_D. This parameter, as was mentioned above, determines the electrical and noise parameters of the optical detector. The capacitance depends on the depletion zone width W (see Figure 5.14) and on the semiconductor material permittivity ϵ (see definitions in Chapter 2), that is, $C_D \sim \epsilon / W$.

For most *p-n* and *p-i-n* photodiodes, the total noise, which influences the forward current of major carriers inside the diode, depends both on the diode current operated in dark conditions, i_s (current in the absence of light radiation), and on the photocurrent of the diode operated in light conditions, i_{ph} (photo-generated current during light radiation). The photodiode has an i-V (total current i – voltage V) relation given by

$$i + i_{ph} + i_s = i_s \cdot \exp(e \cdot V / k_B \cdot T) \qquad (5.17a)$$

or in the form, written in [10], and depicted on the right-side of Figure 5.15 via the corresponding *i-V* dependence:

$$i = i_s \cdot [\exp(e \cdot V / k_B \cdot T) - 1] - i_{ph} \qquad (5.17b)$$

The photo-generated current i_{ph} is proportional to the photon flux Φ and is directed from the *n*-side to the *p*-side due to the outer electric field (see Figure 5.16). Here again, the Boltzmann law for carriers under temperature *T* and outer voltage *V* is available, and the constant $k_B = 1.38 \cdot 10^{-23}$ *J/K* is a Boltzmann constant, introduced after formula (5.2). As illustrated in Figure 5.16 (rearranged from [9–11]), where the generic photodiode and its *i-V* relation is presented, this is usual *i-V* behavior in any *p-n* and *p-i-n* photodiode with an added dark current (when $\Phi = 0$) and with a photocurrent proportional to Φ; V_{ph} is the voltage needed to increase the total current exponentially through the photodetector according to (5.11). As illustrated in Figure 5.16, where a generic photodiode and its *i-V* relation is presented, this is the usual *i-V* relation of a *p-n* junction with an added photocurrent according to (5.17).

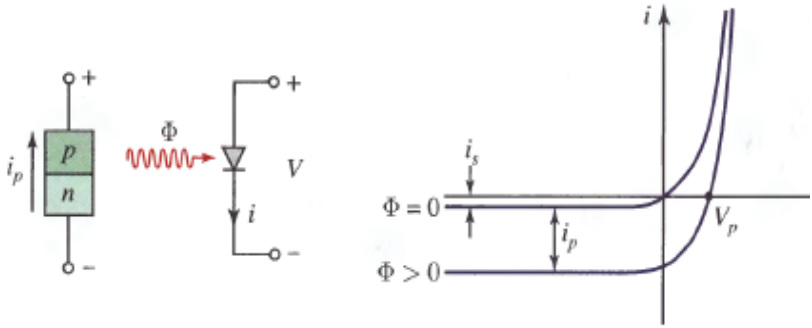


Figure 5.16. From left to right panel: a scheme of photodiode under light radiation of flow and its *i-V* relation: for $\Phi = 0$ and for $\Phi > 0$.

5.4. Multiplication of Photons – Avalanche Diodes

5.4.1 Multiplication of Photons

Multiplication of photons inside a detector can be achieved by using an *avalanche photodiode* (APD) that operates by converting each detecting photon into a *cascade* of moving carrier electron-hole pairs. So a weak light can be converted into a current enough large to be detected by

an electronic circuit following the APD. The device is configured as a strongly reverse-biased photodiode in which the junction electric field is large. The charge carriers can therefore achieve sufficient energy to excite new carriers by the process of *impact ionization*. A schematic representation of a typical electron-hole pair in the depletion region of an APD and the multiplication process by itself is presented in Figure 5.17. Let us say that a photon was absorbed by the middle side of the semiconductor (see Figure 5.17), creating an electron-hole pair (electron "-" in conduction band and hole "+" in valence band). Then, two electrons and two holes are created in the next position, as shown in Figure 5.16 (according to [9, 10]), in conductive and valence regions, 3 and 2, respectively. The holes generated at the previous and the next points also can be accelerated by an outer electric field, moving toward the right along the x-axis (see Figure 5.16), having a chance of creating an impact ionization, generating a hole-initiated electron pair at the next point, and so on.

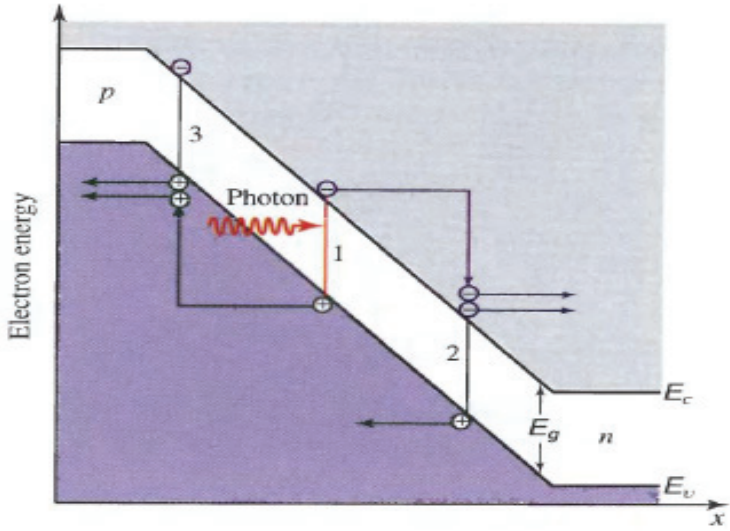


Figure 5.17. The cascade (multiplication) of electrons "-" and holes "+" passing the depletion region of a photodiode.

The process of acceleration caused by a strong electric field can be interrupted by random collisions with the lattice of the semiconductor crystal, in which electrons lose some obtained energy. This process causes electrons to reach an average saturation velocity. But, if the electron obtains energy exceeding the energy E_g of the gap (between the valence and

conductive zones) at any time during the process, it has an opportunity to generate a second electron-hole pair to impact ionization, presented at the left-side of the scheme. Then, these two electrons accelerate under the effect of a strong field, and each of them can be the source for further impact ionization.

The abilities of electrons and holes to impact ionize are characterized by the ionization coefficients for electrons, α_e, and for holes, α_h, as ionization probabilities per unit length (in cm^{-1}). The inverse coefficients $1/\alpha_e$ and $1/\alpha_h$ represent average distances between consecutive ionizations. The ionization coefficients increase with the depletion layer electric field voltage V, providing acceleration, and decrease with increasing temperature (increasing frequency of collisions with the lattice of semiconductor material that loses the energy of accelerated carriers). We will follow a simple explanation, according to which α_e and α_h are constants. On the other hand, the ionization coefficients depend on position and carrier positions, in particular their paths inside the photodetector. An important parameter for characterizing the performance of APD is the ionization ratio, which is defined as the ratio of the ionization coefficients:

$$k = \alpha_h / \alpha_e \tag{5.18}$$

When holes ionize more weakly than electrons (when $\alpha_e >> \alpha_h$, $k << 1$), most of the ionization is achieved by electrons. The avalanching process goes principally from left to right (from the *p*-side to the *n*-side of the device, see Figure 5.17, together with Figure 5.18). This process terminates when all electrons arrive at the *n*-side of the depletion region.

But if electrons and holes ionize in the same order of strength ($k = 1$), those holes that move to the left (from the *n*-side to the *p*-side) create electrons that move to the right, which in turn generate further holes moving to the left, undergoing some circulation. Despite the fact that this process increases the gain of the device (an increase of total generated charge q in the outer circuit of the detector per photo-carrier pair, q/e), there are some drawbacks of this process for several reasons:

- The avalanching process takes time and therefore reduces the device bandwidth.

- The avalanching process is random and therefore increases the device noise.

- The avalanching process can be unstable, finally causing avalanche breakdown.

It is therefore not effective to fabricate APDs from materials that use only one type of carrier (either "+" or "-") to impact ionize. If electrons are injecting carriers with $\alpha_e >> \alpha_h$, then materials where k is small are needed. If holes are major injecting carriers with $\alpha_e << \alpha_h$, then materials where k is large are needed. An ideal case can be achieved when $k = 0$ or is infinite.

The photocurrent passing such an avalanche photodiode can be found via its responsivity R, the corresponding power P_r, and the multiplication factor M:

$$i_{ph} = M \cdot R \cdot P_r \tag{5.19}$$

The relations between R (determined by Eq. (5.14)) and P_r are fully described by Eq. (5.19).

5.4.2 Avalanche Photodiodes

The same as for any photodiode, the geometry of an avalanche photodiode (APD) should maximize the photon absorption. Therefore, it usually takes the form of a *p-i-n* structure as shown in Fig. 5.13, but with some modifications as shown in Fig. 5.18 and Figure 5.19.

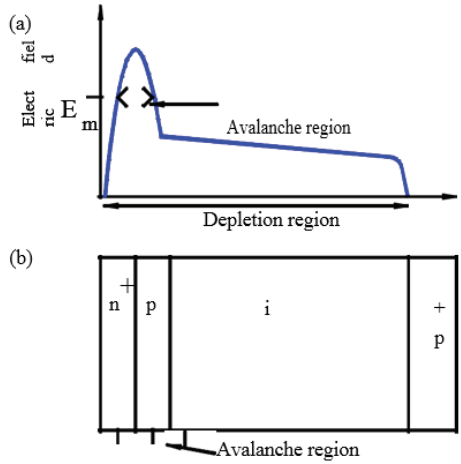


Figure 5.18. a) Electric field distribution inside regions, b) charge multiplication within the *n-p* region.

In APDs, two conflicting requirements are taken into account for their design: the *absorption and multiplication* regions must be separated (see Figure 5.18). Structures of this kind are known as *separate-absorption-multiplication* APD (SAM APD) devices, as shown by Figure 5.19.

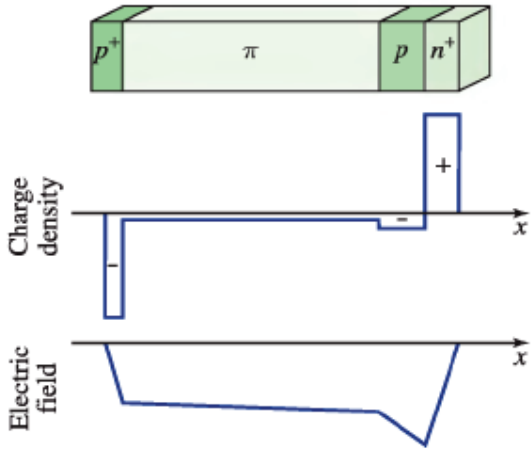


Figure 5.19. Scheme of SAM APD device.

Photons are absorbed in a large intrinsic or lightly doped region. The photoelectrons drift across this region under the influence of a moderate electric field, and then enter a thin multiplication layer with a strong electric field where avalanching occurs. The *reach through* APD structure for these purposes is illustrated in Figure 5.19. Here, photon absorption occurs in the wide π region. Electrons created by photon drift through the region into a thin *p*-*n*+ junction, where they experience a sufficiently strong electric field to cause avalanching (see the bottom panel in Figure 5.19).

The *reverse-bias* voltage applied across the device is large enough for the depletion layer to reach through the *p* and π regions into the *p*+ contact layer. So, we obtain a structure of *p*+ - π - *p* - *n*+ APD seen in the top panel of Figure 5.19. The π region is very lightly doped *p*-type material. The *p*+ and *n*+ regions are *heavily doped*. Finally, the *p*+ region collects multiplied electrons ("-"), and the *n*+ region collects multiplied holes ("+"), which create a net current in the outer circuit of the device.

On the other hand, the multiplication region should be thin to minimize the possibility of localized uncontrolled avalanches being produced by the strong electric field. The electric field uniformity can be

achieved in a thin region of the depletion layers with enough width *W*. These two conflicting requirements are taken into account for an APD design by separating the absorption and multiplication regions. Structures of these types of photodiodes, as was mentioned above, are called in the literature *separate-absorption-multiplication* APD (SAM APD) devices [7–9, 11].
Let us briefly discuss the main characteristics of LEDs.

Ionization Coefficient. The abilities of electrons and holes to impact ionize are characterized by the ionization coefficients α_e and α_h, as *ionization probabilities per unit length* [cm^{-1}]. The inverse coefficients $1/\alpha_e$ and $1/\alpha_h$ represent *average distances* between consecutive ionizations. The *ionization coefficients increase* with the *depletion layer electric field* providing *acceleration*, and decrease (*deceleration*) with increasing temperature (increasing frequency of collisions with lattice that loses energy of accelerated carriers). Presenting a simple theory according to which α_e and α_h are constants, we obtained equation (5.18) for the coefficient of ionization *k*.

When *holes ionize weaker than electrons* (when $\alpha_e >> \alpha_h$ and $k<<1$), most of the *ionization* is achieved by *electrons*. The avalanching process is going principally from *left to right* (from the *p*-side to the *n*-side of the device, see Figures 5.18 and 5.19). This process terminates when all electrons arrive at the *n*-side of the depletion region. But if electrons and holes ionize at the same order of strength ($k \sim 1$), these holes that move to the left (from the *n*-side to the *p*-side) create electrons that move to the right, which in turn generate further holes moving to the left, undergoing some circulation. Despite the fact that this process *increases the gain* of the device (an increase of total generated charge in the circuit per photo-carrier pair, *q/e*), there are some drawbacks of this process for several reasons which were mentioned above, but are repeating for more convenience:

- The avalanching process *takes time* and therefore *reduces the device bandwidth*
- The avalanching process is *random* and therefore *increases the device noise*
- The avalanching process can be *unstable*, causing *avalanche breakdown*

As was mentioned above, it is therefore not effective to fabricate APDs from materials that use only one type of carrier (or "+" or "-") to impact ionize.

If electrons are injecting carriers with $\alpha_e >> \alpha_h$, then they take materials with *k* small. If holes are injecting carriers with $\alpha_e << \alpha_h$, then they

took materials with k large. The ideal case is achieved when $k = 0$ or infinite.

Gain and Responsivity. First of all, we will consider a simple problem where *only one carrier* (let us say, the electron) is "work." In this ideal *single-carrier process*, $\alpha_h = 0$ and $k = 0$. Let $J_e(x)$ be the electric current density carried by electrons at location x as shown in Figure 5.20.

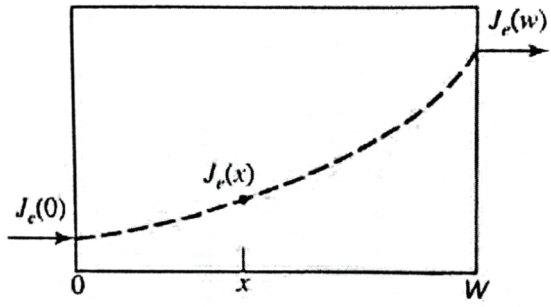


Figure 5.20. Electron current along the multiplication layer width w.

Within the distance dx, on the average, the current differential can be written as:

$$dJ_e(x) = \alpha_e \cdot J_e(x) \cdot dx \tag{5.20}$$

from which yields

$$dJ_e(x)/dx = \alpha_e \cdot J_e(x). \tag{5.21}$$

The solution of Eq. (5.21) is

$$J_e(x) = J_e(0) \exp(\alpha_e \cdot x) \tag{5.22}$$

The gain $G = J_e(\mathrm{w}) / J_e(0)$ is therefore:

$$G = \exp(\alpha_e \cdot x). \tag{5.23}$$

So, the electric current increases exponentially with the product of the *ionization coefficient* α_e and the multiplication layer width w (see Figure 5.20).

Now, a more general problem of *double-carrier multiplication* requires knowledge of both the electron current density $J_e(x)$ and hole current density $J_h(x)$. We will assume that only electrons are injected into

the multiplication region. Since hole ionizations also produce electrons, the growth of $Je(x)$ is described by a differential equation:

$$dJ_e(x)/dx = \alpha_e \cdot J_e(x) + \alpha_h \cdot J_h(x) \tag{5.24}$$

Response Time. The APDs have *additional multiplication time* with respect to other photodiodes, where the total response time is a superposition of the *transit, diffusion and RC time* constants. This additional time is called the *avalanche buildup time*. The response time for a two-carrier multiplication APD is illustrated in Figure 5.21, which follows the process of photoelectrons generated at the edge of the absorption region (point 1, left panel).

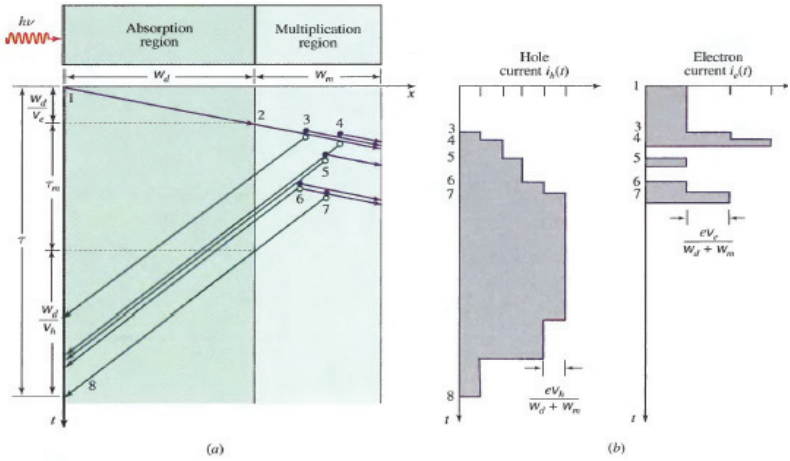


Figure 5.21. a) Schematically presented regions – absorption and multiplication in APD, b) electrons and holes current distribution in the time domain.

In Figure 5.21a, the position-time relation of the total process of avalanche for APD is presented. Blue lines represent traces of the electrons, and the green lines represent holes traces. Electrons move to the right, but holes move to the left. Electron-hole pairs are produced in the multiplication region. The process of movements is terminated when carriers reach the edge of the material. Figure 5.21b presents hole current i_h(t) and electron current i_e(t) induced in an outer circuit. Each carrier pair induces a charge e in the circuit. The total induced charge q, which is an area under the i_e(t) + i_h(t) versus t vertical axis, is $q = Ge$.

As is clearly seen, the electron drifts with a saturation velocity v_e, reaching the multiplication region (point 2) after a transit time w_d/v_e. Within this region electrons also travel with a velocity v_e. Through impact ionization it creates electron-hole pairs, say at points 3 and 4, generating additional electron-hole pairs. The holes travel in the opposite direction with their saturation velocity v_h, also creating the impact ionization resulting in electron-hole pairs as shown, for example, at points 5 and 6, the resulting carriers can cause impact ionization by themselves.

The process is terminated when the last hole leaves the multiplication region (point 7) and crosses the drift region to point 8 (see left panel, Figure 5.21a). The total time τ required for the entire process (between points 1 and 8) is the sum of the transit time (from 1 to 2 and from 7 to 8), and the multiplication time τ_m (see Figure 5.21b), i.e.,

$$\tau = w_d/v_e + w_d/v_h + \tau_m \tag{5.25}$$

Because the process of multiplication is *random*, the multiplication time τ_m is also a *random* value. In the special case k=0 (only electron multiplication, α_h =0) the maximum value of τ_m is clearly seen from Figure 5.21b (middle panel). And can be obtained as:

$$\tau_m = w_m/v_e + w_m/v_h \tag{5.26}$$

For a large gain G, and for $0<k<1$, an order of magnitude of the average value of τ_m is obtained by multiplying the first term in (5.25) by the factor $G \cdot k$, i.e.,

$$\tau_m = G \cdot k \cdot w_m/v_e + w_m/v_h \tag{5.27}$$

In this case, maximum ionization can be achieved by using the material (let us say, Si) with $k \sim 1$. Photons are absorbed in a large intrinsic or lightly doped region. The photoelectrons drift across this region under the influence of a moderate electric field, and then enter a thin multiplication layer with a strong electric field where avalanching occurs.

As for the gain, G, of APDs, its growth with an increase of the product of the ionization factor α_e on multiplication layer width w (assuming pure electron injection) depends on values of the ionization ratio k. Thus, with an increase of the parameter k from 0 to 1, one will get a sharper exponential increase of gain G both for $\alpha_e w$ lies from 0 to unit, and even after $\alpha_e w = 1$.

So, we should find the materials of interest that are closely related to those purposes by use of *p-i-n* photodiodes, accounting for the additional condition that they should have the lowest (for electron injection) or highest (for hole injection) possible value of ionization ratio *k*. Silicon APDs have k = 0.1–0.2, but Si devices with k lower than 0.006 can be fabricated, providing excellent performance in the wavelength region of 700–900 nm (i.e., visual optic and close infrared spectra).

In GaAs APDs are usually used in telecommunications (in 1300–1600 nm), having higher values of k and moderate noise electric fields ~10^5 V/cm, corresponding to tens of volts across the device, initiate the avalanche mechanism. As the reverse-bias voltage increases, the gain and dark current (e.g. noisy current) are also increased. The optimal gains for such materials are G=10 and typical dark currents are ~10^{-11} A, which is too small compared with a photo-induced current.

5.5. Operational Characteristics of Light Photodiodes

As follows from the above, the most common photodetector in electro-optical applications, fiber optic and/or wireless, is the semiconductor junction photodiode, which converts optical power to an electric current called a photocurrent. But now, instead of formula (5.19), we take into account the gain G of the diode described above. So, we get

$$i_{ph} = R \cdot G \cdot P_r \tag{5.28}$$

The cutoff wavelength is determined by bandgap energy (i.e., the depletion zone energy E_g, see Chapter 3) and is given by the following relation:

$$\lambda_g = 1.24 / E_g \tag{5.29}$$

In (5.29), the wavelength is in micrometers (μm) and the bandgap energy is in electron-volts (eV). It is clearly seen that only photons with wavelengths equal to or smaller than the cutoff wavelength can be detected (i.e., their energy should be enough to transfer electrons from the valence region to the conductive region for current generation into the outer electronic circuit consisting of the photodetector).

According to (5.28), the photodetector acts like a constant current source. Therefore, the output voltage $V = I \cdot R_L$ can be increased by increasing the load resistance R_L. However, the receiver bandwidth is not larger than

$$B_\omega = 1 / [2\pi \cdot R_L \cdot C_d] \tag{5.30}$$

Hence, by the increase of R_L, we can decrease the receiver bandwidth (to become narrowband). In (5.30), C_d is a shunt capacitance.

Now, entering into the problems and exercises for homework presented below, it can be useful to use Table 5.1 (according to [4-6, 9, 11]), listing the properties of commonly used semiconductors for optical emitters and diodes construction in electro-optical communication and LIDAR applications.

Table 5.1. Common semiconductors characteristics.

Material	Bandgap (eV)	Band	Mobility @300 K (cm^2/V-s)		Effective Mass		Dielectric Constant, $\varepsilon/\varepsilon_0$	Refractive Index @ $h\nu \approx E_g$
			Electrons	Holes	Electrons (long/trans)	Holes (heavy/light)		
C	5.47	*Indirect*	2,000	2100	1.4/0.36	1.08/0.36	5.7	—
Si	1.124	*Indirect*	1,450	505	0.92/0.19	0.54/0.15	11.9	3.5
AlN	6.2	Direct	—	14	—	—	9.14	2.7
AlP	2.41	*Indirect*	60	450	3.61/0.21	0.51/0.21	9.8	—
AlAs	2.15	*Indirect*	294	—	1.1/0.19	0.41/0.15	10	3.2
AlSb	1.61	*Indirect*	200	400	1.8/0.26	0.33/0.12	12	3.6
GaN	3.44	Direct	440	130	0.22	0.96	10.4	—
GaP	2.27	*Indirect*	160	135	4.8/0.25	0.67/0.17	11.1	3.45
GaAs	1.424	Direct	9,200	320	0.063	0.5/0.076	12.4	3.6
GaSb	0.75	Direct	3,750	680	0.0412	0.28/0.05	15.7	3.8
InN	1.89	Direct	250	—	0.12	0.5/0.17	9.3	—
InP	1.34	Direct	5,900	150	0.079	0.56/0.12	12.6	3.4
InAs	0.353	Direct	33,000	450	0.021	0.35/0.026	15.1	3.5
InSb	0.17	Direct	77,000	850	0.0136	0.34/0.0158	16.8	4.2

Moreover, some important graphically presented relations between parameters and characteristics of semiconductors are presented to be used in computations below. Thus, Figure 5.22 (according to [4-6]) presents the coefficient of absorption depending on the energy of photons and on the wavelength of the corresponding optical ray.

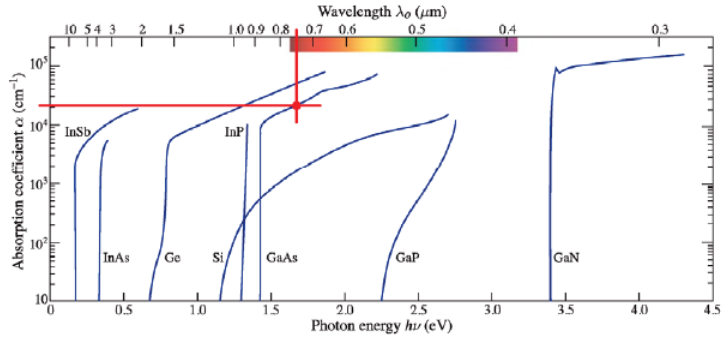


Figure 5.22. Attenuation versus photon energy for pure and binary semiconductors (according to [4–6]).

Figure 5.23 (according to [4-6, 11]) presents the coefficient of absorption of the common mono-, dual -, and poly-semiconductors

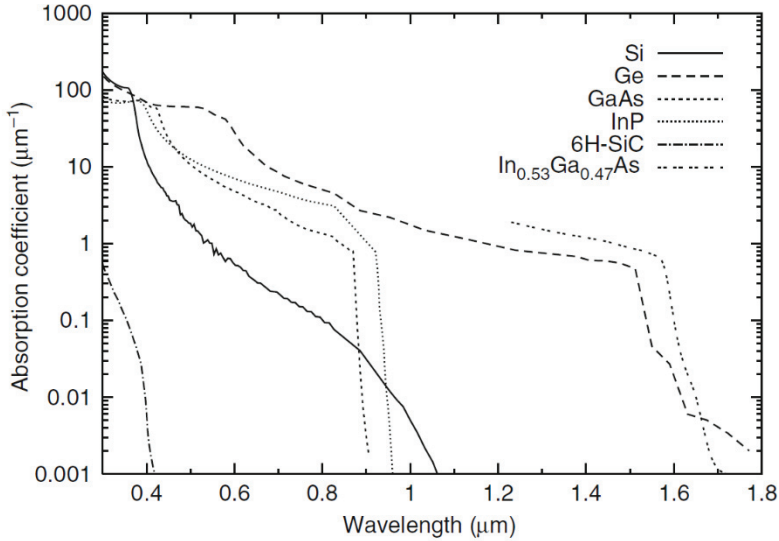


Figure 5.23. Coefficient of absorption vs. the wavelength of optical wave in semiconductor materials commonly used in electro-optics (according to [4-6, 11]).

Figure 5.24 (according to [4-6, 11]) presents the coefficient of

refraction for a composite semiconductor presented in general form $In_xGa_{1-x}As$, where the parameter x characterizes the proportion between pure semiconducting materials, In, Ga, or As, in the composite semiconductor, and so forth.

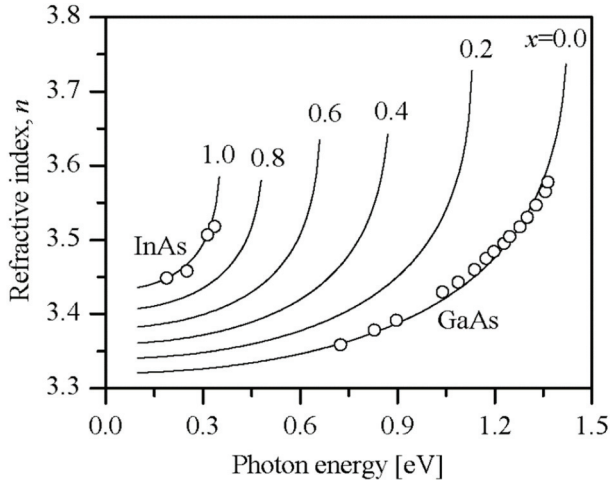


Figure 5.24. Refractive coefficient vs. the photon energy [in eV] in $In_xGa_{1-x}As$ (according to [4-6, 11].

Exercises

Exercise 1.

Given: Binary semiconductor GaAs under the illumination of photon flux with $\lambda = 0.75$ μm and power density of I = 10 W/cm^2. The time of recombination of electron-hole pairs inside p-n junction $\tau_r = 10^{-9}$ sec. Each falling photon gives rise to one electron-hole pair, i.e., the quantum efficiency $\eta_i = 1$.

Find: 1) Generation rate of electron-hole pairs, R;
2) Density of minor carriers ($\Delta n = \Delta p$) in the p-n junction.

Solution

1) The light flux generating electron-hole pairs in the p-n junction satisfies formula (5.12), where the coefficient of absorption α presented in exponent

can be found from Figure 5.22, which presents the difference of the absorption coefficient versus photon energy (in eV) for pure and binary semiconductors (see Table 3.3, Chapter 3) for a room temperature of 300 K.

As follows from Figure 5.22, for $\lambda = 0.75$ μm and for GaAs, crossing straight lines give us $\alpha = 2 \cdot 10^4$ cm^{-3}. If so, the rate of electron-hole pair generation R by photons equals

$$R = \alpha \cdot \Phi = \alpha \cdot I / h \cdot \upsilon = (2 \cdot 10^4 \text{ cm}^{-3} \cdot 10 \text{ W/ cm}^2) \cdot$$
$$\cdot [(1.24 \text{ eV}/0.75)\ 1.6 \cdot 10^{-19} \text{ eV}]^{-1} = 5.65 \cdot 10^{23} \text{ [cm}^{-3} \text{ / s]}$$

2) The density of minor carriers Δn can be found from equation

$$d(\Delta n) / dt = 0 \text{ or } R - \Delta n / \tau_r = 0$$

from which we get:

$$\Delta n = R \cdot \tau_r = 5.65 \cdot 10^{23} \text{ cm}^{-3} \text{ / sec} \cdot 10^{-9} \text{ sec} = 5.65 \cdot 10^{14} \text{ [cm}^{-3}\text{]}$$

Exercise 2.

Given: Photodiode Si has the following parameters: $n = 10^{16}$ cm^{-3}; $\lambda = 600$ nm;

$\mu_n = 1450$ cm^2 / V·s; $\mu_p = 450$ cm^2 / V·s; $n_i = 10^{10}$ cm^{-3}; n=4; $V = 5$ Volt; $A = 1$ μm^2; $d = w = 2$ μm.

Find: 1) The current passed the photodetector.
2) The reflection coefficient.

Solution

1) It is known that conductivity of a photodetector depends on the mobilities of the major carriers, electrons μ_n from n-type and holes μ_p from p-type, as well as their densities, n and p, respectively:

$$\sigma = e \cdot (\mu_n \cdot n + \mu_p \cdot p)$$

Here, from relation $n \cdot p = n_i^2$, we get:

$$p = n_i^2 / n = 10^{20} / 10^{16} = 10^4 \text{ [cm}^{-3}\text{]}$$

Because $n>>p$, we can write:

$$\sigma = e \cdot \mu_n \cdot n = 1.6 \cdot 10^{-19} \cdot 1450 \cdot 10^{16} = 2.32 \text{ [S]}$$

Finally, accounting for the Ohm's law $i = \sigma \cdot E$. and that according to the relation between field strength E and voltage V, $V = E \cdot d$, we get

$$i = \sigma \cdot E \cdot A = \sigma \cdot E \cdot V / d = 580 \text{ [μA]}$$

2) The reflection coefficient for Si material coefficient of refraction n=4 equals:

$$\acute{R} = [(n - 1) / (n + 1)]^2 = (3 / 5)^2 = 0.36$$

Exercise 3.

Given: P-N junction of the laser diode (LD) based on GaAs semiconductor with the following parameters: carriers (electron and hole) diffusion coefficients: $D_n = 20$ cm^2 / sec and $D_p = 15$ cm^2 / sec; the carriers densities equals respectively: $N_n = 5 \cdot 10^{17}$ cm^{-3} and $N_p = 5 \cdot 10^{16}$ cm^{-3}; the corresponding life-times equals: $\tau_n = 10^{-8}$ sec and $\tau_p = 10^{-7}$ sec; The bias voltage is V= 1 Volt; the area of p-n junction $A = 1$ mm^2; the total number of carriers (electrons and holes) $n_i = 2 \cdot 10^6$ cm^{-3}.

Find: 1) The total current *i.*
2) The output flux of emitted light;
3) The output power of light flux;
4) The refractive index of the mirror n_2 if the reflection coefficient Ŕ = 0.1, and the refraction index of the semiconducting material n_1 = 3.66 (for GaAs);
5) The angle of total intrinsic reflection.

Solution

1) The total current can be found by use of formulas (5.2) and (5.3), that is:

$$i_s = e \cdot A \cdot n_i^2 \cdot [(D_n / \tau_n)^{1/2} / N_n + (D_p / \tau_p)^{1/2} / N_p] = 7 \cdot 10^{-21} \text{ [A]}$$

$$i = i_s \cdot \{\exp (e \cdot V / k_B \cdot T) - 1\} = 7 \cdot 10^{-21} \exp (1 / 0.026) - 1\} = 0.35 \text{ [mA]}$$

2) The outer emitted by LD photon flux rate according to (5.6a) equals:

$$\Phi_{out} = \eta_i \cdot i / e = 0.5 \cdot i / e = 1.09 \cdot 10^{15} \text{ [photons/s]}$$

3) The output power of light emitted by LD equals energy of photons and timing of their flux rate

$$P_{out} = \Phi_{out} \cdot h \cdot \upsilon = 0.25 \cdot 10^{-3} \text{ [W]} = 0.25 \text{ [mW]}$$

Here was accounted for 1 eV = $1.6 \cdot 10^{-19}$ J and for Joule /sec = Watt.

4) Accounting for (5.8), but with n_2 not equal to unit, as at the boundary of semiconductor-air, and accounting for reflection coefficient $\acute{R} = 0.1$, we get:

$$\acute{R} = [(n_1 - n_2) / (n_1 + n_2)]^2 = 0.1$$

from which for the refractive index n_2 for given index $n_1 = 0.366$ we get:

$$n_2 = n_1 (1 - 0.316) / (1 + 0.316) = 0.19$$

5) The angle of the total intrinsic reflection from mirrors equals

$$\vartheta_c = \sin^{-1} (n_1 / n_2) = 31^0$$

Exercise 4.

Given: The light-emitted diode (LED) with the parameters: $\mu_n = 3900$ cm^2 / V·s; $\mu_p = 1300$ cm^2 / V·s; $\rho_n = 1$ s·cm; $\rho_p = 0.3$ s · cm; $\tau_n = 10^{-8}$ s; $\tau_p = 10^{-7}$ s. The total number of carriers (electrons and holes) $n_i = 2 \cdot 10^6$ cm^{-3}; the forward-biased voltage is $V = 1$ Volt; the area of p-n junction $A = 1$ mm^2. The temperature is 300K

Find: 1) The current of diffusing minor carriers.

2) The total current of carriers (electrons and holes) at the LED output.

Solution

1) According to formula (5.3)

$$i_s = q \cdot A \cdot n_i^2 \cdot [(D_n / \tau_n)^{1/2} / N_n + (D_p / \tau_p)^{1/2} / N_p]$$

the diffusion coefficients can be found according to formulas (5.4)

$$D_n = \mu_n \cdot k_B \cdot T / q = 101.4 \text{ [cm}^2 / \text{ s]}$$

$$D_p = \mu_p \cdot k_B \cdot T / q = 33.8 \text{ [cm}^2 / \text{ s]}$$

Accounting for relations (5.4b) between densities of carriers, N_n and N_p, via their motilities, μ_n and μ_n, and their partial resistivity, ρ_n and ρ_p, we get:

$$N_n = (e \cdot \rho_n \cdot \mu_n)^{-1} = 1.6 \cdot 10^{15} \text{ [cm}^{-3}]$$

$$N_p = (e \cdot \rho_p \cdot \mu_p)^{-1} = 1.6 \cdot 10^{16} \text{ [cm}^{-3}]$$

Finally, using the above numbers we get:

$$i_s = 4.1 \cdot 10^{-19} \text{ [A]}$$

2) The total current can be found by use of formula (5.2), i.e.,

$$i = i_s \cdot \{\exp (e \cdot V / k_B \cdot T) - 1\} = 21 \cdot 10^{-3} \text{ A} = 21 \text{ [mA]}$$

Exercise 5.

Given: Laser diode (LD) GaAs with absorption coefficient $\alpha_s = 20$ cm^{-1}. Coefficient of refraction $n_{GaAs} = 3.6$.

Find: 1) The coefficient of reflection.
2) The length of the p-n junction, d.
3) Full coefficient of absorption of semiconducting material and two similar mirrors with $\acute{R}_1 = \acute{R}_1 = \acute{R}$ and $\alpha_s = 30$ cm^{-1}.
4) Lifetime of the process of recombination of electron-hole pairs.

Solution

1) According to formula (5.8) we get:

$$\acute{R} = [(n_1 - 1) / (n_1 + 1)]^2 = [(3.6 - 1) / (3.6 + 1)]^2 = 0.33$$

2) Since $\acute{R}_1 = \acute{R}_1 = \acute{R}$, we get

$$\alpha_s = \alpha_m = \ln (1/\acute{R}) / d = - \ln \acute{R} / d$$

which yields:

$$d = - \ln \acute{R} / \alpha s = - \ln (0.33) / 20 = 554 \text{ [μm]}$$

3) The total coefficient of absorption from LD and two similar mirrors is:

$$\alpha_r = \alpha_s + \alpha_m = \alpha_s + \ln (1/\acute{R}) / d = 30 + 20 = 50 \text{ [cm}^{-1}\text{]}$$

4) The total time of the process is inversely proportional to the total coefficient of absorption and the velocity of photons in a semiconductor v = c / n_{GaAs}, $c = 3 \cdot 10^8$ m/s, i.e.,

$$\tau = (\alpha_r \cdot \text{v})^{-1} = (\alpha_r \cdot c / n_{GaAs})^{-1} = 3.6 / (3 \cdot 10^{10} \text{ cm/s} \cdot 50 \text{ cm}^{-1} = 12 \cdot 10^{-12} \text{ sec} = 12 \text{ ps}.$$

Exercise 6.

Given: Laser diode (LD) with the following parameters: τ = 1.25 ns, T = 300 K, $\eta_i = 0.5$; $\Delta nT = 1.25 \cdot 10^{18}$ cm^{-3} ; $\alpha = 600$ cm^{-1} . Geometrical parameters of LD (see Figure 5.9) are: $l = 2$ μm, $d = 20$ μm, and $w = 10$ μm.

Find: 1) Time of emission.
2) Current density inside the p-n junction.
3) Coefficient of emission (amplification), if the total current of major and minor carriers equals $i = 700$ mA.
4) The gain of LD.

Solution

1) Time of emission:

$$\tau_r = \tau / \eta_i = 1.25 \cdot 10^{-9} \text{ sec} / 0.5 = 2.5 \text{ [ns]}$$

2) The current density inside p-n-junction:

$$j_T = e \cdot l \cdot \Delta n_T / \eta_i \cdot \tau_r = 3.2 \cdot 10^4 \text{ [A/ cm}^2\text{]}$$

The total current density through LD

$$j = i / A = i / w \cdot d = 3.5 \cdot 10^4 \ [\text{A/ cm}^2]$$

3) The maximum coefficient of emission (amplification):

$$\gamma_p = \alpha_r \cdot (j / j_T - 1) = 56.25 \text{ cm}^{-1}$$

4) The maximum gain of LD

$$G = \exp(\gamma_p \cdot d) = 3$$

Exercise 7.

Given: Avalanche photodiode (APD) based on semiconductor Si (silica) material with $w_d = 50$ μm; $w_m = 0.5$ μm; $v_e = 10^7$ cm/s; $v_h = 5 \cdot 10^6$ cm/s; G= 100, k=0.1.

Find: 1) Response times of APD: τ_m and τ,
2) Compare the obtained response time with that for a *p-i-n* diode with the same parameters.

Solution

1) For APD from Eq. (5.26) we get

$$\tau_m = w_m / v_e + w_m / v_h = 5 + 10 = 15 \ ps.$$

From Eq. (5.25) we get

$$\tau = 500 + 1000 + 15 = 1515 \ ps = \underline{1.515 \ ns}.$$

On the other hand, Eq. (5.27) yields τ_m = 60 ps, so that Eq. (5.25) now provides

$$\tau = 1565 \ ps = \underline{1.565 \ ns}.$$

2) For a *p-i-n* photodiode with the same values w_d = 50 μm; $v_e = 10^7$ cm/s; $v_h = 5 \cdot 10^6$ cm/s, the transit time

$$\tau = w_d / v_e + w_d / v_h = \underline{1.5 \ ns},$$

which is close to 1.515 ns and 1.565 ns. This is because in the silica (Si)

APD device, the transit time through the multiplicative zone, τ_m, ranges from 15 *ps* to 60 *ps*. i.e., is too small with respect to the transit time through the absorbing zone τ. (1500 *ps*).

Problems

Problem 1.

Given: APD detector based on poly-semiconductor $In_{0.53}Ga_{0.47}As$ with the following parameters:

$$G = 20, \quad \xi = 0.5, \quad \lambda = 1.55\mu m, \quad d = 1.75\mu m$$

Find:

1. What is the detector responsivity *R*?
2. What is the current in the detector for photon outer flux $\Phi = 10^{10}\,s^{-1}$.

Problem 2.

Given: Optical detector based on composite poly-semiconductor $In_{0.5}Ga_{0.5}As$ having the following parameters

$$G = 20, \quad \zeta = 0.5, \quad \lambda_0 = 1550nm, \quad d = 1.75\mu m, \quad \Phi_0 = 10^{10}\,s^{-1}$$

Find:

1. The detector responsiveness.
2. The flux of photons.

Note: Use for these purposes Figures 5.22 to 5.24.

Problem 3.

Given: Non-semiconductor Si (silicon) with the following photovoltaic data and its characteristics:

$$\mu_n = 1450\frac{cm^2}{V\cdot s}, \quad \mu_p = 450\frac{cm^2}{V\cdot s}, \quad n = 10^{16}\,cm^{-3}, \quad \xi = 0.5, \quad \tau = 5\cdot 10^{-11}\,s$$

$$V = 5V, \quad A = 1\mu m^2, \quad d = w = 2\mu m$$

At the first stage, we consider that the silicon is not illuminated.

1. What is the current in the photodetector?

Now a flux of photons of $\Phi = 2.3 \cdot 10^{15}\,s^{-1}$ illuminates silicon at a wavelength of 600nm

2. What is the change in the current in the photodetector?
3. What is the ratio of the current with enlightenment to the current without enlightenment?
4. How can the ratio be improved?

Note: Use for computations Table 5.1.

Problem 4.

Given: A laser based on dual semiconducting material GaAs is presented at room temperature of 300K. The injected current (electron-hole pairs) is created at a rate of $10^{23}\,cm^{-3}s^{-1}$. Concentrating charges in a *p-n* junction equals $n_i = 10^{16}\,cm^{-3}$. The recombination constant, describing the recombination rate of electro-hole pairs, equals $10^{-11}\,cm^{3}s^{-1}$.

Find:

1. Concentration of holes.
2. Time of life of the process of photons creation.
3. Excess in the current carriers, electrons and ions.

Note: Use for computations Table 5.1 and Figures 5.22 to 5.24.

Problem 5.

Given: Avalanche photodiode (APD) based on semiconductor Si (silicon) material with $w_d = 35$ μm; $w_m = 0.4$ μm; $v_e = 5 \cdot 10^7$ cm/s; $v_h = 10^7$ cm/s; $G = 50$, $k = 0.3$.

Find:

1) Response times of APD: for multiplication range, τ_m, and the total time τ.
2) Compare the obtained time with that for a *p-i-n* diode with the same parameters.

Bibliography

[1] Yariv, A. 1976. *Introduction in Optical Electronics*, Chapter 5. NY: Holt, Rinehart, and Winston.
[2] Kressel, H., and J. K. Butler. 1977. *Semiconductor Lasers and Heterojunction LEDs*. NY: Academic Press.
[3] Kressel, H., ed. 1980. *Semiconductor Devices for Optical Communications*. NY: Springer-Verlag.
[4] Sze, S. M. 1985. *Semiconductor Devices: Physics and Technology*. NY: Wiley.
[5] Agrawal, G. P., and N. K. Dutta. 1986. *Long-wavelength Semiconductor Lasers*, NY: Van Nostrand Reinhold.
[6] Coldren, L. A., and C. W. Corzine 1995. *Diode Lasers and Photonic Integrated Circuits*. New York: Wiley.
[7] Marz, R. 1995. *Integrated Optics: Design and Modeling*. Norwood, MA: Artech House.
[8] Morthier, G, and P. Vankwikelberge. 1997. *Handbook of Distributed Feedback Laser Diodes*, Norwood, Ma: Artech House.
[9] Palais, J. C. 2006. "Optical communications." In Handbook: *Engineering Electromagnetics Applications*, edited by R. Bansal. NY: Taylor and Frances.
[10] Blaunstein, N., S. Engelberg, E. Krouk, and M, Sergeev. 2020. *Fiber Optic and Atmospheric Optical Communication*. Hoboken, New Jersey: Wiley.
[11] Saleh, B. E. A., and M. C. Tech. 2012. *Fundamentals of Photonics,* 2nd ed. New York: Wiley & Sons, Inc.

CHAPTER 6

NOISE IN LIGHT EMITTERS AND DIODES

We will start to analyze different types of noise occurring in the light sources (e.g., lasers) and detectors (e.g., diodes), as the initial and the later terminals of any optical communication link, wired (e.g., fiber optic) and wireless (e.g., atmospheric). Noise occurring in fiber optics will be discussed later.

6.1. Noise in Photodiodes and Light Emitters

As mentioned in [1–7], noise is a fundamental characteristic of all kinds of photodetectors and optical sources that characterizes the photoconductive process. Here we briefly introduce the reader to some main kinds of noise occurring inside each photodiode, working as a source or a detector, mentioned above, and will describe their main operational characteristics. Thus, as was shown in Chapter 5, any photodetector is responsive to photon flux Φ, and therefore, on optical power $P=h\nu\Phi$. This flux gives rise to a proportional photocurrent $I_{ph} = \eta \cdot e \cdot \Phi = R \cdot P$, where η is quantum efficiency, and R is responsivity of the photodetector. However, the electric current generated in the device is a random quantity I, whose value fluctuations, determined as *noise*, around the average value $\langle I \rangle$, characterized by a standard deviation σ_I^2 or variance $\sigma_I = [\langle (I - \langle I \rangle)^2 \rangle]^{1/2}$. For zero-mean photocurrent fluctuations $\langle I \rangle = 0$, the standard deviation can be reduced by use of the root mean square (*rms*) definition, $\sigma_I = [\langle I^2 \rangle]^{1/2}$.

We will now briefly describe the types of noise that can corrupt the optical signal data recorded by photodetectors and lead to fading phenomenon and data bit errors.

Photon Noise. This noise is related to the random arrival of photons themselves and can be described by *Poisson statistics* [1–5].

Let us suppose the existence of an assembly of n atoms, and the probability of any one of them emitting a photon in time τ is p. Then, the average number of photons detected in this time would be np, but the actual

number for r photons will vary statistically around this mean according to the Poisson law with probability

$$P_r = \exp(-np)\frac{(np)^r}{r!} \tag{6.1}$$

For example, the probability of receiving zero photons is exp(- np), and two photons is exp(- np)·(np) ^{2} / 2!, and so forth.

We can relate np to the mean optical power received by the detector, P_m, for np is just the mean number of photons received in time τ. Hence

$$P_m = np\frac{h\nu}{\tau} \tag{6.2}$$

and then the mean of the Poisson distribution, i.e., the mean number of photons, becomes [1–3]

$$np = \frac{P_m\tau}{h\nu} = \frac{P_m}{h\nu B_\omega} \tag{6.3}$$

where B_ω is the detector bandwidth.

Now we need to measure the spread from this mean, which is called the variance, which, according to the Poisson law, is equal:

$$\sigma_n^2 = np \tag{6.4}$$

Finally, we obtain the variance as a measure that gives us the noise of the optical signal. This noise is usually called *quantum noise* or *photon noise*, the power of which equals:

$$N \equiv \sigma_n = \left(\frac{P_m}{h\nu B_\omega}\right)^{1/2} \tag{6.5}$$

Consequently, the signal-to-noise ratio (SNR or S/N) will be [9, 10]

$$SNR = \frac{P_m}{B_\omega h\nu}\cdot\frac{1}{N} = \left(\frac{P_m}{h\nu B_\omega}\right)^{1/2} \tag{6.6}$$

This is an important result since it provides the ultimate limit on the accuracy with which a light power level can be measured by a laser detector or photodetector. We should notice that the measurement accuracy improves as $(P_m)^{1/2}$, and for lower power, the accuracy will be poor enough. Correspondingly, if frequency ν of photon emission is larger for a given power, the accuracy becomes worse.

Finally, it must be stated that these conclusions only apply when the probability of photon emission is small enough. The results obtained above are no longer valid for intense laser beams of power density $w > 10^6$ W/m^2. Such light is sometimes said to have non-Poisson statistics [1–6].

Photoelectron Noise. Since in the process of generation of a photon, an electron-hole pair is random and going with probability $1 - \eta$, it is a source of noise, η is quantum efficiency, as introduced above. An incident photon on a photodetector with quantum efficiency η creates the electron-hole pair or liberates a photoelectron, with probability η or fails to do so with probability $1 - \eta$. The carriers are selected at random from the photon stream. An incident mean photon flux Φ (photons/s) therefore results in a mean photoelectron flux $\eta\Phi$ (photoelectrons per second). The number of photoelectrons n_{ph} detected in the time interval τ is random

$$< n_{ph} > = \eta < n > = \eta \cdot \Phi \cdot \tau \tag{6.7}$$

Assuming, as above, that photons are distributed according to Poisson law, then the photoelectron-number variance, which describes the electron noise, is equal $< n_{ph} >$, that is,

$$\sigma_m^{\,2} = < n_{ph} > = \eta < n > \tag{6.8}$$

It is seen that the photoelectron noise differs from the photon noise [compare (6.8) and (6.4)]. Accounting for the photon noise, as a fundamental noise when using light to transmit signals through the detector, we can easily determine the photoelectron SNR as:

$$\text{SNR} = < n_{ph} > = \eta \cdot < n > \tag{6.9}$$

The minimum-detectable photoelectron number is $< n_{ph} >$ equals one photoelectron, corresponding to $1/\eta$ photons. If so, it can be easily shown that for SNR = 10^3 (or for SNR = 30 dB) the receiver sensitivity equals 10^3

photons per second or $10^3/\eta$ photoelectrons per second.

Generation - Recombination Noise. The generation- recombination noise arises from fluctuations in the generation and recombination rates of electron-hole pairs due to the process of photoemission, its spectral density can be presented according to [4] as

$$N_{ph} = 4 \cdot e \cdot G \cdot I_{ph} / [1 + 4 \cdot \pi^2 \cdot f^2 \cdot \tau_r^2] \tag{6.10}$$

where τ_r is a mean electron-hole recombination time, and f is 3dB-bandwidth that can be defined as [4]

$$f|_{3dB} = 1 / [2\pi \cdot G \cdot t_T] \tag{6.11}$$

where t_T is the detector transit time.

Photocurrent Noise. When induced in a circuit a random photoelectron stream with mean $\eta \cdot \Phi$ results in a stream of current pulses with amplitude a_e and time τ_d in the outer electric circuit of the photodetector, which add together to constitute the photocurrent $I(t)$. The randomness of the photon stream is transformed into a fluctuating electric current. If, as above, the incident photons are Poisson distributed, these fluctuations are known as *shot noise* [4, 5, 11]. Let us consider that the random number of photoelectrons counted within a characteristic time interval, $T_r = 1/2f$, called the *resolution time* of the circuit [4], generates a photocurrent $I_{ph}(t)$, where t is the current time following the interval T_r. For rectangular current pulses of direction T_r, the current and the photoelectron number random variables are related by $I_{ph} = (a_e / T_r) \cdot < n_{ph} >$. The photocurrent mean and variance are therefore given by:

$$< I_{ph} > = a_e \cdot < n_{ph} > / T_r \tag{6.12}$$

and, finally, the noise introduced by a photocurrent inside the detector equals

$$\sigma_i^2 = (a_e / T_r)^2 \cdot \sigma_m^2 \tag{6.13}$$

where, again, $< n_{ph} > = \eta \cdot \Phi \cdot T = \eta \cdot \Phi / B$ is a mean number of photoelectrons collected in the resolution time $T_r = 1/2B$; B is the bandwidth, and σ_m^2 is defined by (6.8).

Gain Noise. The photocurrent mean and variance for a device with fixed (*deterministic*) gain G is determined by the generated pulse $q = G \cdot a_e$. In this case, the mean photocurrent can be written as:

$$< I_{ph} > = a_e \cdot G \cdot \eta \cdot \Phi = a_e \cdot G \cdot \eta \cdot P / h \cdot \nu \tag{6.14}$$

and the corresponding variance is

$$\sigma_G{}^2 = 2 \cdot a_e \cdot G \cdot < I_{ph} > B = 2 \cdot a_e{}^2 \cdot G^2 \cdot B \cdot \eta \cdot \Phi \tag{6.15}$$

The SNR then equals:

$$SNR = < I_{ph} > / (2 \cdot a_e \cdot G \cdot B) = \eta \cdot \Phi / 2 \cdot B = < n_{ph} > \tag{6.16}$$

Now, when G is a *random variable*, the derivation is more complicated, and we present here only their modified formula. First, instead of the above equations, we account for $G = <G>$ and will introduce F as the *excess noise factor*:

$$F = <G^2> / <G>^2 = 1 + \sigma_G / <G>^2 \tag{6.17}$$

Finally, we get

$$\begin{aligned} SNR &= <I>^2 / \sigma_G{}^2 = <I> / [2 a_e \cdot <G> \cdot B \cdot F] = \\ &= \eta \cdot \Phi / (2 \cdot B \cdot F) = < n_{ph} > / F \end{aligned} \tag{6.18}$$

The difference between formulas (6.16) and (6.18) is the existence in the denominator of the noise factor F, an increase of which decreases SNR at the output of the optical detector or source.

Thermal Noise. The thermal noise (called *Johnson noise* or *Nyquist noise* [1–4]), occurring in the outer *electric circuit*, consisting of any photodetector or laser emitted source, is the last noise that must be taken into account, when we discuss the terminal assembled in both ends of a wired or wireless communication link (see Chapter 1). It arises from the random motions of mobile carriers in the resistive electrical material at finite temperature T giving rise to a random electric current $I(t)$. Even in the absence of an external electrical power source the thermal electric current at the bulk resistance of the photoconductor or laser emitted source, R_s, is a random function $I(t)$ whose mean value $<I(t)> = 0$. The Johnson noise spectral density is directly proportional to the absolute temperature T (in

Kelvin, K) via the Boltzmann constant $k_B = 1.38 \cdot 10^{-23}$ J / K, and inversely proportional to the bulk resistance of the photoconductor or laser emitted source, R_s, that is,

$$N_T = 4 \cdot k_B \cdot T / R_s \tag{6.19}$$

If we again take into account the Boltzmann statistics according to (5.7) [see Chapter 5], we immediately obtain the variance of the circuit current, σ_r^2 (for $B << k_B \cdot T / h$) as

$$\sigma_r^2 = 4 \cdot k_B \cdot T \cdot B / R \tag{6.20}$$

It is clear that the thermal noise increases with the temperature T.

Circuit Noise. Additional noise is observed inside a photodiode circuit in the form of a random electric current i_r of Gaussian probability distribution with zero-mean and variance σ_q^2. Within a time interval T, the accumulated charge $q = i_r \cdot T / e$ (units of electrons) has an RMS value $\sigma_q = \sigma_r \cdot T/e$. The parameter σ_q, called the *circuit – noise parameter*, depends on the receiver bandwidth B. The total accumulated charge per bit $s = \eta + q$ (units of electrons) is the sum of a Poisson random variable η and independent Gaussian random variable q. Its mean is the sum of the averages:

$$\mu = <\eta> = \eta \cdot <n> \tag{6.21}$$

Its variance is the sum of the variances:

$$\sigma^2 = <\eta> + \sigma_q^2 \tag{6.22}$$

For large $<\eta>$, the Poisson distribution can be approximated by the Gaussian one, with mean μ and variance σ^2 (see above). According to this approximation, which is valid mostly for avalanche photodiode (APD) [see definition in Chapter 5] of gain $<G>$, the mean number of photoelectrons amplified by factor $<G>$, but with additional noise introduced in the amplification process, finally, we get for the mean of the total collected charge per bit $s = \eta + q$ (units of electrons):

$$\mu = <\eta> \cdot <G> \tag{6.23a}$$

$$\sigma^2 = <\eta> \cdot <G>^2 \cdot F + \sigma_q^2 \tag{6.23b}$$

where F is the excess noise factor.

Finally, we should emphasize that the above does not illuminate special aspects regarding types of semiconducting materials and special engineering techniques of light diodes and sources design. For precise and extensive information on light sources and detectors, the reader is referred to the corresponding works [1–6].

The simplest measure of the quality of any detection and recording optical device is the signal-to-noise ratio (SNR).

To find the SNR in a noiseless circuit, we should divide the variance of the total input current by the sum of variances of the constituent sources of noise written above:

$$SNR = \langle I \rangle^2 / [2 \cdot a_e \cdot \langle G \rangle \cdot B \cdot F + \sigma_r^2] = \\ = (a_e \cdot \langle G \rangle \cdot \eta \cdot \Phi)^2 / [2 \cdot a_e \cdot \langle G \rangle^2 \cdot \eta \cdot \Phi \cdot B \cdot F + \sigma_r^2] \quad (6.24)$$

In the denominator of (6.24), the first term represents photoelectron and gain noise, the second one represents circuit noise. For the optical detector or source without gain and having resistance against noise, we can deduce formula (6.24) introducing in it $\langle G \rangle = 1$ and $F = 1$. We notice that in [4], for characterizing the circuit noise, another parameter was introduced: $\sigma_q = \sigma_r / 2B \cdot a_e$. Accounting now for the relations

$$\langle n_{ph} \rangle = \eta \cdot \Phi \cdot T_r \quad (6.25a)$$

and

$$T_r = 1/(2 \cdot B) \quad (6.25b)$$

allows us to rewrite (6.24) in more compact form:

$$SNR = (\langle G \rangle^2 \cdot \langle n_{ph} \rangle^2) / [\langle G \rangle^2 \cdot F \cdot \langle n_{ph} \rangle + \sigma_q^2] \quad (6.26)$$

The SNR for an optical receiver described above has a simple interpretation. The numerator is the square of the mean number of multiplied photoelectrons detected in the receiver at resolution time T_r = 1/2B. The denominator is the sum of the variances of the number of photoelectrons and the number of circuit noise electrons collected in time T_r. For $\langle G \rangle = F = 1$, for the noiseless receiver in the absence of gain yields:

$$SNR = (\langle n_{ph} \rangle^2) / [\langle n_{ph} \rangle^2 + \sigma_q^2] \quad (6.27)$$

As was mentioned in [4], this formula is useful for a resistance limited optical receiver with a temperature of T = 300 K, when $\sigma_q \sim B^2 / 100$ (bandwidth B in Hz).

Now, we will introduce the circuit noise parameter at room temperature T = 300 K, the gain noise $\sigma_q = \sigma_r / 2B \cdot e$, which for a resistance limited optical receiver can be simplified as (for B in Hz): $\sigma_q \sim B^{1/2} / 100$. Again, if B=100 MHz (T = 300 K), then σ_q ~100. For B ranging from 100 MHz to 2 GHz, σ_q typically ~500, provided that the corresponding transistors have optimal biased conditions between the resistivity and transistor.

As follows from a general formula (6.24), SNR depends on all photo-electrical processes occurring in optical sources and detectors, namely on the photon flux and quantum efficiency of excited photons, on the type of receiver and photo emitter, LED or avalanche (APD), and amplifier (see Chapters 5). For practical applications in optical communication and optical radars (LIDAR), designers mostly deal with the dependence of SNR on the bandwidth of the optical device. For example, for a *resonance resistor*, SNR~1/B, whereas for an amplifying receiver with *bipolar transistor*, SNR~$(B + s \cdot B^2)^{-1}$, and for an amplifying receiver with *forward-emitted transistor* (FET), SNR~$(B + s \cdot B^3)^{-1}$, where s is a constant defined empirically. These relations are illustrated in Figure 6.1 for all three kinds of receivers (according to [4]).

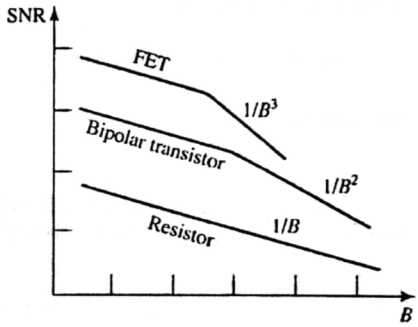


Figure 6.1. A plot of SNR vs. bandwidth B in logarithmic scale for three types of receivers (according to [4]).

The SNR always decreases with increasing B. For sufficiently small bandwidths, all three receivers exhibit an SNR that varies as B^{-1}. For large bandwidths, the SNR of the FET and bipolar transistor-amplifier receivers declines more sharply with bandwidth with respect to the resistor-limited receiver.

Finally, we should emphasize that the above has not illuminated special aspects regarding types of semiconducting materials. For precise and extensive information on light sources and detectors, the reader is referred to the corresponding works [1–7].

6.2. Noise in Optical Receivers

Noise inside Photodetector. In photodetectors, the noise arises from two kinds of noise: Johnson noise associated with the thermal noise from the bulk resistance of the photodiode slab described above by (6.8) and generation-recombination noise described above by (6.10).

Noise inside Optical Receivers. As was mentioned in Chapters 1 and 5, the optical receiver comprises the photodiode, a bias circuit, forward or inverse, a preamplifier and filtering. It can be depicted by the corresponding equivalent electronic circuit, as shown in Figure 6.2 (according to [5]). This equivalent circuit is similar for a *p-i-n* diode, avalanche photodiode (APD), and photoconductor-biased receivers (see definitions in Chapter 5).

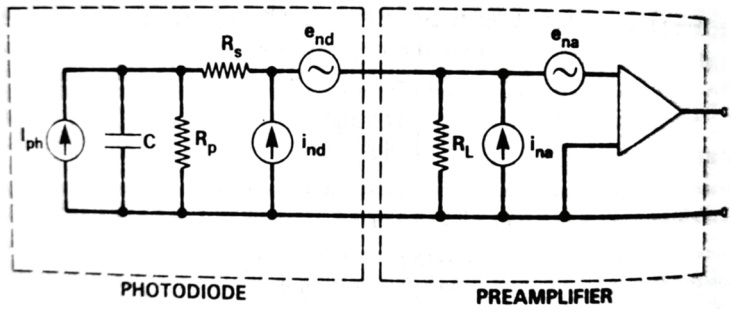


Figure 6.2. Equivalent electronic circuit of an optical receiver: photodiode (input) and preamplifier (output) (according to [5]).

Now we will summarize these types of noise for most detectors by representing such kinds of noise via the corresponding shunt noise current generators and the series noise voltage generators, as shown in Figure 6.2.

Figure 6.2 depicts, in addition to the equivalent noise current photodetector, i_{ph}, also the equivalent noise current generators, i_{nd}, and, i_{na}, and the equivalent noise voltage generators, e_{nd}, and, e_{na}, the equivalent noise current generator for the detector and preamplifier, respectively. The photodetector noise was discussed above.

1. Thus, for a *p-i-n* photodetector the noise current spectral density:

$$<i_{nd}^2>/B = 2\cdot e\cdot (I + 2\cdot I_D + I_{ph}) \tag{6.28a}$$

and the noise voltage spectral density:

$$<e_{nd}^2> / B = 4\cdot k_B \cdot T \cdot B / R_s \tag{6.28b}$$

2. For avalanche photodiode the noise current spectral density:

$$<i_{nd}^2> / B = 2\cdot e\cdot (I + 2I_D + I_{ph})\, M^2 \cdot F(M) \tag{6.29a}$$

and the noise voltage spectral density:

$$<e_{nd}^2> / B = 4\cdot k_B \cdot T \cdot B / R_s \tag{6.29b}$$

3. For photoconductor the noise current spectral density:

$$<i_{nd}^2> / B = 4\cdot k_B \cdot T \cdot B / R_p + 4\cdot e \cdot I_{ph} \cdot G / (1 + 4\cdot \pi^2 \cdot f^2 \cdot \tau_c^2) \tag{6.30}$$

Here, as above, T is the temperature (in Kelvin), k_B is the Boltzmann constant defined from the beginning, B is the rate of signals inside the receiver, G is the gain of the detector, M is the parameter of multiplication of the avalanche diode called the *average gain* (see Chapter 5), $F(M)$ is the excess noise factor [this parameter depends on material and junction characteristics via the ionization coefficient, and the nature of electron and hole injection, see Chapter 5], R_p is the photoconductive resistance of photodetector, τ_c is a mean lifetime of major carriers (see Chapter 5), and R_s is the bulk resistance of photoconductor or laser emitted source.

As for an amplifier, the corresponding equivalent circuit representation was proposed in [5, 6] and is presented in Figure 6.3, where R_F denotes the feedback resistance.

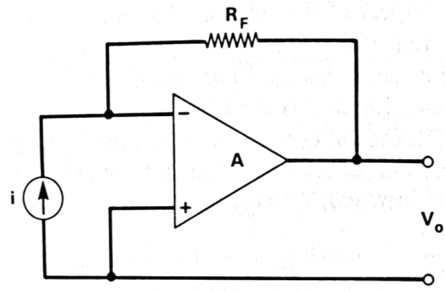


Figure 6.3. Circuit representation of preamplifier with feedback resistor R_F and amplifier gain A (according to [5, 6]).

The amplifier gain A shown in Figure 6.3 relates to the receiver bandwidth B as:

$$A = 2\pi \cdot B \cdot C_T \cdot R_F \tag{6.31}$$

where C_T is the total capacitance of the receiver, and R_F is the feedback resistor depicted in Figure 6.3.

The SNR of a digital receiver with preamplifier described by the equivalent circuit shown in Fig. 6.3, can be expressed in the following form [6]:

$$\text{SNR} = M^2 \cdot R^2 \cdot P_r^2 / [(S_1 + 4 \cdot k_B \cdot T / R_L) \cdot K_2 \cdot B + \langle I_c^2 \rangle] \tag{6.32}$$

In this expression, S_1 refers to the noise current spectral density for the detector, the second term describes the thermal noise associated with the bias resistor R_L. Here also $\langle I_c^2 \rangle$ represents the noise contribution from the preamplifier, which can be related to the noise current generation i_{na} and the noise voltage generation e_{na} (see Figure 6.2).

The quantity K_2 is a dimensionless parameter and denotes a noise integral defined in such a way that, at a data rate B, the product $K_2 \cdot B$ represents the effective receiver noise bandwidth.

As for $\langle I_c^2 \rangle$, it is described in many works, which will not be repeated here. Instead, we give its form containing the dominant noise term for the case of a preamplifier, which was evaluated in [7]:

$$\langle I_c^2 \rangle = 2 \cdot e \cdot I_b \cdot K_2 \cdot B + 2e \cdot I_c \cdot (2C_T)^2 \cdot B^2 \cdot K_3 / g_m^2 \tag{6.33}$$

The noise integral parameter, K_3, which depends on the input pulse shapes at the receiver, was described and evaluated in [7], and refers to the basic I_b and collector I_c current by bipolar transistors, respectively, and is the trans-conductance of the field-effect-transistor (FET) of GaAs semiconductor (see Table 5.1 in Chapter 5), which is usually used for the fabrication of bipolar transistors.

Now, following [5], we can present the noise current spectral density and the noise voltage spectral density for the noise generators i_{na} and e_{na}, respectively, depicted in Figure 6.2 for a preamplifier, that is:

$$<i_{na}^2>/B = 2 \cdot e \cdot I_G \tag{6.34}$$

and

$$<e_{na}^2> / B = (4 \cdot k_B \cdot T \cdot \Gamma / g_m) / (1 - f_k / f) \tag{6.35}$$

In (6.34), the FET of GaAs gate leakage noise is described by means of short noise of the leakage current I_G. The channel thermal noise in (6.35) is described by use of the FET trans-conductance g_m, as well as the empirical factor Γ close to unity for GaAs FET semiconductor [5]. In the denominator of (6.35) the second term ~ $1/f$ relates to the FET channel noise, which is characterized by a corner frequency f_k in the receiver spectrum [5].

Bibliography

[1] Marz, R. 1995. *Integrated Optics: Design and Modeling*. Norwood, MA: Artech House.

[2] Morthier, G, and P. Vankwikelberge. 1997. *Handbook of Distributed Feedback Laser Diodes*. Norwood, Ma: Artech House.

[3] Palais, J. C. 2006. "Optical communications." In Handbook: *Engineering Electromagnetics Applications*, edited by R. Bansal. NY: Taylor and Frances.

[4] Saleh, B. E. A., and M. C. Tech. 2012. *Fundamentals of Photonics*, 2nd Ed. Norwood, Ma: Artech House.

[5] Debney, B. T., and A. C. Carter. 1988. "*Optical Detectors and Receivers*." In *Optical Fiber Sensors: Principles and Components*, Vol. 1, 107–149. Norwood, Ma: Artech House.

[6] Haykin, S. 1983. *Communication Systems*, 2nd Ed., Chapter 9. New York: Wiley.

[7] Muoi, T. V. 1984. "Receiver design for high speed optical fiber systems." *IEEE J. Lightwave Technol*. 2(3), pp. 243–267.

CHAPTER 7

OPTICAL AMPLIFIERS

7.1. Principles of Optical Amplification

Given the photon material interaction processes described above in Chapter 4, it is obvious that *only stimulated emission* can lead to optical amplification. In the system depicted by Figure 7.1, stimulated emission competes with *absorption* to determine whether the incident beam is *amplified* or *attenuated*. Spontaneous emission results in background light emitted randomly into a 4π-*steroidal* sphere, a proportion of which reaches the detector as background noise. So, simulated emission gives an impact in terms of noise and does not play a positive role in optical signals with data amplification and transmission along the link [1–6].

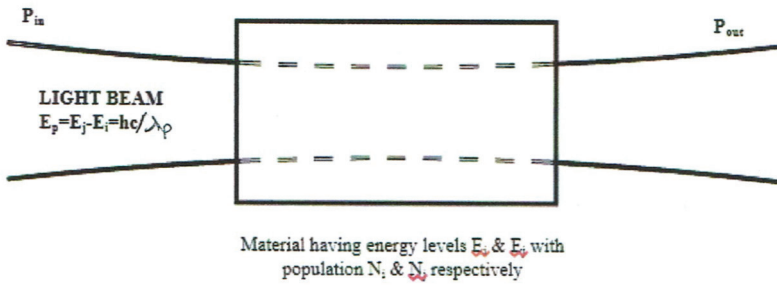


Figure 7.1. Interaction of light beam with any material-filled body.

As was shown in Chapter 4, to achieve *amplification*, the rate of stimulated emission must be greater than the rate of absorption, i.e., $B_{ji}N_j > B_{ij}N_i$, which generally means that $N_j > N_i$. This situation ($N_j > N_i$) is referred to as a *population inversion* since at thermal equilibrium, the populations are highest for lower order states as defined by the Boltzmann distribution. For example, the relative population, N_j/N_i, at 295 K = 22 ºC of two states differing in energy equivalent to the photon energy of light at $\lambda = 1\ \mu m$ is $6.0 \cdot 10^{-22}$ Joule (assuming $g_j = g_i$, see Chapter 4). Hence, at thermal

equilibrium *absorption* completely dominates and the beam is attenuated.

Let us first consider the process of *stimulated emission* in a slab of material of thickness Δz in the body of the gain medium (see Figure 7.2, according to [6-8]).

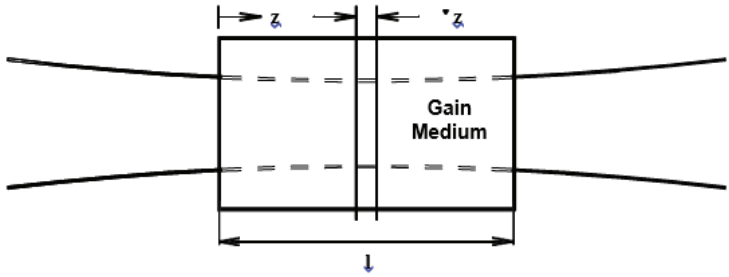


Figure 7.2. Stimulated emission in the slab filled by the gain medium (according to [6-8]).

Using the rate equations obtained in Chapter 4, we can derive expressions in terms of the population densities of the states involved and the incident light intensity.

7.2. Amplification with Small Signal Gain

The photon energy density in the slab is $\rho_\upsilon(z)$ and the incident light intensity is $I_\upsilon(z)$ (where $\rho_\upsilon(z) = I_\upsilon(z) \cdot n/c$, see Chapter 5). Applying (4.4) and (4.5) presented in Chapter 4, the rate of reduction of the number of atoms in energy level E_j due to *stimulated emission* in the slab is [6–8]:

$$(dN_j/dt) \cdot A \cdot \Delta z = B_{ji} \cdot N_j \cdot I_\upsilon(z) \cdot (n/c) \cdot g(\upsilon) \cdot A \cdot \Delta z \qquad (7.1)$$

Each transition adds a photon of energy $h\upsilon$ to the beam. Hence, we simply multiply by $h\upsilon$ to get an expression for the incremental power, ΔP, added to the beam by *stimulated emission* in the slab, and divide by the cross-sectional area to get the incremental intensity, $\Delta I_\upsilon(z)$ [6–8]:

$$\Delta I_\upsilon(z) = -(dN_j/dt) \cdot h \cdot \upsilon \cdot \Delta z = B_{ji} \cdot N_j \cdot I_\upsilon(z) \cdot (n/c) \cdot g(\upsilon) \cdot h \cdot \upsilon \cdot \Delta z \qquad (7.2)$$

Similarly, each *absorption transition* from state *i* to state *j* annihilates a photon from the beam and by analogy we can derive the incremental

reduction in intensity, - $\Delta I_\upsilon(z)$, due to *absorption* as [6–8]:

$$\Delta I_\upsilon(z) = (dN_i/dt) \cdot h \cdot \upsilon \cdot \Delta z = B_{ij} \cdot N_i \cdot I_\upsilon(z) \cdot (n/c) \cdot g(\upsilon) \cdot h \cdot \upsilon \cdot \Delta z \qquad (7.3)$$

In addition, we must consider the contribution of *spontaneous emission* from the atoms of the slab to the total radiation field. Since spontaneous emission is random and omnidirectional, only a fraction of the light emitted by any element of the slab is collected at the detector, that fraction being $\Omega/4\pi$, where Ω is the solid angle subtended by the detector at the plane of the slab.

Hence, the incremental intensity provided by the slab to the detected beam from *spontaneous emission* is:

$$\Delta I_\upsilon(z) = (dN_j/dt) \cdot (\Omega/4\pi) \cdot h \cdot \upsilon \cdot \Delta z = A_{ji} \cdot N_j \cdot (\Omega/4\pi) \cdot h \cdot \upsilon \cdot \Delta z \qquad (7.4)$$

In the formulation of Eq. (7.4), it is assumed that *spontaneous emission* contributes over the entire line shape function of the atomic transition. If a narrow band filter is used in front of the detector to reduce the level of spontaneous emission, then we must multiply (7.4) by $g(\upsilon)\cdot\Delta\upsilon$, where $\Delta\upsilon$ is the *linewidth* of the filter (see Chapters 4 and 5).

The total contribution of the slab to the signal intensity is simply the summation of the incremental intensities contributed by *stimulated* and *spontaneous emission* minus the *absorbed intensity*. Hence, the rate of change of intensity with distance z through the gain medium is given by:

$$(dI_\upsilon/dz) = (h \cdot \upsilon/c) \cdot g(\upsilon) \cdot n \cdot [B_{ji} \cdot N_j - B_{ij} \cdot N_i] \cdot I_\upsilon(z) + A_{ji} \cdot N_j \cdot (\Omega/4\pi) \cdot h \cdot \upsilon \qquad (7.5)$$

The second term at the end of the right side of Eq. (7.5) is the contribution of *spontaneous emission* to the collected signal. It is basically a source of noise. Neglecting the noise term and using the relationship (5.7a), obtained in Chapter 5, between the Einstein coefficients and the Max Plank's law (see Chapter 3), we obtain the most widely used expression to describe the process of amplification/attenuation arising from the competing processes of *stimulated emission* and *absorption,* respectively:

$$(dI_\upsilon/dz) = \{A_{ji} \cdot g(\upsilon) \cdot (\lambda_0/n)^2 / 8\pi \cdot [N_j - N_i(g_2/g_1)]\} \cdot I_\upsilon(z) = \gamma_0(\upsilon) \cdot I_\upsilon(z) \qquad (7.6)$$

Equation (7.6) is only valid for $I_\upsilon(0)$ sufficiently small to ensure negligible perturbation of N_j and N_i. For this reason, $\gamma_0(\upsilon)$ is referred to as the small

signal gain coefficient, which is frequency-dependent through the lineshape function $g(\upsilon)$. Clearly the condition for amplification is that the term $[N_j - N_i(g_2/g_1)]$, referred to as the population inversion, is greater than 0, i.e., $N_j > N_i(g_2/g_1)$.

Integrating Eq. (7.6) over coordinate of slab z, the intensity as a function of z for an input signal of $I_\upsilon(0)$ can be obtained [6, 7]:

$$I_\upsilon(z) = I_\upsilon(0)\cdot\exp\{\gamma_0(\upsilon)\cdot z\} \tag{7.7}$$

For a gain medium of length l, Eq. (7.7) becomes

$$I_\upsilon(l) = G_0(\upsilon)\cdot I_\upsilon(0) \tag{7.8}$$

where $G_0(\upsilon) = \exp\{\gamma_0(\upsilon)\cdot l\}$ is the *overall gain* and $\gamma_0(\upsilon)$ is the *small signal gain* of the amplifier of length l, respectively. In decibels, the overall gain of the amplifier can be presented as:

$$G(\text{dB}) = 10\log G_0(\upsilon) = 10\log[\exp(\gamma_0(\upsilon)\cdot l)] = 4.34\cdot\gamma_0(\upsilon)\cdot l \tag{7.9}$$

Equation (7.9) shows that the small signal gain (in dB) of an optical amplifier increases linearly with the gain coefficient and the pump power. Indeed, the increase in the pump power leads to an increase in the population of the upper gain state linearly, as well as to an increase of the population inversion and the gain coefficient.

7.3. Pumping Mechanism in Optical Amplifiers

As follows from previous discussions and from Eq. (7.9), at *weak pump powers*, the population inversion is insufficient to provide gain and the signal is attenuated by an amount depending on the population of the lower gain state.

With an increase of the pump power, the population inversion and stimulated emission increase. At the same time, the attenuation decreases, and the system becomes transparent. Beyond the point of transparency (the gain threshold), the gain (in dB) increases linearly with pump power according to Eq. (7.9). It must be noted that the approach presented in the previous section only applies under weak pumping conditions for which we can assume insignificant depletion of the ground state.

For *strong pumping*, the ground state becomes severely depleted, and further increases in pump power result in minimal improvements to the gain and the output power. Let us briefly consider the process of pumping accounting

for the fact that under thermal equilibrium, the condition of most of the atoms of any solid material is in the ground state and the relative populations of the higher allowed states are given by the Boltzmann distribution (see Figure 7.3).

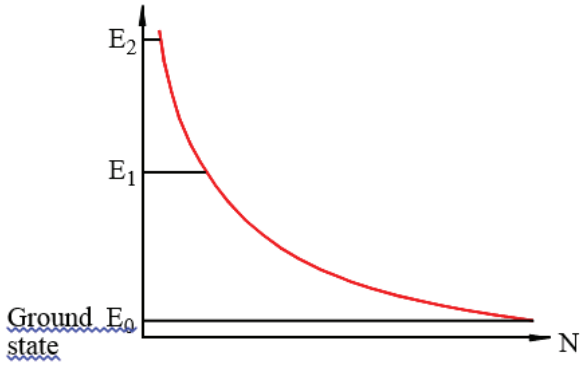


Figure 7.3. Distribution of energy levels of atoms according to Boltzmann's law.

In most semiconducting materials, laser and amplifier gain are pumped optically. In this process, atoms in the ground state (ground state energy level E_o, see Figure 7.3) of the material are raised to a higher energy level (E_j) by the absorption of photons of energy, $E_j - E_0$, supplied from an external light source. If only two energy levels were involved in the process, then the population of the upper state would increase until the rates of absorption and stimulated emission of pump photons were equal. Hence, a population inversion *cannot be created* by pumping from the ground state in an only *two-level system*!

Most optical amplifiers and lasers are based on either a three- or a four-level gain medium and pumping system. Figure 7.4, shows the simplified energy level diagram and pumping scheme for a typical *three-level system*.

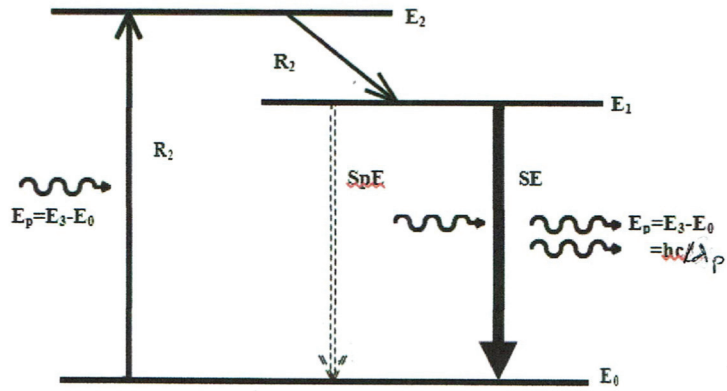


Figure 7.4. Three-level pumping system occurring in optical amplifiers (according to [6, 7]).

Atoms are pumped by photons of energy, $E_2 - E_0$, from the ground state to some higher energy level, E_2, from where they make rapid transitions into energy level E_1. Provided that the rate of decay of the atomic population in the E_1 level is slow relative to the pumping rate (i.e., the E_1 level is metastable), the population of E_1 will increase to exceed that of the ground state, thus creating a population inversion. Light of a wavelength satisfying the relationship $E_{photon} = E_1 - E_0 = h \cdot c/\lambda_p$ can then be amplified by this gain medium.

The rates of pumping and decay of the various populations are also indicated in Figure 7.4, where R_2 is the rate ($dN_2/dt_{[pump]}$) at which atoms are being pumped into state E_2 from the ground state as a result of absorption of the pump light. Since the transition rate, N_2/τ_{21}, to state E_1 is very rapid (τ_{21} is short), R_2 is also the pumping rate ($dN_1/dt_{[pump]}$) of the upper lasing level, E_1.

Figure 7.5 presents a *four-level pumping system*, where ground state atoms are pumped by photons of energy $E_3 - E_0$, to energy level E_3 from where they rapidly make the transition to the metastable state, E_2. Due to the long lifetime / slow decay of the atomic population in the metastable E_2 level, its population builds up, creating an inversion relative to level E_1 and providing amplification of light of λ satisfying the relationship $E_{photon} = E_2 - E_1 = h \cdot c/\lambda_p$.

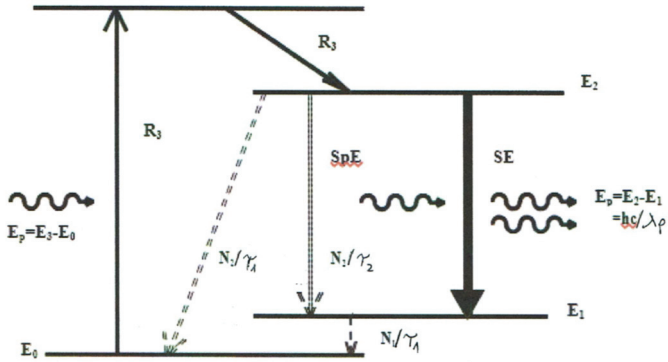


Figure 7.5. Four-level pumping system occurring in optical amplifiers (according to [6, 7]).

In efficient gain media, the E_1 level is sufficiently higher than the ground state. In addition, its transition rate to the ground state is usually very fast to ensure that its population remains negligible under high rates of emission from the E_2 level. Here again, the pump and decay rates are indicated on the energy level diagram. R_3 is the pump rate of level E_3 and the upper lasing level, E_2, since the transition rate, N_3/τ_{32}, is rapid. In the absence of stimulated emission, the population of the metastable E_2 level decays slowly by spontaneous emission to E_1 and to the ground state at the rates N_2/τ_{21} and N_2/τ_{20}, respectively. The decay rate of the E_1 population, N_1/τ_{10}, is very rapid, thus maintaining low N_1.

In the analysis of *small signal gain* presented above following [6, 7], it is obviously assumed that the population inversion is constant, remaining unperturbed by the low levels of stimulated emission arising from amplification of a weak input signal. As the signal strength increases, the stimulated emission process begins to significantly reduce the population inversion and the gain decreases, a phenomenon referred to as gain saturation.

To analyze *large signal gain* and *gain saturation*, we must consider the coupled rate equations for all of the transitions which influence the populations of the two energy levels involved in the amplification process. The analysis and results are different for three- and four-level pumping systems (see Figures 7.4 and 7.5) was carried out in Ref. [6] and we do not enter into the precise analysis illuminated there. We will only emphasize that knowledge of stimulated emission cross-section, σ_{SE}, allows us to find a large signal gain of the amplifier $\gamma(\upsilon)$ via intensity I_υ and times of relaxation

τ_1 and τ_2, respectively. Thus, for four-level pumping the optical amplifier yields:

$$\gamma(\upsilon) = \gamma(\upsilon) \cdot [1 + (\tau_1 + \tau_2 - \tau_1 \cdot \tau_2 / \tau_{21}) (\sigma_{SE} \cdot I_\upsilon / h\upsilon)]^{-1} \quad (7.10)$$

For *homogeneously broadened gain* material (see definition in Chapter 5) under *intense radiation* at any wavelength under the gain curve, the high level of stimulated emission simply depletes the population of the upper state and the entire gain curve diminishes but maintains its shape (see Figure 7.6). This means that the gain for all wavelengths under the gain curve is reduced uniformly.

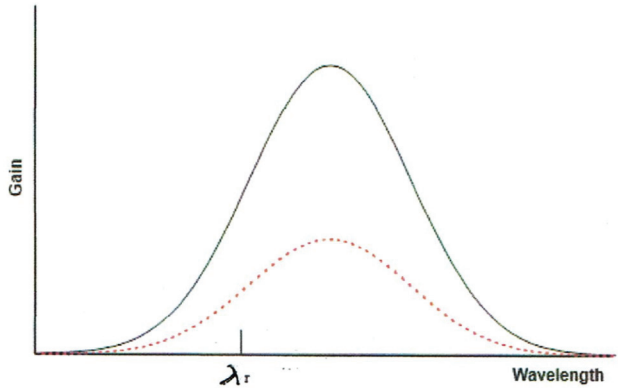


Figure 7.6. A typical gain curve under small signal conditions (solid line) and under internal radiation at λ_r for *homogeneously* broadened transition (dashed lines).

For gain media with *inhomogeneously* broadened transitions (see definition in Chapter 5) under *intense radiation*, the population of the upper state only decreases for that group of atoms with a homogeneous lineshape that overlaps the radiation wavelength. Hence, the gain is only diminished in a narrow range of wavelengths (the homogeneously broadened linewidth for these atoms) around the radiation wavelength (see Figure 7.7).

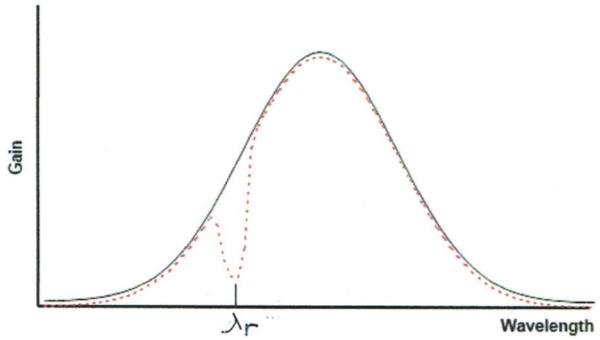


Fig. 7.7. The same, as in Fig. 7.6, but for *inhomogeneously* broadened transition.

This phenomenon is referred to as spectral hole burning. The gain for wavelengths under the gain curve but out with this region is unaffected.

7.4. Noise in Optical Amplifiers

As was mentioned in Chapter 5 and above, the amplified spontaneous emission (ASE), is a random process and when mixed with the signal on the detector, it is a source of noise. Noise associated with the ASE is the limiting factor in determining the ultimate signal-to-noise ratio in any system using optical amplifiers [4], particularly in long haul periodically amplified systems using EDF, in which the ASE accumulates through the system (see discussions on EDFA below).

Let us consider the spontaneous emission from a cylindrical gain medium of cross-sectional area A and length l (see Figure 7.8, rearranged from [6, 7]). A cross-sectional slab of material of infinitesimal thickness, dz, will spontaneously emit a total power $A_{21}N_2 \cdot g(\upsilon) \cdot d\upsilon \cdot h\upsilon \cdot A \cdot dz$ in the frequency range υ to $\upsilon + d\upsilon$.

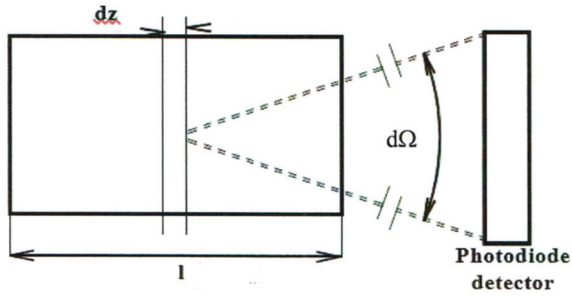


Figure 7.8. Schematically sketched spontaneously emitted light falling at the detector under angle dΩ (rearranged from [6, 7]).

Generally, we are only concerned with power emitted through the end face of the cylinder and confined within a given solid angle, dΩ. The solid angle dΩ may be the angle subtended at the center of the cylinder by a remote receiver, as shown in Figure 7.8. Alternatively, if the gain medium is in the form of a waveguide as in optical fiber amplifiers, dΩ can be associated with the numerical aperture of the guide, i.e., $d\Omega = \pi A^2/4$.

$$dP_{ASE} = A_{21} \cdot N_2 \cdot g(\upsilon) \cdot d\upsilon \cdot h\upsilon \cdot A \cdot dz \cdot d\Omega / 4\pi \tag{7.11}$$

For small dΩ power emitted spontaneously from a slab at position z along the length of the cylinder l will be amplified by the remaining gain medium of length l-z by a factor $e^{j(l - z)}$ before it leaves the exit face. Hence, the ASE power, in the frequency interval dυ, emerging from the cylinder end face is the summation of the contributions from each slab of thickness dz and can be found after integration of (7.11) over z, which yields:

$$P_{ASE} = [G\text{-}1] \cdot A_{21} \cdot N_2 \cdot g(\upsilon) \cdot h\upsilon \cdot A \cdot d\Omega \cdot d\upsilon / (\gamma \cdot 4\pi) \tag{7.12}$$

where, as above, G is the total gain of the amplifier given as $G = \exp(\gamma \cdot l)$.

Accounting for well-known signal gain after spontaneous emission and the cross-section of such an emission, given, respectively, by

$$\gamma(\upsilon) = \sigma_{SE}(\upsilon) \cdot [N_2 - N_1 \cdot (g_2 / g_1)] \tag{7.13}$$

$$\sigma_{SE}(\upsilon) = A_{21} \cdot g(\upsilon) \cdot (\lambda_0 / n)^2 / 8\pi \tag{7.14}$$

and substituting them in (7.12), we finally get:

$$P_{ASE} = 2\mu\cdot[G\text{-}1]\cdot h\upsilon \cdot A\cdot n^2\cdot d\Omega\cdot d\upsilon / \lambda_0^2 \tag{7.15}$$

where $\mu = N_2 / [N_2 - N_1 \cdot (g_2 / g_1)]$ which is known as the *population inversion factor*. The term $A\cdot n^2\cdot d\Omega / \lambda_0^2$ characterizes the geometry of the light emission and collection system relative to the wavelength. To minimize the ASE traveling with the beam to the detector, one can place an aperture stop at the output facet with a radius, a, equal to the beam radius. If so, the output beam diverges by diffraction at a half angle, θ, given by:

$$\theta = \lambda_0 / \pi\cdot n\cdot a \tag{7.16}$$

For small angles θ, this corresponds to a solid angle, $\Omega_{\min}$

$$\Omega_{\min} = \pi\cdot\sin^2\theta = \pi\cdot\theta^2 = \pi\cdot(\lambda_0 / \pi\cdot n\cdot a)^2 = \lambda_0 / n^2\cdot A \tag{7.17}$$

For small angles and assuming transmission through a linear polarizer, the minimum ASE power incident on the receiver with the signal is obtained by substituting $\Omega_{\min}$ for $d\Omega$ in (7.15) and dividing by 2 to give:

$$P_{ASE} = \mu\cdot[G\text{-}1]\cdot h\upsilon \cdot d\upsilon = \rho_{ASE} \cdot d\upsilon \tag{7.18}$$

where $\rho_{ASE} = \mu[G - 1]h\cdot\upsilon$ is the spectral power density of the ASE contained within the amplified signal beam and reaching the detector via a linear polarizer. Here, we showed the resulting expression of the ASE power for a single linear polarization state, because only the *E*-field components in the ASE, which are polarized parallel with the signal *E*-field result in nonzero beat terms. For non-polarized ASE power, the right-hand side of (7.18) is simply doubled.

Hence, the ASE power per frequency interval $d\upsilon$ propagating in the fiber optic channel with the signal, polarized in the same direction as the signal and inseparable from it, is fully described by (7.18).

At the output of an optical amplifier, the total optical power, P_R, incident on a receiver is thus the summation of the received signal power, P_S, and the total ASE power, $\rho_{ASE}\cdot B_o$, which has accumulated from the amplifier:

$$P_R = P_S + \rho_{ASE} \cdot B_o \tag{7.19}$$

where B_o is the optical bandwidth of the system or of an optical filter placed in front of the receiver.

The presence of the ASE gives rise to additional optical noise at the receiver output, over and above the signal shot noise current. The ASE has its own shot noise, and it beats both with itself and the signal in the square law detector to generate ASE-ASE beat noise and signal (S) -ASE beat noise. These optical noise components must be considered in addition to the intrinsic noise of the receiver, which is usually dominated by thermal noise from the load resistance.

Noise is characterized by the variance of the current fluctuations, which is equivalent to the mean square current fluctuations $<i^2>$. The noise sources discussed above are uncorrelated and the total variance of the receiver current fluctuation, σ_N, is simply the sum of the variances associated with each noise source, i.e.:

$$\sigma_N = \sigma_S + \sigma_{ASE} + \sigma_{S\text{-}ASE} + \sigma_{ASE\text{-}ASE} \tag{7.20}$$

where the terms on the right are, in order of appearance, the mean square current fluctuations (the current variance) associated with thermal noise in the receiver, signal shot noise, ASE shot noise, S-ASE beat noise and ASE-ASE beat noise. The optical noise terms are given by the following expressions:

$$\sigma_S = 2e \cdot I_S \cdot B_e = 2e \cdot R \cdot P_S \cdot B_e = 2e \cdot R \cdot G \cdot P_0 \cdot B_e \tag{7.21a}$$

$$\sigma_{ASE} = 2e \cdot I_{ASE} \cdot B_e = 2e \cdot R \cdot P_{ASE} \cdot B_e \tag{7.21b}$$

$$\sigma_{S\text{-}ASE} = 4 \cdot R^2 \cdot G \cdot B_0 \cdot \rho_{ASE} \cdot B_e = 4 \cdot R^2 \cdot G \cdot P_0 \cdot P_{ASE}^{Be} \tag{7.21c}$$

$$\sigma_{ASE\text{-}ASE} = R^2 \cdot \rho_{ASE}^2 \cdot B_0 \cdot B_e = 2 \cdot R^2 \cdot P_{ASE}^{Be} \cdot P_{ASE}^{B_0} \tag{7.21d}$$

Here e is the electronic charge, R is the photodiode responsivity ($R = \eta q / h\upsilon$, η being the quantum efficiency, defined in Chapter 5), B_e is the receiver bandwidth, B_o is the optical bandwidth, I_S and I_{ASE} are the photodetector currents arising from the signal and ASE, respectively; P_0 is the signal input power to the amplifier, P_S is the received signal power and $P_{ASE}^{B_0} = \eta_{ASE} \cdot B_0$ and $P_{ASE}^{Be} = \eta_{ASE} \cdot B_e$ are the single polarization ASE powers in the optical and electrical (receiver) bandwidths respectively.

Clearly from (7.21a) to (7.21d) for any significant level of gain, G, and input signal, P_0, the S-ASE beat noise and/or the ASE-ASE beat noise terms represent the largest optical contributions to the total noise. In most applications of optical amplifiers, one or other or both of these noise components limits the overall performance of any system.

Signal-to-Noise Ratio. As was mentioned in Chapter 6, for most applications the received signal-to-noise ratio, SNR_{out}, at the output of an optical amplifier is:

$$SNR_{out} = (R \cdot G \cdot P_0)^2 / [4 \cdot R^2 \cdot G \cdot P_0 \cdot P_{ASE}^{B_0} + 2 \cdot R^2 \cdot P_{ASE}^{Be} \cdot P_{ASE}^{B_0}] \tag{7.22}$$

For some applications, particularly for low signal levels and when the amplifier is not in saturation, the ASE-ASE beat noise becomes important and dominant, and we can neglect the first term in the denominator. Conversely, in many other practical applications the amplifiers have significant output signal levels and operate at or near saturation, implying that we can neglect the second term. In such cases, the received SNR is given by:

$$SNR_{out} = (R \cdot G \cdot P_0)^2 / [4 \cdot R^2 \cdot G \cdot P_0 \cdot P_{ASE}^{B_0}] = (G \cdot P_0) / (4 \cdot P_{ASE}^{B_0}) \tag{7.23}$$

Generally speaking, our suggestions are realistic, because the shot noise and thermal noise influence shown in a system (7.21) were proven experimentally, their expressions are well known, and their derivations may be found in most textbooks on optical communications. The two beat noise terms in a system (7.21) are more particular to systems using optical amplifiers and are less familiar. Nevertheless, we present SNR via these two terms, accounting for ASE-ASE beat noise, as it is described by (7.22).

Noise Figure. Often it is convenient to characterize the noise and SNR of optically amplified systems using a parameter known as the *amplifier noise figure*, *NF*. Usually, NF is defined as the ratio of the optical SNR at the amplifier input to the optical SNR at the output, as detected by a receiver whose intrinsic noise level (thermal noise) is less than the optical noise in both cases. The optical noise at the input is simply the signal shot noise and the SNR (using (7.21a) and $R = e/h\upsilon$, $\eta = 1$) is given by:

$$SNR_{out} = (R \cdot P_0)^2 / [2e \cdot R \cdot P_0 \cdot B_0] = P_0 / (2h \cdot \upsilon \cdot B_e) \tag{7.24}$$

Using (7.23), the *NF* is given by

$$NF = SNR_{out} / SNR_{in} = 2 \cdot P_{ASE}^{B_0} / (G \cdot h \cdot \upsilon \cdot B_e) \tag{7.25}$$

Substituting in (7.23) $P^{Be} = \eta_{ASE} \cdot B_e$ and using (7.18) by assuming gain $G >> 1$, we get

$$NF = 2\mu \tag{7.26}$$

This implies that the minimum possible noise figure (*NF*) is 2 (3dB) is for an ideal amplifier having a complete population inversion (i.e., $N_1 = 0$, $\mu = 1$). That is, even for an ideal amplifier, the output SNR is degraded by 3dB relative to the input SNR. Typically, in practice, amplifiers operate with a noise factor greater than 3dB, and it can be as high as 7–8 dB. If the NF is known under the conditions at which the amplifier is operated, then we can use it to calculate the output SNR. Applying (7.24) and (7.25), SNR_{out} in terms of the noise figure is:

$$SNR_{out} = P_0 / (2h \cdot \upsilon \cdot B_e \cdot NF) \tag{7.27}$$

From measurements of the total ASE power and the ASE spectrum (see Figure 7.9) we can readily calculate P^{Be} (i.e., the ASE power within the receiver bandwidth).

Noise Figure (*NF*). In many systems the ASE-ASE beat noise is significant and must be included in the measurements of noise figure and expressions for SNR_{out} presented by (7.22) and by the following expression:

$$NF = 2 \cdot PASE^{Be} / (G \cdot h \cdot \upsilon \cdot B_e) + P_{ASE}^{B0} \cdot P_{ASE}^{B0} / (G^2 \cdot h \cdot \upsilon \cdot B_e) \tag{7.28}$$

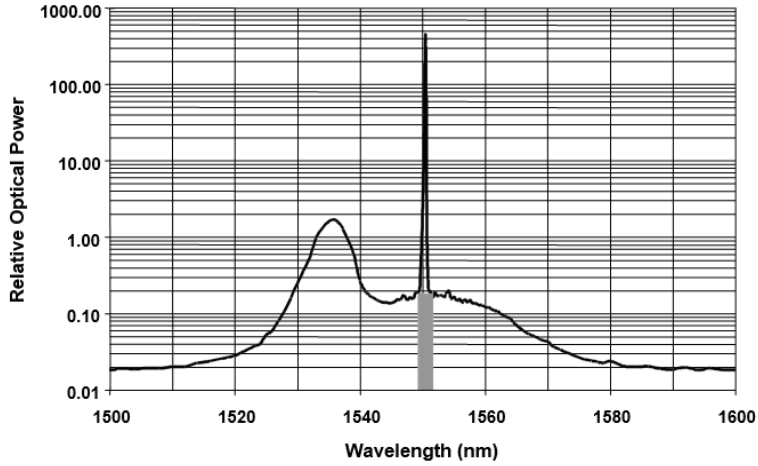


Figure 7.9. ASE power spectrum showing the ASE falling within the bandwidth of the receiver; B_e; P^{Be} is integral of the ASE spectrum over the shaded area.

Hence, measurement of the total ASE power, plus the ASE spectrum and the gain allow the noise figure (*NF*) to be calculated.

7.5. Erbium Doped Fiber Amplifier (EDFA)

7.5.1. Structure and Principle of Operation of EDFA

The main goal of researchers in the late nineteen eighties was to find the preferred wavelength for long optical communications systems, starting with 1550 nm for the construction of semiconductor optical amplifiers at 1550 nm. Spectroscopic studies had shown that *erbium* atoms may be suitable as an active 3 level species for optical amplification at 1550 nm. Figure 7.10 shows a partial energy level diagram for erbium atoms doped into a glass host. Regarding fiber optic communication links we will discus in Chapter 10.

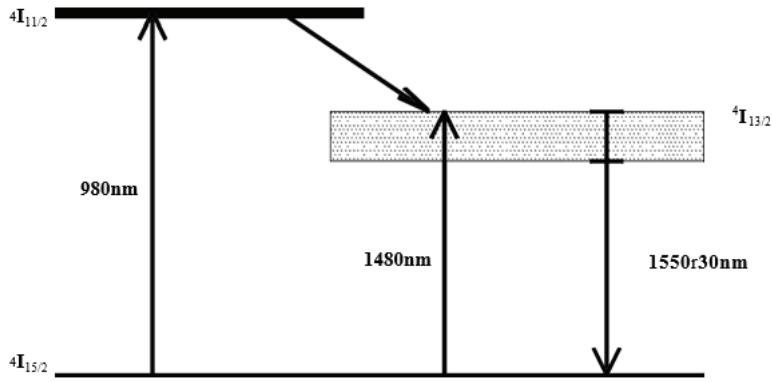


Figure 7.10. Energy level diagram and pumping scheme for erbium doped silica glass – a three-level system giving gain at 1550 ± 30 nm (according to [6, 7]).

Three-level systems (see definition above) with the ground state as the lower gain state require very strong pumping to achieve a population inversion and the erbium doped glass system is further impaired by the inability to achieve high doping concentrations, implying the need for long lengths of material to achieve significant gain. The broad band of levels denoted $^4I_{13/2}$ are metastable with long spontaneous emission lifetimes in the region of a few milliseconds and transitions to the ground state produce photons in the wavelength range 1520–1580 nm, providing the possibility

of optical amplification centered on 1550 nm.

As follows from the discrete zone diagram depicted in Figure 7.10, the population of the $^4I_{13/2}$ levels could be pumped by irradiation at 980 nm or 1480 nm. Photons at 980 nm are absorbed by ground state atoms which make the transition to energy level $^4I_{11/2}$. Further non-radiative transitions from the $^4I_{11/2}$ level to the $^4I_{13/2}$ level are very rapid and the population of the $^4I_{13/2}$ metastable levels builds up.

Alternatively, irradiation by 1480nm light allows direct pumping into the upper levels of the $^4I_{13/2}$ band with rapid transitions to the lower, long-lived levels allowing this pumping scheme to operate as a three-level system.

The structure of a typical EDFA is shown in Figure 7.11, according to [6, 7]. Erbium ions are the active gain species which, when doped into silica glass form a useful 3-level gain medium.

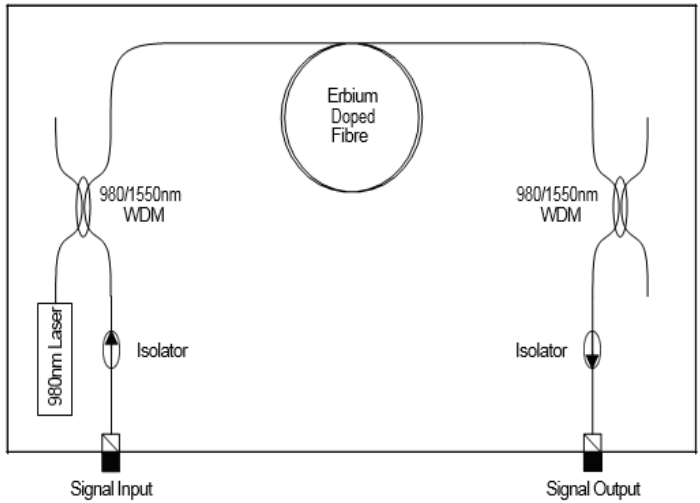


Figure 7.11. Schematic diagram of EDFA (according to [6, 7])..

As was shown in Fig. 7.10, depicted following [6, 7], pumping at 980 nm or 1480 nm results in a population inversion between an intermediate state and the ground state providing gain in a band of wavelengths from 1520 to 1580 nm. In single mode fiber form, the high intensity and strong confinement of the pump and signal light creates an extremely efficient amplifier offering gains in excess of 40 dB for modest pump power, gain efficiencies in the region of 2–10 dB per mW of pump,

saturated output powers in excess of 100 mW and low noise figure (*NF*) down to < 4 dB. With such performance EDFAs are finding widespread use in fiber optic systems.

Finally, EDFAs, based on an all fiber structure, are readily spliced into optical fiber systems to provide in-line gains of up to 40 dB or greater with the added advantages of low power consumption and high reliability. In addition, they offer the benefits of bit rate transparency and wavelength transparency over a 30nm range, leading to high bit rate operation, and the possibility of simultaneous multiple wavelength amplification for wavelength division multiplexing [3–5]. However, EDFAs provide only signal amplification without regeneration of the pulse shape or its width.

With such advantages and suitably few disadvantages, EDFAs are replacing optoelectronic repeaters as the in-line signal conditioning elements in communications systems, and they have found numerous applications as power amplifiers and receiver pre-amplifiers in many other fields of fiber optics.

7.5.2 Gain Characteristics of EDFA

We start to briefly discuss some of the principles and equations directly required in the following study. For small signals of insufficient intensity to significantly perturb the population of the upper gain state, the gain in intensity per unit length (dI_n/dz) at a given distance z along a uniformly pumped amplifier is given by (7.6), where the term in the brackets is the population inversion, $I_\upsilon(z)$ is the intensity at z defined by expression (7.7), and $\gamma_o(\upsilon)$ is referred to as the small signal gain coefficient given by formula the following formula:

$$\gamma_o(\upsilon) = \sigma_{SE}(\upsilon) \cdot [N_j - N_j(g_i / g_j)] \tag{7.29}$$

Here σ_{SE} was defined above as a constant referred to as the stimulated emission cross-section determined by expression (7.14). As for intensity, defined by (7.7), it can be described via the overall gain of the EDFA amplifier of the length l via an input intensity $I_\upsilon(0)$ as [6, 7]:

$$I_\upsilon(l) = G_0(\upsilon) \cdot I_\upsilon(0) \tag{7.30}$$

where

$$G_0(\upsilon) = \exp\{\gamma_0(\upsilon) \cdot l\}$$

or in decibels, the overall gain of the EDFA will be expressed by the same formula, as formula (7.9), which we present again for the readers' convenience:

$$G(\text{dB}) = 10\log G_0(\upsilon) = 10\log[\exp(\gamma_0(\upsilon) \cdot l)] = 4.34 \cdot \gamma_0(\upsilon) \cdot l \qquad (7.31)$$

7.5.3 Noise Characteristics of EDFA

Often it is convenient to characterize the noise and SNR of EDFA as usually presented above for optically amplified systems using a parameter known as the amplifier noise figure, *NF* as the ratio of the optical SNR at the EDFA input to the optical SNR at the output, detected by a receiver whose intrinsic noise level (thermal noise) is less than the optical noise in both cases. The optical noise at the input is simply the signal shot noise and the SNR is given by expression (7.24), which we repeat for the readers' convenience:

$$SNR_{\text{out}} = (R \cdot P_0)^2 / [2e \cdot R \cdot P_0 \cdot B] = P_0 / (2h \cdot \upsilon \cdot B_e) \qquad (7.32)$$

where we have used the substitution $R = e/h\upsilon$ (we have assumed a quantum efficiency η=1, see details in Chapter 5). Repeating the ratio (7.25), we present the noise figure by [6. 7]:

$$NF = SNR_{\text{out}} / SNR_{\text{in}} = 2 \cdot P_{ASE}^{B_e} / (G \cdot h \cdot \upsilon \cdot B_e) \qquad (7.33)$$

Substituting (7.21b) for P_{ASE} and assuming significant gain, G>>1, gives the same expression for EDFA, as generally was obtained above, that is,

$$NF = 2\mu \qquad (7.34)$$

This gives us that the minimum possible NF is 2 (3dB) for an ideal amplifier having a complete population inversion (i.e., $N_1 = 0$, $\mu = 1$). That is, even for an ideal amplifier, the output SNR is degraded by 3dB relative to the input SNR. Typically, in practice, EDF-amplifiers operate with a noise factor greater than 3dB, and it can be as high as 7–8 dB.

Following initial data (see Figures 7.10 and 7.11), it was quickly found that EDFAs were highly compatible with 1550nm fiber optic telecommunications systems and that their potential performance could significantly enhance such kinds of fiber systems, namely using fiber optic systems by providing high gain, excellent reliability, and low power consumption.

In addition, they proved to be bit rate transparent and wavelength transparent (within their optical bandwidth), implying that system capacities could be increased by increasing the bit rate capabilities of terminal equipment or by introducing wavelength division multiplexing (WDM) (see [3–5] and bibliography therein). EDFAs have thus found numerous applications, turning previously attenuation limited systems into much higher performance dispersion limited systems. The operational characteristics of EDFA are shown in Table 7.1, according to [6, 7]:

Table 7.1. Main Characteristics of EDFA.

Input power	2mW	EDFA Gain	26dB
Signal bandwidth	2.5GHz	Required SNR	288
Fibre attenuation	0.26dB/km	Distance	640km
Optical Bandwidth	2.0nm	Receiver responsivity	0.65mA/mW
Receiver sensitivity	-25dBm		

Of course, EDFAs provide *only gain* but *do not reshape the signals,* which spread and degrade by dispersion. However, assembled with the *dispersion shifted* and *dispersion compensating fiber*, the dispersion limits have been extended dramatically even to *transoceanic distances*. For these reasons EDFAs were developed from proof of principle to fully engineered products for deployment under the ocean.

Exercises

On the basis of a specific Erbium Doped Fiber Amplifier (EDFA), we will show, as examples, how to compute each of its characteristics, described by system of equations (7.21)-(7.22).

Exercise 1.

Given: The optical bandwidth of the optical amplifier is B_o, which is defined by an optical band pass filter centered on the signal of frequency ν_o; the optical gain G and the amplifier with spontaneous emission (ASE, see definition above) are uniform over that bandwidth. ASE is described by a summation of sinusoidal electric field components of infinitesimally small bandwidth $\delta\upsilon$ and ranging in frequency from $\Omega_o - B_o/2$ to $\Omega_o + B_o/2$.

Find: The beat noise current main terms.

Solution

The ASE spectral power density $\rho_{ASE}(\upsilon)$ transmitted through a linear polarizer is given by [7]:

$$\rho_{ASE}(\upsilon) = h \cdot \upsilon \cdot \mu \cdot (G - 1) \tag{1E}$$

The total ASE power in a single linear polarization state is

$$P_{ASE}(B_o) = \rho_{ASE}(\upsilon) \cdot B_0 \tag{2E}$$

Given that the optical power is proportional to the time averaged electric field amplitude squared, i.e.

$$<P> = c \cdot \varepsilon_0 \cdot <E^2> = c \cdot \varepsilon_0 \cdot E_0^2/2 \tag{3E}$$

In Eq. (3E), c is the speed of light and ε_0 is the vacuum permittivity. Finally, the narrow band sinusoidal field components associated with the ASE may be written:

$$E_{ASE}(t) = \sum_{k=-B_o/2\delta\nu}^{+B_o/2\delta\nu} \sqrt{\frac{2\rho_{ASE}\delta\nu}{c\varepsilon_0}} \cdot \cos[(\omega_o + 2\pi k\delta\nu)t + \phi_k] \tag{4E-1}$$

Denoting now $M = B_0 / 2\delta\upsilon$, yields

$$E_{tot}(t) = \sqrt{\frac{2GP_o}{c\varepsilon}} \cdot \cos(\omega_o t) + \sum_{k=-M}^{+M} \sqrt{\frac{2\rho_{ASE}\delta\nu}{c\varepsilon_o}} \cdot \cos[(\omega_o + 2\pi k\delta\nu)t + \phi_k] \tag{4E-2}$$

The photo current, $i(t)$, generated in the detector of responsivity R by incident power, P_{in}, can be found as [7]

$$i(t) = RP_{in} = Rc\varepsilon_o \left\langle E_{Tot}^{\ 2}(t) \right\rangle \tag{5E}$$

where the brackets denote time averaging over the optical frequencies.

Substituting (4E2) into (5E) yields:

$$i(t) = RGP_o + 4R\sum_{k=-M}^{+M}\cdot\sqrt{GP_o\rho_{ASE}\delta\nu}\cdot\cos(\omega_o t).\cos[(\omega_o + 2\pi k\delta\nu)t + \phi_k] \qquad (6E)$$
$$+ 2R\rho_{ASE}\delta\nu\left[\sum_{k=-M}^{+M}\cos[(\omega_o + 2\pi k\delta\nu)t + \phi_k]\right]^2$$

So, a general expression was found consisting of three terms on the right-hand side of (6E): the signal noise, the signal (S)-ASE beat noise, and the ASE-ASE beat noise.

Exercise 2.

Given: The same, as above, optical ASE with bandwidth B_o, and with the optical gain G. The electric field of ASE is ranging in frequency from Ω_o - $B_o/2$ to $\Omega_o + B_o/2$ over infinitesimally small bandwidth δv.

Find: The signal-ASE beat noise current.

Solution

The second term in Eq. (6E) gives the signal-ASE beat noise current. By use of the trigonometrical identity 2cosA·cosB, we can rewrite the second term obtaining a series of sum and difference frequency terms proportional to ($2\acute{\omega}0 + 2k\cdot\pi\cdot\delta v$). The sum frequencies can be filtered out by low passband of the optical receiver allowing us to present the second term in Eq. (6E) in the following form:

$$i_{S-ASE}(t) = 2R\sum_{k=-M}^{+M}\cdot\sqrt{GP_o\rho_{ASE}\delta\nu}\cdot\cos(2\pi k\delta\nu t + \phi_k)] \qquad (7E)$$

For each frequency, $2k\cdot\pi\cdot\delta v$, there are two contributions of random relative phase, one from each ASE component symmetrical in frequency space about the signal frequency, $\acute{\omega}_0$. Now, as was given from beginning, the ASE spectrum is uniformly distributed over the bandwidth, B_o, then the power spectrum of the beat noise is also will be uniform over the frequency interval Ω - $B_o/2$.

To obtain the mean square noise current, associated with each frequency component, square each of the components of (7E), take the time average and multiply by 2 (to account for the 2 components symmetrical on

either side of the signal frequency), which yields

$$< i^2_{S\text{-}ASE}(\nu)> = 4R^2 \cdot G \cdot P_0 \cdot \rho_{ASE}\, \delta\nu \tag{8E}$$

Finally, the mean square noise current density is:

$$\rho_{S\text{-}ASE} = < i^2_{S\text{-}ASE}(\nu)> = 4R^2 \cdot G \cdot P_0 \cdot \rho_{\text{ASE}} \tag{9E}$$

Hence, the total mean square noise current within the electronic bandwidth, B_e, of the detector can be written

$$\begin{aligned}\sigma_{S\text{-}ASE} = 4 \cdot R^2 \cdot G \cdot P_0 \cdot \rho_{\text{ASE}} \cdot B_o = 4 \cdot R^2 \cdot G \cdot P_0 \cdot P_{ASE}^{Be} = \\ = 4 \cdot e^2 \cdot \eta^2\, G \cdot P_0 \cdot \mu \cdot (G\text{-}1) \cdot B_e\end{aligned} \tag{10E}$$

where P_{ASE}^{Be} is the single polarization ASE power within the receiver bandwidth and η is the quantum efficiency of the optical detector in the receiver.

Exercise 3.

Given: The same, as above, optical ASE with bandwidth B_o, and with the optical gain G. The electric field of ASE ranges in frequency from $\Omega_o - B_o/2$ to $\Omega_o + B_o/2$ over infinitesimally small bandwidth $\delta\upsilon$.

Find: The ASE-ASE beat noise current.

Solution

The third term in Eq. (6E) represents the ASE-ASE beat noise current and can be written, according to Ref. [7], as

$$\begin{aligned} i_{ASE-ASE}(t) &= 2R\rho_{ASE}\delta\nu\left[\sum_{k=-M}^{+M}\cos[(\omega_o + 2\pi k\delta\nu)t + \phi_k]\right]^2 \\ &= 2R\rho_{ASE}\delta\nu\left[\sum_{k=-M}^{+M}\cos\beta_k \cdot \sum_{j=-M}^{+M}\cos\beta_j\right] \end{aligned} \tag{11E}$$

For simplification of expression (11E), here the total phases for harmonics of electric field with numbers *j* and *k* is given as:

$$\begin{aligned}\beta_k &= (\omega_o + 2\pi k\delta\nu)t + \varphi_k \\ \beta_j &= (\omega_o + 2\pi j\delta\nu)t + \varphi_j\end{aligned} \tag{12E}$$

Using the standard trigonometric identity for 2cosA·cosB, we rewrite Eq. (11E) as:

$$i_{ASE-ASE}(t) = 2R\rho_{ASE}\delta\nu \sum_{k=-M}^{+M}\sum_{j=-M}^{+M}\left[\frac{1}{2}\cos(\beta_k - \beta_j) + \frac{1}{2}\cos(\beta_k + \beta_j)\right] \tag{13E}$$

Squaring and time averaging, following straightforward computations as carried out in [7], it can be easily found that each component at frequency $\delta\upsilon$ gives its contribution to the mean square noise current and for (2M - 1) components at frequency $\delta\upsilon$, the mean square noise current density close to DC current

$$i_{ASE}^{DC} = R\rho_{ASE}\delta\nu 2M = \mu(G-1)eB_o \tag{14E}$$

and finally can be presented as

$$\rho_{ASE-ASE} = \left\langle i_{ASE-ASE}^2(\nu)\right\rangle = 4R^2\rho_{ASE}^2\delta\nu.(2M-1).\frac{1}{2} \tag{15E}$$

For $B_o >> \delta\upsilon$, accounting that $M = B_0 / 2\delta\upsilon$, and assuming a detector quantum efficiency $\eta = 1$, Eq. (15E) finally becomes:

$$\begin{aligned}\sigma_{ASE\text{-}ASE} &= R^2 \cdot \rho^2_{\mathrm{ASE}} \cdot B_0 \cdot B_e = 2\cdot R^2 \cdot P_{ASE}^{Be} \cdot P_{ASE}^{B_0} \\ &= 2\cdot e^2 \cdot \mu^2 \cdot (G\text{-}1)^2 \cdot B_e \cdot B_0\end{aligned} \tag{16E}$$

Here, as above, P_{ASE}^{Be} and $P_{ASE}^{B_0}$ are the single polarization ASE powers within the optical and receiver bandwidth, respectively.
So, we finally found all terms of beat noise current occurring in the EDFA presented in the system (7.21a-d) and in expression (7.22).

Bibliography

[1] Dakin, J. and B. Culshaw, eds. 1988. Handbook: *Optical Fiber Sensors: Principles and Components*. Boston-London, Artech House.
[2] Digonet, M. J. F., and B. Y. Kim. 1988. "Fiber Optic Components." In *Optical Fiber Sensors: Principles and Components*, Vol. 1, edited by J. Dankin and B. Culshaw, 209–248. Boston-London: Artech House.
[3] Kersten, R. Th. 1988. "*Integrated Optics for Sensors*." In *Optical Fiber Sensors: Principles and Components*, Vol. 1, edited J. Dankin and B. Culshaw, 278–317. Boston-London: Artech House.
[4] Yurek, A. M., and Dandridge. 1988. "*Optical Sources*." In *Optical Fiber Sensors: Principles and Components*, Vol. 1, edited by J. Dankin and B. Culshaw, 278–317. Boston-London: Artech House.
[5] Palais, J. C. 1998. *Fiber Optic Communications*, 4th Ed. New Jersey: Prentice-Hall.
[6] Johnston, W. 1997. *Erbium Doped Fiber Amplifiers,* Student Manual, Optoelectronic Systems: Glasgow.
[7] Johnston, W. 1995–2011. *Fiber Optic Communications,* Student Manual, Optoelectronic Systems: Glasgow.
[8] Palais, J. C. 2006. "Optical Communications." In Handbook: *Engineering Electromagnetics Applications*, edited by R. Bansal. New York: Taylor and Frances.

CHAPTER 8

TYPES OF SIGNALS IN OPTICS

In optical wire (e.g. fiber optic) or wireless (e. g. atmospheric) links, the same kinds of signal are formed and transmitted, as in similar radio wire and wireless communication channels. They are continuous and discrete (e.g., pulses). Therefore, the same mathematical tool can be used for the description of such kinds of signals, radio and optical. Let us briefly present a mathematical description of both types of signals – continuous wave (CW) and pulses. In communications, wired and wireless, there are other definitions of these kinds of signals that researchers have used regarding their presentation in the frequency domain. Thus, if we deal with a continuous signal in the time domain, let us say, $x(t) = A(t)\cos\omega t$ that occupies a wide time range along the time axis, its Fourier transform $F[x(t)]$ converts this signal into a narrowband signal, that is, $F[x(t)]=y(f)$, which occupies a very narrow frequency band in the frequency domain. Conversely, if we deal initially with a pulse signal in the time domain that occupies a very narrow time range along the time axis, its Fourier transform $F[x(t)]$ converts this signal into a wideband signal, that is, $F[x(t)]=X(f)$, which occupies a wide frequency band in the frequency domain. Therefore, in the terminology usually used in communication systems and LIDAR design, the continuous signals and the pulses are often called the *narrowband* and *wideband*, respectively. In our description below we will follow both terminologies where the usage of different definitions is more suitable.

8.1. Narrowband or Continuous Wave Optical Signals

A voice modulated *continuous wave* (CW) signal occupies a very narrow bandwidth surrounding the carrier frequency f_c of the signal (e.g., the carrier), which can be expressed as:

$$x(t) = A(t)\cos\left[2\pi f_c t + \phi(t)\right] \tag{8.1}$$

where $A(t)$ is the signal envelope (i.e., slowly-varied amplitude) and $\phi(t)$ is its signal phase. Since all information in the signal is contained within the phase and envelope-time variations, an alternative form of a bandpass signal $x(t)$ is introduced [1, 4–10]:

$$y(t) = A(t)\exp\{j\phi(t)\} \tag{8.2}$$

which is also called the *complex baseband* representation of $x(t)$. By comparing (8.1) and (8.2), we can see that the relation between the *bandpass* and the *complex baseband* signals are related by:

$$x(t) = \mathrm{Re}\left[y(t)\exp\left(j2\pi f_c t\right)\right] \tag{8.3}$$

Relations between these two representations of the narrowband signal in the frequency domain are shown schematically in Figure 8.1.

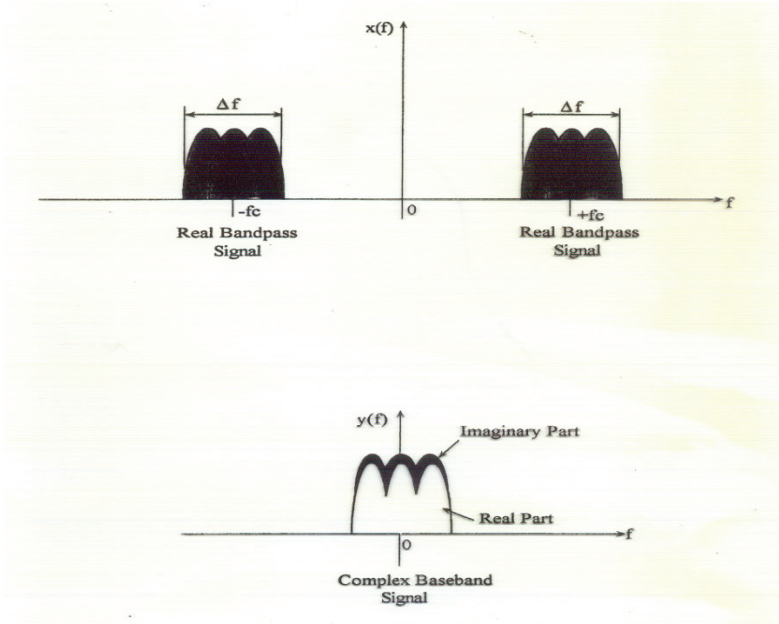


Figure 8.1. Comparison between baseband and bandpass signals.

One can see that the complex baseband signal is a frequency shifted version of the bandpass signal with the same spectral shape but centered on

a zero-frequency instead of the f_c [6–10]. Here, $X(f)$ and $Y(f)$ are the Fourier transform of $x(t)$ and $y(t)$, respectively and can be presented in the following manner [1–3, 12–14]:

$$Y(f) = \int_{-\infty}^{\infty} y(t)e^{+j2\pi ft}\,\mathrm{d}t = \mathrm{Re}\big[Y(f)\big] + j\,\mathrm{Im}\big[Y(f)\big] \tag{8.4}$$

and

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}\,\mathrm{d}t = \mathrm{Re}\big[X(f)\big] + j\,\mathrm{Im}\big[X(f)\big] \tag{8.5}$$

Substituting for $x(t)$ in integral (8.5) from (8.3) gives

$$X(f) = \int_{-\infty}^{\infty} \mathrm{Re}\Big[y(t)e^{j2\pi f_c t}\Big]e^{-j2\pi ft}\,\mathrm{d}t \tag{8.6}$$

Taking into account that the real part of any arbitrary complex variable w can be presented as

$$\mathrm{Re}[w] = \frac{1}{2}\Big[w + w^*\Big]$$

where w^* is the complex conjugate, we can rewrite (8.5) in the following form:

$$X(f) = \frac{1}{2}\int_{-\infty}^{\infty}\Big[y(t)e^{j2\pi f_c t} + y^*(t)e^{-j2\pi f_c t}\Big]\cdot e^{-j2\pi ft}\,\mathrm{d}t \tag{8.7}$$

After comparing expressions (8.4) and (8.7), we get

$$X(f) = \frac{1}{2}\Big[Y(f - f_c) + Y^*(-f - f_c)\Big] \tag{8.8}$$

In other words, the spectrum of the real bandpass signal $x(t)$ can be represented by the real part of that for the complex baseband signal $y(t)$ with a shift of $\pm f_c$ along the frequency axis. It is clear that the baseband signal has its frequency content centered on the "zero" frequency value.

Now we notice that the mean power of the baseband signal $y(t)$ gives the same result as the mean-square value of the real bandpass signal $x(t)$, that is,

$$\left\langle P_y(t)\right\rangle = \frac{\left\langle \left|y(t)\right|^2 \right\rangle}{2} = \frac{\left\langle y(t)y^*(t)\right\rangle}{2} \equiv \left\langle P_x(t)\right\rangle \tag{8.9}$$

The complex envelope $y(t)$ of the received narrowband signal can be expressed according to (8.2) and (8.3), within the multipath wireless channel, as a sum of phases of N baseband individual multiray components arriving at the detector with their corresponding time delay, τ_i, i=0,1,2,…, N-1 [6–10].

$$y(t) = \sum_{i=0}^{N-1} u_i(t) = \sum_{i=0}^{N-1} A_i(t)\exp\left[j\phi_i(t,\tau_i)\right] \tag{8.10}$$

If we assume that during the subscriber movements through the local area of service, the amplitude A_i time variations are small enough, whereas phases ϕ_i vary greatly due to changes in propagation distance between the source and the desired detector, then there are great random oscillations of the total signal $y(t)$ at the detector during its movement over a small distance. Since $y(t)$ is the phase sum in (8.10) of the individual multipath components, the instantaneous phases of the multipath components result in large fluctuations, that is, fast fading, in the CW signal. The average received power for such a signal over a local area of service can be presented according to [1, 4–10] as:

$$\left\langle P_{\mathrm{CW}}\right\rangle \approx \sum_{i=0}^{N-1}\left\langle A_i^2\right\rangle + 2\sum_{i=0}^{N-1}\sum_{i,j\neq i}\left\langle A_iA_j\right\rangle\left\langle \cos\left[\phi_i-\phi_j\right]\right\rangle \tag{8.11}$$

8.2. Wideband or Impulse Optical Signals

The typical *wideband* or *impulse* signal passing through the multipath communication channel is shown schematically in Figure 8.2a following [4–10].

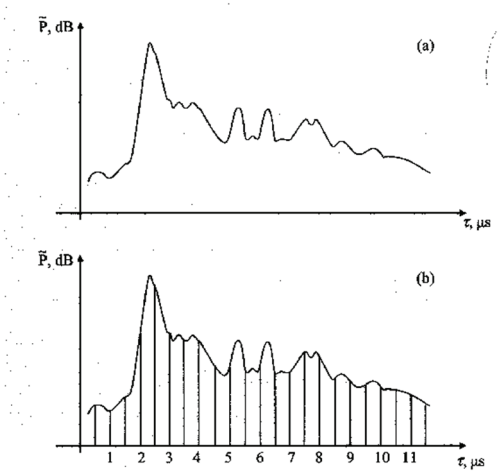


Figure 8.2: (a) A typical impulse signal passing through a multipath communication channel; (b) The use of bins, as vectors, for the impulse signal with spreading.

If we divide the time-delay axis into equal segments, usually called bins, then there will be a number of received signals in the form of vectors or delta functions. Each bin corresponds to a different path whose time of arrival is within the bin duration, as depicted in Figure 8.2b. In this case, the time-varying discrete-time impulse response can be expressed as:

$$h(t,\tau)=\left\{\sum_{i=0}^{N-1}A_i(t,\tau)\exp\left[-j2\pi f_c\tau_i(t)\right]\delta\left(\tau-\tau_i(t)\right)\right\}\exp\left[-j\phi(t,\tau)\right] \qquad (8.12)$$

If the channel impulse response is assumed to be time-invariant, or is at least stationary over a short-time interval or over a small-scale displacement of the detector or source, then the impulse response (8.12) reduces to

$$h(t,\tau)=\sum_{i=0}^{N-1}A_i(\tau)\exp\left[-j\theta_i\right]\delta\left(\tau-\tau_i\right) \qquad (8.13)$$

Where $\theta_i=2\pi f_c\tau_i+\phi(\tau)$. So, the received power delay profile for a wideband or pulsed signal averaged over a small area can be presented simply as a sum of the powers of the individual multipath components,

where each component has a random amplitude and phase at any time, that is,

$$\left\langle P_{pulse} \right\rangle = \left\langle \sum_{i=0}^{N-1} \left\{ A_i(\tau) \left| \exp\left[-j\theta_i \right] \right| \right\}^2 \right\rangle \approx \sum_{i=0}^{N-1} \left\langle A_i^2 \right\rangle \tag{8.14}$$

The received power of the wideband or pulse signal does not fluctuate significantly when the subscriber moves within a local area because, in practice, the amplitudes of the individual multipath components do not change widely in a local area of service,

Comparison between small-scale presentations of the average power of the narrowband (CW) and wideband (pulse) signals, that is (8.11) and (8.14), shows that:

In the cases when $\left\langle A_i A_j \right\rangle = 0$ *or/and* $\left\langle \cos\left[\phi_i - \phi_j \right] \right\rangle = 0$ *, the average power for CW signal and that for pulse are equivalent.*

This can occur when either the path amplitudes are uncorrelated, that is, each multipath component is independent after multiple reflections, diffractions, and scattering from obstructions surrounding both the detector and the source. It can also occur when multipath phases are independently and uniformly distributed over the range of $\left[0,\ 2\pi \right]$. This property is correct for optical wavebands when the multipath components traverse differential paths having hundreds and thousands of wavelengths [6–10].

Bibliography

[1] Marcuse, O. 1972. *Light Transmission Optics*. New York: Van Nostrand-Reinhold Publisher.
[2] Clarke, R. H. 1968. “A statistical theory of mobile-radio reception.” *Bell Systems Technical Journal* 7:957–1000.
[3] Aulin, T. 1979. “A modified model for the fading signal at a mobile radio channel.” *IEEE Trans. Veh. Technol*. 28(3): 182–203.
[4] Jakes, W. C. 1974. *Microwave Mobile Communications*. New Jersey: IEEE Press.
[5] Lee, W. Y. C. 1985. *Mobile Communication Engineering*. New York: McGraw Hill Publications.
[6] Saunders, S. R. 1999. *Antennas and Propagation for Wireless Communication Systems*. New York: John Wiley & Sons.

[7] Feuerstein, M. L., and T. S. Rappaport. 1992. *Wireless Personal Communication*. Boston-London: Artech House.
[8] Steele, R. 1992. *Mobile Radio Communication*. New Jersey: IEEE Press.
[9] Rappaport, T. S. 1996. *Wireless Communications*. New York: Prentice Hall PTR.
[10] Proakis, J. G. 1995. *Digital Communications*. NY: McGraw Hill.
[11] Leon-Garcia, A. 1994. *Probability and Random Processes for Electrical Engineering*. New York: Addison-Wesley Publishing Company.
[12] Stark, H., and J. W. Woods. 1994. *Probability, Random Processes, and Estimation Theory for Engineers*. New Jersey: Prentice Hall.
[13] Blaunstein, N. 2004. "Wireless Communication Systems." In *Handbook of Engineering Electromagnetics*, edited by Rajeev Bansal. New York: Marcel Dekker.
[14] Krouk, E., and S. Semionov, eds. 2011. *Modulation and Coding Techniques in Wireless Communications*, Chichester: Wiley & Sons.

CHAPTER 9

MODULATION OF SIGNALS IN OPTICS

As was shown in Chapter 8, there are two main types of optical signals propagating in fiber optic or atmospheric communication links, time-continuous or analog, which correspond to narrowband channels, and time-discrete or pulse-shaped, which correspond to wideband channels [1–5]. Therefore, there are different types of modulation that are usually used for such types of signals. First of all, we will define the process of modulation and demodulation.

Modulation is the process where the message information is added to the optical carrier. In other words, modulation is the process of encoding information from a message source in a manner suitable for transmission. This process involves translating a baseband message signal, the source, to a bandpass signal at frequencies that are very high with respect to the baseband frequency. The bandpass signal is called the *modulated* signal and the baseband message signal is called the *modulating* signal [3–10].

Modulation can be achieved by varying the amplitude, phase, or frequency of a high frequency carrier in accordance with the amplitude of the baseband message signal. These kinds of analog modulation have been employed in the first generation of wireless systems and have continued until nowadays for LIDAR and optical imaging applications. Further, digital modulation has been proposed for use in current radio and optical communication systems. Because this kind of modulation has numerous benefits compared with conventional analog modulation, the primary emphasis of this topic is on digital modulation techniques and schemes (see the next section). However, since analog modulation techniques are still in widespread use today and will continue to be used in future, they are treated first.

Demodulation is the process of extracting the baseband message from the carrier so that it may be processed and interpreted by the intended radio or optical receiver [1–3]. Since the main goal of a modulation technique is to transport the message signal through an optical communication channel, wire or wireless, with the best possible quality while occupying the least amount of frequency band spectrum, many modern practical modulation

techniques have been proposed to increase the quality and efficiency of various optical communication links, including fiber-optical links.

Below, will be described briefly the main principles of both kinds of modulation, analog and digital, and some examples will be given of the most useful types of modulation adapted for both kinds of channels, narrowband and wideband.

9.1. Analog Modulation of Optical Signals

Each analog signal consists of three main time-varying characteristics: the amplitude $a(t)$, the phase $\phi(t)$, and the angular frequency $\omega(t) = 2\pi f(t)$, since there is a simple relation between the phase and the frequency $\phi(t) = \omega(t) \cdot t + \phi_0$, where ϕ_0 is the initial phase of the signal. In other words, any signal can be presented via these three parameters as

$$x(t) = a(t)e^{j\phi(t)} = a(t)e^{j[\omega(t)\cdot t+\phi_0]} \tag{9.1}$$

Consequently, there are three types of modulation, depending on what characteristic is time-varied in the modulating signal (called the *message*, see above definitions) – amplitude (AM), phase (PM) and frequency (FM) modulation.

9.1.1 Analog Amplitude Modulation

In the amplitude modulation (AM) technique, the amplitude of a high frequency carrier signal is varied in accordance with the instantaneous amplitude of the modulating message signal. The AM signal can be represented through the carrier signal and the modulating message signal as

$$s_{AM}(t) = A_c[1 + m(t)]\cos(2\pi f_c t) \tag{9.2}$$

where $x_c(t) = A_c \cos(2\pi f_c t)$ is a carrier signal with amplitude A_c and high frequency f_c, $m(t) = (A_m / A_c)\cos(2\pi f_m t)$ is a sinusoidal modulating signal with amplitude A_m and low frequency f_m. Usually, the modulation index $k_m = (A_m / A_c)$ is introduced, which is often expressed as a percentage and is called *percentage modulation*. Figure 9.1 shows a sinusoidal modulating signal $m(t)$ and the corresponding AM signal $s_{AM}(t)$ for the case $k_m = (A_m / A_c) = 0.5$ – that is, the signal is said to be 50% modulated.

If $k_m(\%) > 100\%$, the message signal will be distorted at the envelope detector. Equation (9.2) can be rewritten as

$$s_{AM}(t) = \mathrm{Re}[g(t)\exp(j2\pi f_c t)] \tag{9.3}$$

where $g(t)$ is the complex envelope of the AM signal given according to (9.2) by

$$g(t) = A_c[1 + m(t)] \tag{9.4}$$

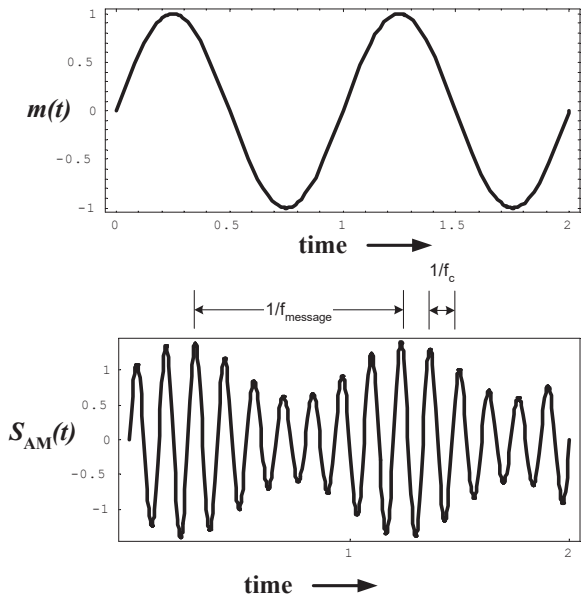


Figure 9.1. The amplitude modulating (top panel) and modulated (bottom panel) signals for the modulation index $k_m = 0.5$.

The corresponding power spectrum of an AM signal can be shown to be [1–4]

$$S_{AM}(f) = \frac{1}{2}A_c\left[\delta(f - f_c) + S_M(f - f_c) + \delta(f + f_c) + S_M(f + f_c)\right] \tag{9.5}$$

where $\delta(\bullet)$ is the unit impulse function, and $S_M(f)$ is the message signal spectrum.

The bandwidth of an AM signal is equal to $B_{AM} = 2f_m$ where f_m is the maximum frequency contained in the modulating message signal. The total power in an AM signal can be obtained as [1–4]

$$P_{AM}(t) = \frac{1}{2} A_c \left[1 + 2 < m(t) > + < m^2(t) > \right] \tag{9.6}$$

where $< m(t) >$ represents the average value of the message signal. Using the expression of the message signal through the modulation index presented above, one can simplify expression (9.6) as

$$P_{AM}(t) = \frac{1}{2} A_c \left[1 + P_m \right] = P_c \left[1 + \frac{k_m^2}{2} \right] \tag{9.7}$$

where $P_c = \frac{1}{2} A_c^2$ is the power of the carrier signal and $P_m = < m^2(t) >$ is the power of the modulating message signal. It can be shown that

$$\frac{1}{2}\left[P_{AM} - P_c \right] = \frac{1}{2}\left[P_c + P_c \frac{k_m^2}{2} - P_c \right] = \frac{1}{2}\left[P_c \frac{k_m^2}{2} \right] = \frac{1}{2}\left[\frac{A_c^2}{2} \frac{A_m^2}{2A_c^2} \right] = \frac{1}{8} A_m^2 = \frac{1}{4} P_m , \tag{9.8}$$

from which follows that $\left[P_{AM} - P_c \right] = P_m / 2$.

9.1.2 Analog Frequency and Phase Modulation

Frequency modulation (FM) is a part of an *angle modulation* technique where the instantaneous frequency of the carrier, $f_c(t)$, varies linearly with the baseband modulating waveform, $m(t)$, i.e.,

$$f_c(t) = f_c + k_f m(t) \tag{9.9}$$

where k_f is the frequency sensitivity (the frequency deviation constant) of the modulator measured in Hz/volt. To understand what it means, let us first of all explain the *angle modulation* technique.

Angle modulation varies a sinusoidal carrier signal in such a way that the phase θ of the carrier is varied according to the amplitude of the modulating baseband signal. In this technique of modulation, the amplitude of the carrier wave is kept constant (called the *constant envelope* modulation). There are several techniques to vary the phase $\theta(t)$ of a carrier signal in accordance with the baseband signal. The well-used techniques of angle modulation are *frequency* modulation and *phase* modulation. In an *FM signal* the instantaneous carrier phase is

$$\theta(t) = 2\pi \int_0^t f_c(t')dt' = 2\pi \left[f_c t + k_f \int_0^t m(t')dt' \right] \tag{9.10}$$

So, the bandpass FM signal can be presented in the following form:

$$s_{FM}(t) = \mathrm{Re}[g(t)\exp(j2\pi f_c t)] = A_c \cos\left[2\pi f_c t + 2\pi k_f \int_0^t m(t')dt' \right] \tag{9.11}$$

Here the envelope $g(t)$ is the complex lowpass FM signal:

$$g(t) = A_c \exp\left[2\pi k_f \int_0^t m(t')dt' \right] \tag{9.12}$$

whereas before, Re[w] is the real part of w. The process of frequency modulation is illustrated in Figure 9.2.

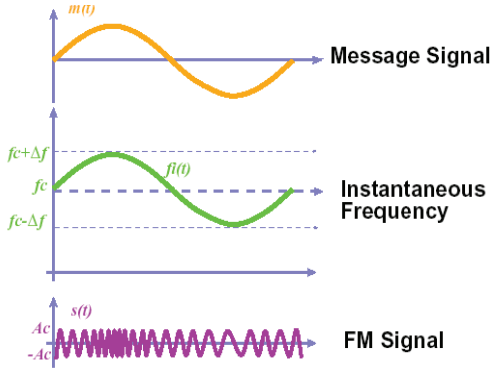


Figure 9.2. The modulating signal (top plot), time-varied modulation frequency (middle plot), and FM signal (bottom plot).

We notice that FM is a constant envelope modulation technique making it suitable for nonlinear amplification. If, for example, the modulating baseband signal has sinusoidal amplitude and frequency, then the FM signal can be expressed as

$$s_{FM}(t) = A_c \cos\left[2\pi f_c t + \frac{k_f A_m}{f_m}\sin(2\pi f_m t)\right] \tag{9.13}$$

In *phase modulation* (PM) signals the angle $\theta(t)$ of the carrier signal is varied linearly with the baseband message signal $m(t)$, and can be presented in the same manner as the FM signal, that is,

$$s_{PM}(t) = A_c \cos\left[2\pi f_c t + k_\theta m(t)\right] \tag{9.14}$$

In (9.14) k_θ is the phase sensitivity of the modulator (the phase deviation constant) measured in *radian per volt*. From (9.11) and (9.14) it follows that an FM signal can be regarded as a PM signal in which the lowpass modulating wave is integrated before modulation. So, an FM signal can be generated by first integrating $m(t)$ and then using the result as an input to a phase modulator. Conversely, a PM signal can be generated by first differentiating $m(t)$ and then using the result as the input to a frequency modulator.

The frequency modulation index defines the relationship between the message amplitude and the bandwidth of the transmitted signal, which is presented in the following form

$$\beta_f = \frac{k_f A_m}{B_f} = \frac{\Delta f}{B_f} \tag{9.15}$$

where, as above, A_m is the peak value of the modulating message signal, Δf is the peak frequency deviation of the transmitter, and B_f is the maximum bandwidth of the modulating lowpass signal (usually B_f is equal to the highest frequency component f_m presented in the modulating signal and simply $\beta_f = \Delta f / f_m$).

The phase modulation index is defined as

$$\beta_\theta = k_\theta A_m = \Delta\theta \tag{9.16}$$

where $\Delta\theta$ is the peak phase deviation of the transmitter.

9.1.3 Spectrum and Bandwidth of FM or PM Signals

Since PM and FM signals have the same form of presentation of modulated signal, we will pay attention to one of them, let us say an FM signal. An FM signal is a nonlinear function of the modulating waveform $m(t)$ and, therefore, the spectral characteristics of $s(t)$ cannot be obtained directly from the spectral characteristics of $m(t)$. However, the bandwidth of $s(t)$ depends on $\beta_f = \Delta f / f_m$. If $\beta_f < 1$, then a narrowband FM signal is generated, where the spectral widths of $s(t)$ and $m(t)$ are about the same, i.e., $2 f_m$. If $\beta_f >> 1$, then a wideband FM signal is generated, where the spectral width of $s(t)$ is slightly greater than $2\Delta f$. For an arbitrary frequency modulation index, the approximate bandwidth of the FM signal (in which this signal has 98% of the total power of the transmitted optical frequency (OF) signal), which continuously limits these upper and lower bounds, is [3–5]

$$B_T = \begin{cases} 2\Delta f\left(1+\dfrac{1}{\beta_f}\right) \approx 2(\beta_f + 1) f_m, & \beta_f < 1 \\ 2\Delta f\left(1+\dfrac{1}{\beta_f}\right) \approx 2\Delta f, & \beta_f >> 1 \end{cases} \tag{9.17}$$

This approximation of FM bandwidth is known as *Carson's rule* [3–5]. It states that for the upper bound, the spectrum of the FM signal is limited to the carrier frequency f_c of the carrier signal, and one pair of sideband frequencies at $f_c \pm f_m$. For the lower bound, the spectrum of the FM signal is simply slightly greater than $2\Delta f$.

There are two variants of FM signals generation, the direct and indirect, as well as many methods of its demodulation by use of different kinds of detectors. This specific subject is out of the scope of the current book, therefore we propose that the reader refers to the special literature [1–5], where these questions are fully described.

9.1.4 Relations Between SNR and Bandwidth of AM and FM Signals

In the angle modulation systems, the signal-to-noise ratio (SNR) *before* detection is a function of the receiver intermediate frequency (IF) filter bandwidth (see sections 9.1.1 and 9.1.2, where optical signal AM and FM modulation is described), of the received carrier power, and of the received interference [3–5], that is,

$$(SNR)_{in} = \frac{A_c^2 / 2}{2N_0(\beta_f + 1)B_F} \tag{9.18}$$

Where A_c is the carrier amplitude, N_0 is the white noise power spectral density, and B_F is the equivalent bandwidth of the bandpass filter at the front end of the receiver. Note that $(SNR)_{in}$ uses the carrier signal bandwidth according to Carson's rule (9.17).

However, the SNR *after* detection is a function of the maximum frequency of the message, f_m, the modulation index, β_f or β_θ, and the given SNR at the input of the detector, $(SNR)_{in}$. For example, the SNR at the output of an FM receiver depends on the modulation index and is given by [5]

$$(SNR)_{out} = 6(\beta_f + 1)\beta_f^2 \left\langle \left(\frac{m(t)}{V_p} \right)^2 \right\rangle (SNR)_{in} \tag{9.19}$$

where V_p is the peak-to-zero value of the modulating signal $m(t)$.

For comparison purposes, let us present here the $(SNR)_{in}$ for an AM signal which, according to [3], is defined as the input power to a conventional AM receiver having bandwidth equaling to $2B_F$, that is,

$$(SNR)_{in,AM} = \frac{P_c}{N} = \frac{A_c^2}{2N_0 B_F} \tag{9.20}$$

Then, for $m(t) = A_m \sin 2\pi f_m t$, equation (9.20) can be simplified to

$$(SNR)_{out,FM} = 3\left(\beta_f + 1\right)\beta_f^2 (SNR)_{in,FM} \tag{9.21}$$

At the same time

$$(SNR)_{out,FM} = 3\beta_f^2 (SNR)_{in,AM} \tag{9.22}$$

Expressions (9.18) and from (9.21) to (9.22) are valid only if $(SNR)_{in}$ exceeds the threshold of the FM detector. The minimum received value of $(SNR)_{in}$, needed to exceed the threshold is around 10dB [3]. Below this threshold, the demodulated signal becomes noisy. Equation (9.21) shows that the SNR at the output of the FM detector can be increased with an increase of the modulation index β_f of the transmitted signal. At the same time, the increase in modulation index β_f leads to an increased bandwidth and spectral occupancy. In fact, for large values of β_f, Carson's rule gives the channel bandwidth of $2\beta_f f_m$. As also follows from (9.21), the SNR at the output of the FM detector is $(\beta_f + 1)$ times greater than the input SNR for an AM signal with the same bandwidth. Moreover, it follows from (9.21) that $(SNR)_{out,FM}$ for FM is much greater than $(SNR)_{out,AM}$ for AM.

Finally, we should notice that, as follows from (9.21), the term $(SNR)_{out,FM}$ increases as a cube of the bandwidth of the message. This clearly illustrates why FM offers very good performance for fast fading signals when compared with AM. As long as $(SNR)_{in,FM}$ remains above threshold, $(SNR)_{out,FM}$ is much greater than $(SNR)_{in,FM}$. A technique called *threshold extension* is usually used in FM demodulators to improve detection sensitivity to about $(SNR)_{in,FM} = 6\text{dB}$ [5].

9.2. Digital Signal Modulation

As was mentioned above, *modulation* is the process where the baseband message information is added to the bandpass carrier. In *digital modulation* the digital beam stream is transmitted as a *message*, and then is converted into the analog signal of the type described by (9.1) that modulates the digital bit stream into a carrier signal. As was mentioned above, the analog signal described by (9.1) has amplitude, frequency, and phase. Changing these three characteristics, we can formulate three kinds of digital modulation. They are [3–10]:

Amplitude shift keying (ASK) for phase and frequency keeping being constant;

Frequency shift keying (FSK) for amplitude and phase keeping being constant;

Phase shift keying (PSK) for amplitude and frequency keeping being constant.

In so-called hybrid modulation methods combinations of these three kinds of modulation are usually used. Namely, if frequency is constant, but amplitude and phase are not constant, quadrature amplitude modulation (QAM) is used. Some modulation methods are linear, as binary phase shift keying (BPSK), quadrature phase shift keying (QPSK), including $\pi/4-$ QPSK, DQPSK and $\pi/4-$DQPSK, and so on. At the same time, FSK as well as, minimum shift keying (MSK) and Gaussian minimum shift keying (GMSK) are nonlinear modulation techniques [3–10].

Because digital modulation offers many advantages over analog modulation, it is often used in modern electro-optical systems. Some advantages include greater noise immunity and robustness to channel impairments, easier multiplexing of various forms of information (such as voice, data, and video), and greater security. Moreover, digital transmissions accommodate digital error-control codes, which detect and correct transmission errors, and support complex signal processing techniques such as coding, encryption, etc. (see [12]).

9.2.1 Types of Linear Digital Modulation Techniques

We present now only a few examples of such kinds of modulation, transferring the reader to the excellent books [4–10].

Linear Modulation. This is a type of modulation where the amplitude of the transmitted signal varies linearly with the modulating digital signal $m(t)$ according to the following law:

$$\begin{aligned} s(t) &= \mathrm{Re}\left[A\,m(t)\exp(j2\pi f_c t)\right] \\ &= A\left[m_R(t)\cos(2\pi f_c t) - m_I(t)\sin(2\pi f_c t)\right] \\ m(t) &= m_R(t) + jm_I(t) \end{aligned} \tag{9.23}$$

This kind of modulation has a good spectral efficiency, but linear amplifiers have *poor* power efficiency [3–10]. Side lobes are generated, increasing adjacent channel interference and canceling the benefits of linear modulation.

Amplitude Shift Keying (*ASK) Modulation.* This is a modulation where keying (or switching) the carrier sinusoid *on* if the input bit is "1" and *off* if "0" (so-called On-Off-Keying-OOK [3, 5, 10]). This kind of modulation is shown in Figure 9.4.

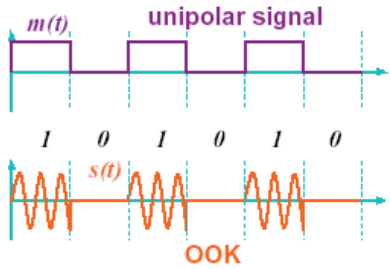


Figure 9.4. The message $m(t)$ unipolar signal [top plot] and the baseband modulated OOK signal [bottom plot].

Binary Phase Shift Keying (*BPSK*) *Modulation*. BPSK modulated signals $g_1(t)$ and $g_2(t)$ can be presented as:

$$g_1(t) = \sqrt{\frac{2W_b}{T_b}} \cos(2\pi f_c t), \qquad 0 \le t \le T_b \tag{9.24a}$$

and

$$g_2(t) = -\sqrt{\frac{2W_b}{T_b}} \cos(2\pi f_c t), \qquad 0 \le t \le T_b \tag{9.24b}$$

where W_b is the energy per bit, T_b is the bit period, and a rectangular pulse shape $p(t) = \Pi\left((t - T_b / 2) / T_b\right)$ is assumed. Basis signals ϕ_i for this signal, setting in 2D-vector-space, simply contain a single wave form ϕ_1, where

$$\phi_1(t) = \sqrt{\frac{2}{T_b}} \cos(2\pi f_c t), \qquad 0 \le t \le T_b \tag{9.25}$$

The result of such a kind of modulation is presented in Figure 9.5.

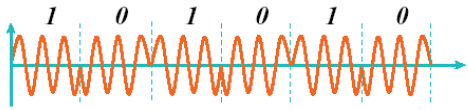


Figure 9.5. BPSK signal presentation.

Using this basis signal, the BPSK signal set can be represented as

$$\mathbf{g}_{iBPSK} = \left\{ \sqrt{W_b}\varphi_1(t), -\sqrt{W_b}\varphi_1(t) \right\} \tag{9.26}$$

Such a mathematical representation of a vector, consisting of two points that are then placed at the *constellation diagram*, as shown in Figure 9.6, provides a graphical representation of the complex envelope of each possible symbol state. The distance between signals on a constellation diagram relates to how different the modulation waveforms are and how well a receiver can differentiate between all possible symbols when random noise is present.

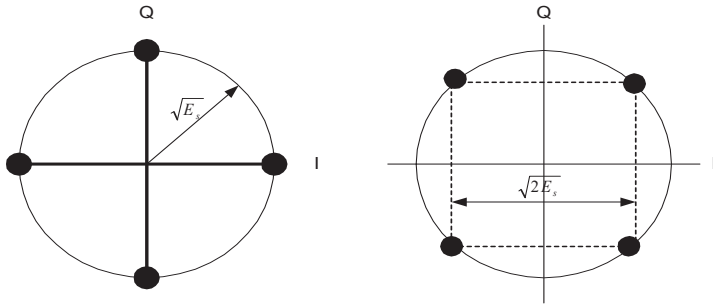


Figure 9.6. Constellation diagram of QPSK and $\pi/4-$QPSK modulated signals.

As was mentioned in Chapter 8, the number of basis signals will always be less than or equal to the number of signals in the set. The number of basis signals required to represent the complete modulation signal set is

called the *dimension* of the vector space (in the example above - it is two-dimensional (2-D) vector space). If there are many basis signals in the modulation signal set, then all of them must be orthogonal according to (9.24).

Quadrature Phase Shift Keying (*QPSK*) *Modulation*. The QPSK signal has the advantage that it has twice the bandwidth efficiency or two bits at a time

$$\begin{aligned} s_{QPSK}(t) &= \sqrt{\frac{2E_s}{T_s}}\cos\left[2\pi f_c t + i\frac{\pi}{2}\right] \qquad 0 \le t \le T_s \quad i = 0,1,2,3 \\ &= \sqrt{\frac{2E_s}{T_s}}\cos\left(i\frac{\pi}{2}\right)\cos(2\pi f_c t) - \sqrt{\frac{2E_s}{T_s}}\sin\left(i\frac{\pi}{2}\right)\sin(2\pi f_c t) \end{aligned} \tag{9.27}$$

This signal set is shown geometrically in Fig. 9.5, where the left diagram is for pure QPSK and the right one for $\pi/4-$QPSK modulation, that is with angle shift at $\pi/4$.

9.2.2 Nonlinear Digital Modulation

As was mentioned from the beginning of this section, the frequency shift keying signals are examples of a nonlinear type of digital modulation. We shall briefly describe it.

Frequency Shift Keying (*FSK*) *Modulation*. FSK modulated signals, where switching the carrier sinusoid frequency f_c to $f_c - \Delta f$ occurs, if the input bit is "0", and to $f_c + \Delta f$, if input bit is "1". Results of modulation are shown in Figure 9.7. Finishing this chapter, we should notice that usage of each kind of modulation depends on the conditions of propagation inside a channel, effects of fading inside, and on what kinds of detectors and corresponding filters are used.

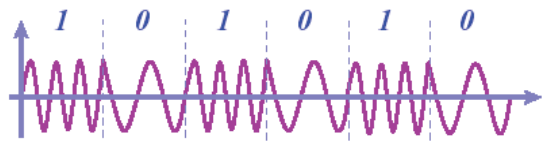


Figure 9.7. FSK modulated signal presentation.

All these aspects are fully described in excellent books [3–5, 7–10].

Exercises

Exercise 1.

A zero mean (<$m(t)$>=0) sinusoidal message is applied to a transmitter that radiates the AM signal with a power of 10 kW. The modulation index $k_m = 0.6$.

Find: 1) The carrier power.
2) What is it as a percentage of the total power in the carrier?
3) What is the power of each sideband?

Solution:

1) $P_c = \dfrac{P_{AM}}{1 + k_m^2 / 2} = \dfrac{10}{1 + 0.18} = 8.47 \ kW$

2) $\dfrac{P_c}{P_{AM}} \cdot 100\% = \dfrac{8.47}{10} \cdot 100\% = 84.7\%$

3) $\dfrac{1}{2}(P_{AM} - P_c) = 0.5 \cdot (10 - 8.47) = 0.765 \ kW$

Exercise 2.

A sinusoidal modulating signal, $m(t) = 4\cos 2\pi f_m t$, that is, $f_m = 4 \ kHz$ and the maximum amplitude $A_m = 4$ Volt, is applied to an FM modulator which has a frequency deviation constant gain $k_f = 10 \ kHz / Volt$.

Find: 1) The peak frequency deviation, Δf
2) The modulation index, β_f.

Solution:

1) $\Delta f = k_f \cdot A_m = 4 \ V \cdot 10 \ kHz / V = 40 \ kHz$

2) $\beta_f = \dfrac{\Delta f}{f_m} = \dfrac{40 \ kHz}{4 \ kHz} = 10$

Example 3.

A frequency modulated signal with the carrier frequency $f_c = 880\ MHz$ and with sinusoidal modulating waveform of $f_m = 100\ kHz$ has a peak deviation $\Delta f = 500\ kHz$.

Find: The receiver bandwidth necessary to pass such a signal.

Solution:

The modulation index equals: $\beta_f = \Delta f / f_m = 500/100 = 5$

According to Carson's rule (9.21)

$$B_T = 2(\beta_f + 1) f_m = 2(5+1)\cdot 100\ kHz = 1200\ kHz$$

Exercise 4.

An FM signal with $f_m = 5\ kHz$ has modulation index $\beta_f = 3$.

Find: 1) The bandwidth required for such an analog frequency modulation.

2) How much output SNR improvement would be obtained if the modulation index is increased to $\beta_f = 5$. What is the trade-off bandwidth of this improvement?

Solution:

1) for $\beta_f = 3$: $B_T = 2\cdot(\beta_f + 1)\cdot f_m = 2\cdot(3+1)\cdot 5\ kHz = 40\ kHz$

   for $\beta_f = 5$: $B_T = 2\cdot(\beta_f + 1)\cdot f_m = 2\cdot(5+1)\cdot 5\ kHz = 60\ kHz$

2) from (9.24) the output SNR improvement factor is approximately for $3\beta_f^3 + 3\beta_f^2$, that is,

for $\beta_f = 3$: $3\beta_f^3 + 3\beta_f^2 \approx 3\cdot(3)^3 + 3\cdot(3)^2 = 108 \Rightarrow 20.33\ dB$

for $\beta_f = 5$: $3\beta_f^3 + 3\beta_f^2 \approx 3\cdot(5)^3 + 3\cdot(5)^2 = 450 \Rightarrow 26.53\ dB$

Therefore, the improvement in output SNR by increasing the modulation index from 3 to 5 is 26.53–20.33 = 6.2 *dB*.

This improvement is achieved at the expense of bandwidth (1.5 times wider): for $\beta_f = 3$ is $B_T = 40\ kHz$ and for $\beta_f = 5$ is $B_T = 60\ kHz$.

Bibliography

[1] Jakes, W. C. 1974. *Microwave Mobile Communications*. New Jersey: IEEE Press.
[2] Steele, R. 1992. *Mobile Radio Communication*. New Jersey: IEEE Press.
[3] Rappaport, T. S. 1996. *Wireless Communications*. New York: Prentice Hall PTR.
[4] Stuber, G. L. 1996. *Principles of Mobile Communication*. Boston-London: Kluwert Academic Publishers.
[5] Couch, L. W. 1993. *Digital and Analog Communication Systems*. Macmillan, New York.
[6] Lusignan, B. B. 1978. "Single-sideband transmission for land mobile radio." *IEEE Spectrum* July: 33–37.
[7] Ziemer, R. E., and R. L. Peterson. 1992. *Introduction to Digital Communications*. New York: Macmillan Publishing Co.
[8] Saunders, S. R. 1999. *Antennas and Propagation for Wireless Communication Systems*. New York: John Wiley & Son.
[9] Proakis, J. G. 1989. *Digital Communications*. McGraw-Hill: New York.
[10] Krouk, E., and S. Semenov, eds. 2011. *Modulation and Coding Techniques in Wireless Communications*. Chichester, England: Wiley.
[11] Dakin, J., and B. Culshaw, eds. 1988. *Optical Fiber Sensors: Principles and Components*, vol. 1. Boston-London: Artech House.
[12] Blaunstein, N., S. Engelberg, E. Krouk. and M. Sergeev. 2020. *Fiber Optic and Atmospheric Optical Communication*. Hoboken, NJ: IEEE Press & Wiley.

CHAPTER 10

OPTICAL WAVES PROPAGATION IN FIBEROPTIC STRUCTURES

Below, we pay attention to the description of optical wave propagation in fiber optic structures, the dispersive properties of optical signals caused by non-homogeneous material phenomena, and multimode propagation of optical signals in such kinds of wired links. We illustrate these phenomena, based on the corresponding computational results obtained below for such guiding optical structures accounting for arbitrary refractive indices of the inner (core) and outer (cladding) elements of the optical cable and on the features accompanying propagation of light inside fiber. In our discussions, we will follow the corresponding literature [1–6].

10.1. Types of Optical Fibers

The fiber optic 3-D guiding structure consists of two parts: the inner, called the *core*, and the outer, called the *cladding* (see Figure 10.1). Light propagates inside the core which guides optical modes inside the optical cable.

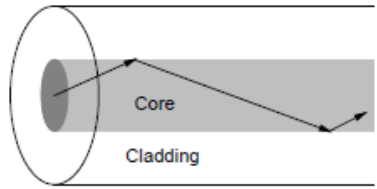


Figure 10.1. Two structures, inner and outer of 3-D optical fiber, called the core and the cladding, respectively.

The first commonly used kind of fiber optic structure is the *step-index fiber* (see Figure 10.2, left panel).

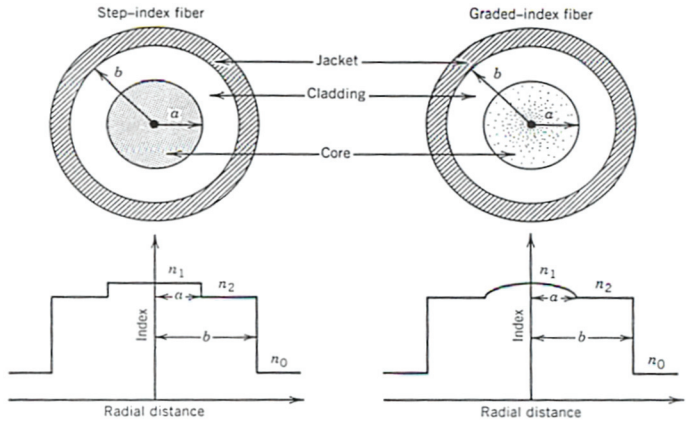


Figure 10.2. Difference between the refractive index profiles for step-index and graded-index fibers.

As clearly seen from Figure 10.2, such fibers consist of a central core of radius a and refractive index n_1, surrounded by a cladding of radius b and refractive index n_2.

According to the definition of Total Intrinsic Reflection (TIR) (see definition in Chapter 2), to obtain the total reflection from the cladding, its refractive index should be lower than that for the core, i.e., $n_1 > n_2$. Figure 10.1 shows the geometry of optical ray propagation within the core on the assumption that the cladding width is thick enough to exclude the evanescent field decay inside the cladding depth. So, from the beginning we can suppose that the effects of a finite cladding thickness are negligible, and a ray field is small enough to penetrate to the outer edges of the cladding. As will be described below, in multimode *step-index fiber* a large modal distortion occurs.

To avoid such drawbacks of this kind of fiber, a new type, called *graded-index fiber*, was developed [1–6] that has the same configuration as the previous fiber, and is shown in Figure 10.2, right panel. The difference between both kinds of fiber is defined by differences in the profiles of the refractive indexes of the core and cladding, as illustrated in Figure 10.2. Thus, as clearly seen from the illustrations, in the *step-index* fiber the index change at the core-cladding interface is abrupt, whereas in the *graded-index* fiber the refractive index decreases gradually inside the core.

To understand the effects of optical wave propagation in both kinds of fibers, let us introduce the main operational parameters of the fiber optic guiding structures usually used in electro-optics and optical engineering.

10.2. Main Operational Parameters of Optical Fibers

In fiber optics, there is an important parameter that is usually used, called the *numerical aperture* of the fiber optic guiding structure, denoted as N.A. [1–6]

$$N.A. = n_1 \sin\theta_c \equiv \sin\theta_a \tag{10.1}$$

where $\theta_{full} = 2 \cdot \theta_a$ is called in the literature the *angle of minimum light energy spread outside the cladding* or *angle of full transfer of optical energy along the core* [1–6], when total internal reflection (TIR) occurs in a fiber optic structure. Accounting for $\cos^2\theta = 1 - \sin^2\theta$, we finally get

$$N.A. = \left(n_1^2 - n_2^2\right)^{1/2} \tag{10.2}$$

The second parameter usually used in fiber optic physics is the *relative refractive index* difference [1–6]. It has two definitions depending on the type of fiber optic structure, as shown in Figure 10.2.

Thus, for the *graded-index* fiber:

$$\Delta = \frac{\left(n_1^2 - n_2^2\right)}{n_1^2} \equiv \frac{\left(N.A\right)^2}{n_1^2} \tag{10.3}$$

Using the above formulas, we can find relations between these two engineering parameters for the fiber optic with a graded-index refractive index profile (see Figure 10.2):

$$N.A. = n_1 \cdot (\Delta)^{1/2} \tag{10.4}$$

For the *step-index* fiber optic:

$$\Delta = \frac{n_1 - n_2}{n_1} \tag{10.5}$$

As follows from the geometry of the refractive index profile presented in Figure 10.2 for the step-index fiber optic structure, the relation between N.A., according to (10.3) and Δ according to (10.5), is not so trivial, as (10.4) for this kind of optical cable.

10.3. Propagation of Optical Rays in a 2-D Plane Dielectric Guiding Structure

Before entering into discussions of optical wave propagation inside a 3-D fiber optic structure, let us consider the simpler case of a 2-D plane dielectric guiding structure, a *slab*, shown in Figure 10.3 on the basis of geometrical optic presentation of rays and the corresponding Snell's laws, as was seen in Chapter 2. Such a simplified presentation of a 3-D fiber optic structure, presented in Figure 10.1, can model the plane core structure covered by the plane cladding structure, as shown in Figure 10.3.

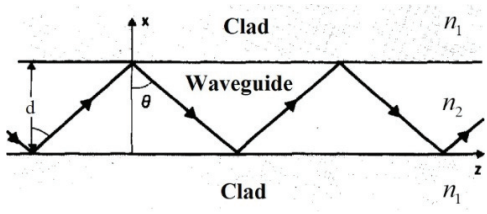


Figure 10.3. 2-D plane model of the 3-D fiber optic structure shown in Figure 10.1.

In such a slab, for the description of the guiding modes of propagation, either the transverse electric (TE) or vertical polarized or the transverse magnetic (TM) or horizontal polarized [1–6], a new parameter is always used called the normalized frequency, denoted by V and defined as:

$$V = 2\pi d \cdot \ N.A. = \left(n_1^2 - n_2^2\right)^{1/2} / 2\lambda_0 \tag{10.6}$$

where all parameters in Eq. (10.6) are shown in Figure 10.2 or defined above in Section 10.1.

We will discuss TE and TM modes later, but now, after a simplified assumption, we will present a description of such modes propagation for these two kinds of wave polarization, vertical and horizontal, briefly introduced in Chapter 2.

The number of wave guiding modes propagating along such a dielectric slab, according to geometrical optic postulates, corresponds to the number of specular reflections (regulated by Snell's second law, see Chapter 2), as it is clearly illustrated in Figure 10.4 for TE modes propagation along the slab.

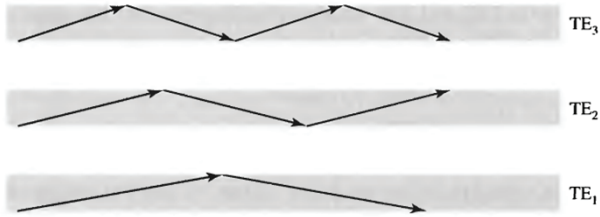


Figure 10.4. Illustration of coincidence between numbers of reflection and indexes of TE modes.

There are two kinds of optical wave propagating inside the guiding structure, *one-mode*, and *multiple-mode*, as shown in Figure 10.5. In the first case, only one ray propagates inside the optical structure with multiple reflections from both its boundaries (left panel) according to Snell's law (see Chapter 2), whereas in the second case, many rays are reflecting from the upper and lower boundaries of the optical structure and propagate inside it (right panel).

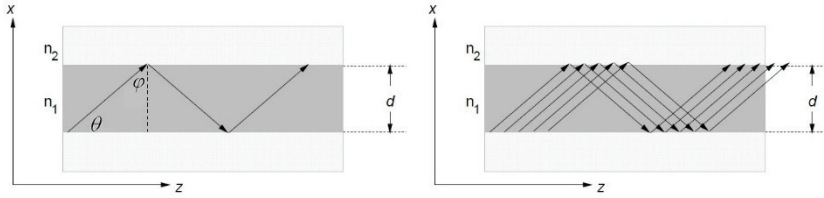


Figure 10.5. Presentation of single-mode and multiple-mode propagation inside a 2-D slab.

In Figure 10.5, which models the real 2-D fiber optic structure, n_1 is the refractive index of the core and n_2 is the refractive index of the cladding (illustrated by Figure 10.2). In an ideal case (i.e., without losses), propagation of a ray (left panel) and the rays (right panel) occur both along the vertical (along the x-axis) and the horizontal (along the z-axis) without any losses, if the law of total intrinsic reflection is satisfied (see Chapter 2),

according to which, the incident angle must exceed the critical angle of TIR, that is:

$$\varphi > \varphi_c = \text{arcsine}\ (n_2 / n_1) \tag{10.7}$$

Let us now consider the physical meaning of multi-ray propagation inside a 2-D slab, that models the fiber optic structure presented in Figure 10.1, based on a simple geometrical optic presentation and on Snell's TIR law. Thus, the multimode propagation can be described via wave vector **k** and its components, k_x and k_z, along the x-axis and z-axis, respectively (as illustrated in Figure 10.6):

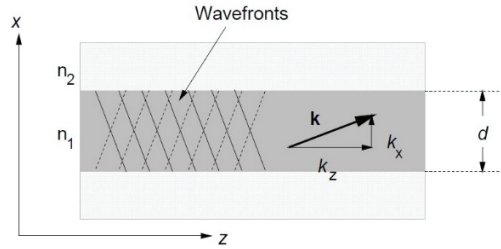


Figure 10.6. Geometrical presentation of multi-ray propagation with the wave vector **k** via its components along *x*-axis and *z*-axis.

$$\boldsymbol{k} = \mathbf{k}_z + \mathbf{k}_x = k_z\,\mathbf{z} + k_x\mathbf{x} \tag{10.8}$$

where

$$k_x = k\cos\varphi = k\sin\theta = (2\pi\, n_1/\lambda)\sin\theta \tag{10.9}$$

During propagation, each mode after its reflection obtains the phase difference not only $2k_x$, but also the angle δ, which is the result of differences between the inner and outer refractive indexes, n_1 and n_2, inside the Fresnel coefficients which depend on the phase difference and the wave polarization, vertical or horizontal (see Chapter 2). According to the *transverse resonance condition* of total intrinsic reflection of rays with vertical polarization (e.g., for *TE* – modes)

$$2k_x d - 2\delta = 2\pi m \tag{10.10a}$$

Or

$$4\pi n_1 \, d \sin\theta_m / \lambda - 2\delta = 2\pi m \tag{10.10b}$$

In Eq. (10.10b), angle θ was denoted by θ_m because for each mode with number m (e.g., for each of m reflections), it has its own separate meaning and number. Accounting now for $\theta + \varphi = \pi / 2$, after straightforward computations, we finally get:

$$2\delta = 4\tan^{-1}\sqrt{\frac{\sin^2\varphi_c}{\sin^2\theta} - 1} \tag{10.11}$$

Substituting expression (10.11) in Eq. (10.10b), yields:

$$\frac{4\pi n_1 d}{\lambda}\sin\theta_m - 4\tan^{-1}\sqrt{\frac{\sin^2\varphi_c}{\sin^2\theta_m} - 1} = 2\pi m$$

The latter equation can be rearranged in a more convenient form:

$$\tan^2\left(\frac{\pi n_1 d}{\lambda}\sin\theta_m - \frac{\pi}{2}m\right) = \frac{\sin^2\varphi_c}{\sin^2\theta_m} - 1 \tag{10.12}$$

Equation (10.12) has a physically vivid explanation. Thus, the *left-hand side* (denoted by LHS) of Eq. (10.12) gives the roots of trigonometrical expressions of tangents via the mode parameter m, as shown by light points along the horizontal axis (indicated by $\sin\theta$ in Figure 10.7), whereas the *right-hand side* (denoted as RHS) of Eq. (10.12) gives the roots of trigonometrical expression via $\sin\theta$ in Figure 10.7 lining the discrete curve, which monotonically decreases (with an increase of m from 1 to 8) until the condition of $\sin\theta = \sin\varphi_c$. The two sides of Eq. (10.12) are equal in the cases when the RHS curves cross the LHS curves (for each m from 0 to 8) at the bold points denoted by a dark color.

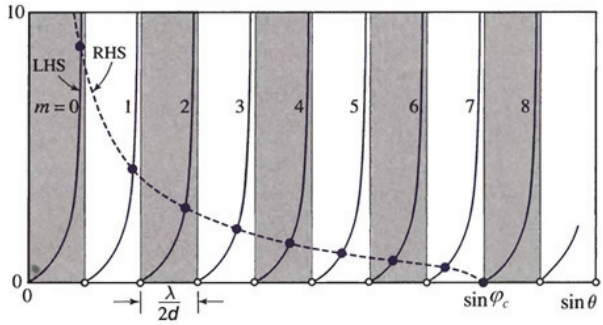


Figure 10.7. Geometrical presentation of the solutions of Eq. (10.12) depicted by bold points where the left-hand side (LHS) of this equation equals the right-hand side (RHS) of this equation; the difference between roots of LHS equals $\lambda/2d$ and is depicted by light points.

It is clearly seen that after mode number m>8, that is, when $\sin\theta_m$ exceeds $\sin\varphi_c$, the law of TIR is valid, and the further propagation of optical modes (with m>8) without loss of energy due to penetration in the outer region (with n_2) becomes unacceptable. We should also notice that (10.12) has roots for m = 0, 2, 4,.., i.e., *even*, with solutions of its LHS of $\tan(\pi d n_1 \sin\varphi / \lambda)$, and roots for m = 1, 3, 5…, i.e., *odd*, with solutions of $\cot(\pi d n_1 \sin\varphi / \lambda)$.

So, crossing bold points for even and odd modes of number m enables finding $\sin\theta_m$ for each m which are regulated by condition $\sin\theta \le \sin\varphi_c$, and, finally, allow obtaining the component of the wave number along the z-axis, that is,

$$k_z = \beta_m = n_1 k \cos\theta_m \tag{10.13}$$

A step between two light points (i.e., between two boundaries, left and right, of shadow vertical plates) was denoted in Figure 10.6, as $\lambda/2d\,n_1$, and a number of solutions, N, when LHS = RHS in Eq. (10.12), can be found according to the above condition $\sin\theta \le \sin\varphi_c$ and geometrical consideration $\sin\theta = (N\cdot\lambda)/(2d\,n_1)$. Then, we get:

$$(N\cdot\lambda)/(2d\cdot n_1) \le \sin\varphi_c \tag{10.14a}$$

or

$$N \leq \sin\varphi_c / (\lambda / 2d \cdot n_1) \tag{10.14b}$$

Accounting now for definition (10.7) for φ_c and that $\sin\varphi_c = (1 - \cos^2\varphi)^{1/2}$ $= [1-(n_2/n_1)]^{1/2}$, we finally get:

$$N = \{(2d\, n_1/\lambda_0)\, [1-(n_2/n_1)^2]^{1/2} + 1\} = [(2d/\lambda_0)\, (n_1^2 - n_2^2)^{1/2} + 1] \tag{10.15a}$$

Here, at the right side of (10.15a), and in further expressions for N, the rectangular parentheses indicate the integer part of the number N. This expression can be rewritten via the parameter numerical aperture, defined by (10.2), which is useful for future engineering computation of the relation between N and NA:

$$N = [NA \cdot (2d/\lambda_0) + 1] \tag{10.15b}$$

Remembering the relation between $N.A.$ and the normalized frequency parameter V, defined by expression (10.6), we can present the additional useful engineering formula that defines N via V:

$$N = \left[2d \frac{\sqrt{n_1^2 - n_2^2}}{\lambda_0} + 1 \right] = \left[\frac{1}{\pi} \frac{2\pi}{\lambda_0} d \sqrt{n_1^2 - n_2^2} + 1 \right] = \left[\frac{2V}{\pi} + 1 \right] \tag{10.15c}$$

accounting now for relations between NA and V

$$V = k \frac{d}{2} \sqrt{n_1^2 - n_2^2} = \frac{2\pi}{\lambda_0} d \sqrt{n_1^2 - n_2^2} = \frac{\pi d}{\lambda_0} NA \tag{10.16}$$

So, based on the simple geometrical optic model of ray modes propagation inside the 2-D guiding structure that models the real case of a 3-D optical fiber, we presented several variants of how to find a number of propagating ray modes' solutions N (according to Eq. (10.12)), through the knowledge of different operational parameters introduced above for the description of fiber optic guiding structures.

10.4. Propagation of Optical Wave Along the 3-D Fiber Optic Structure

Let us now consider the cylindrical dielectric fiber optic structure as shown in Figure 10.8. This is just the geometry of the optical fiber, where the central region is known as the *core* and the outer region as the *cladding*. In this case, the same basic principles, as for the dielectric slab, but the circular rather than planar symmetry changes the mathematics. We use the solution of Maxwell's equation in the cylindrical coordinates for both the coaxial cable and the circular waveguide, where we deal mostly with guiding modes rather than the ray concept [1–4, 6].

The wave equation that describes such propagation of light within a cylindrical waveguide can be presented in cylindrical coordinates as follows for $\mu_r = 1$:

$$\frac{\partial^2 E_z}{\partial r^2} + \frac{1}{r}\frac{\partial E_z}{\partial r} + \frac{1}{r^2}\frac{\partial^2 E_z}{\partial \phi^2} + \frac{\partial^2 E_z}{\partial z^2} + n^2 k_0^2 E_z = 0 \tag{10.17}$$

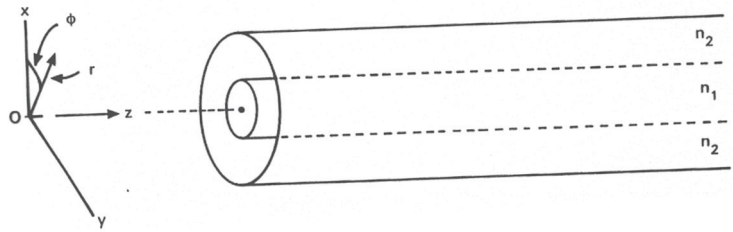


Figure 10.8. Presentation of fiber optic structure in a 3-D cylindrical coordinate system.

We can present the solution taking into account the separation of variables:

$$E_z(r,\phi,z) = F(r)\Phi(\phi)Z(z) \tag{10.18}$$

From the well-known mathematical approaches [2-5] by use of the independence of each separated variable function, we immediately get for each of them its own simplified equation with the corresponding solution:

$$\frac{d^2Z}{dz^2}+\beta^2 Z=0 \quad \longrightarrow \quad Z=\exp(j\beta z)$$
$$\frac{d^2\Phi}{dl^2}+l^2\Phi=0 \quad \longrightarrow \quad \Phi=\exp(jl\phi) \tag{10.19}$$

As for coordinate r, the equation for it is more complicated and can be presented as

$$\frac{d^2F}{dr^2}+\frac{1}{r}\frac{dF}{dr}+\left(n^2k_0^2-\beta^2-\frac{l^2}{r^2}\right)F=0 \tag{10.20}$$

where l is an azimuthal integer. Equation (10.20) has the form of Bessel's equation, and its solutions are Bessel functions [7, 8]. We finally can obtain solutions for the field of rays through the modified Bessel functions of first and second order, $J(qr)$ and $K(pr)$, via wave parameters h and q as propagation parameters inside the core and cladding, respectively. These parameters can be presented in the following form [2-5]:

$$h^2=n_1^2k_0^2-\beta^2$$
$$q^2=\beta^2-n_2^2k_0^2 \tag{10.21}$$

This finally gives at the core ($r \le a$) and at the cladding ($r>a$) the following expressions for z0 components of the electric and magnetic fields:

$$E_z=\begin{cases} AJ_l(hr)\exp(jl\phi)\exp(j\beta z) & , \quad r\le a \\ CK_l(qr)\exp(jl\phi)\exp(j\beta z) & , \quad r>a \end{cases} \tag{10.22a}$$

$$H_z=\begin{cases} BJ_l(hr)\exp(jl\phi)\exp(j\beta z) & , \quad r\le a \\ DK_l(qr)\exp(jl\phi)\exp(j\beta z) & , \quad r>a \end{cases} \tag{10.22b}$$

As for other components of the total field, they can be presented through the z-components in the following manner:

$$E_r = \frac{j}{h^2}\left(\beta\frac{\partial E_z}{\partial r} + \mu_0\frac{\omega}{r}\frac{\partial H_z}{\partial \phi}\right)$$
$$E_\varphi = \frac{j}{h^2}\left(\frac{\beta}{r}\frac{\partial E_z}{\partial \phi} - \mu_0\omega\frac{\partial H_z}{\partial r}\right) \tag{10.23}$$

The solution for each component can be expressed via the corresponding Bessel $J_l(hr)$ function of the first kind and via the modified Hankel $K_l(hr)$ function (e.g., Bessel function of the second kind [7, 8]). Namely, for r-components of the field solutions can be presented in the following form [7, 8]:

$$E_r = E_c J_l(hr) \tag{10.24a}$$

$$E_r = E_{cl} J_l(hr) \tag{10.24b}$$

Roots of $J_l(hr) = J(v)$, $l = 0, 1, 2,\ldots$, are shown in Figure 10.9.

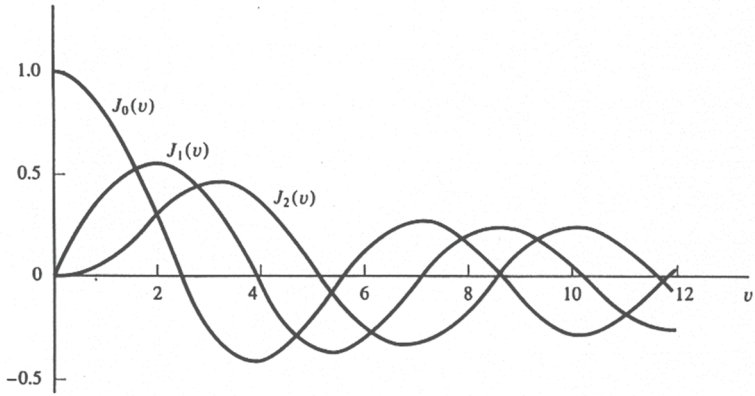


Figure 10.9. Bessel function of the first kind vs. variable v.

At the boundary between the core and the cladding, one can obtain relations between the first order roots of these specific functions and the parameters of propagation described by (10.21) and the refractive index of the core and the cladding respectively, i.e.,

$$\left[\frac{J_l'(ha)}{hJ_l(ha)}+\frac{K_l'(qa)}{qK_l(qa)}\right]\cdot\left[\frac{J_l'(ha)}{hJ_l(ha)}+\frac{n_2^2}{n_1^2}\frac{K_l'(qa)}{qK_l(qa)}\right]=\frac{l^2}{a^2}\left(\frac{1}{h^2}+\frac{1}{q^2}\right)\left(\frac{1}{h^2}+\frac{n_2^2}{n_1^2}\frac{1}{q^2}\right) \tag{10.25}$$

The solution of the first kind of Hankel function can be presented for qr >>1, as [8. 9]:

$$K_l(qr)=\sqrt{\frac{\pi}{2qr}}\exp(-qr) \quad , \quad qr\gg 1 \tag{10.26}$$

For practical applications the effective refractive index is usually introduced

$$n_{eff}=\frac{\beta}{n_2} \tag{10.27}$$

$$n_1>n_{eff}>n_2$$

and the normalized mode propagation constant

$$b=\frac{n_{eff}-n_2}{n_1-n_2} \tag{10.28}$$

Now, according to Ref. [9], we can determine, for an optical fiber, the corresponding values for given propagation parameters k (in free space) and β (inside core), by imposing the boundary conditions at $r = a$. The result is a relationship that provides the β versus k or *dispersion curves* shown in Figure 10.10.

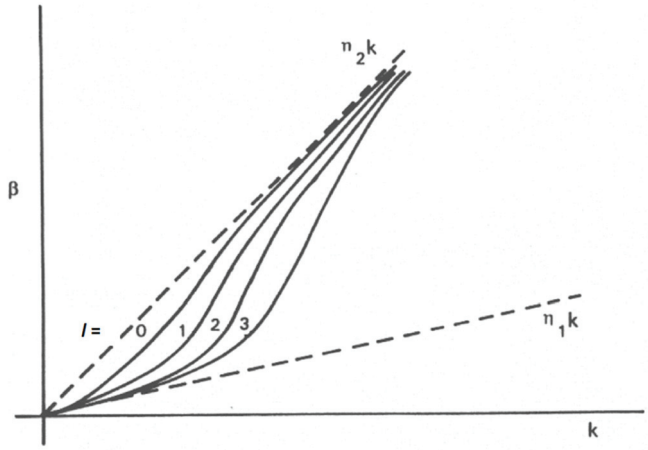


Figure 10.10. Dispersion diagram of optical modes in fiber optic structure.

It is clearly seen that the modes with numbers from $l = 0$ to $l = 3$ (the same property occurring for modes with numbers $l > 3$) propagate between the core upper and cladding lower boundaries of the fiber with wavelengths depending on the refractive properties of these two fiber structures. It is vividly seen that with an increase of wave propagation number k, these modes with increasing number l propagate inside the inner (core) structure.

It should be noticed that the full mathematical approach is very complicated, and so-called "weakly guiding" approximation can only be used for analysis of the processes occurring with light in a fiber optic cable [1, 2]. This makes use of the fact that if $n_1 \approx n_2$ the ray's angle of incidence at the boundary "core-cladding" must be very large, if TIR is to occur. The ray must bounce down the core almost at grazing incidence. This means that the wave is very nearly a transverse wave, with very small z-components. Let us briefly consider the weak guiding approximation of modes propagation in a fiber optic structure, which satisfies the following conditions:

$$n_1 \approx n_2 \longrightarrow \Delta \ll 1 \tag{10.29}$$

Accounting for relations between the parameters of propagation described by Eq. (10.21) and accounting for the constraint

$$h, q \ll \beta$$

we can present the components of the electric and magnetic fields in the 3-D Cartesian coordinate system (see Fig. 10.8) inside and outside the core as follows:

$$E_x = \begin{cases} AJ_l(hr)e^{jl\phi}e^{j(\omega t-\beta z)} & , \quad r < a \\ BK_l(qr)e^{jl\phi}e^{j(\omega t-\beta z)} & , \quad r > a \end{cases} \tag{10.30a}$$

$$E_z = \begin{cases} j\dfrac{h}{\beta}\dfrac{A}{2}\left[J_{l+1}(hr)e^{j(l+1)\phi} - J_{l-1}(hr)e^{j(l-1)\phi}\right]e^{j(\omega t-\beta z)} & , \quad r < a \\ j\dfrac{q}{\beta}\dfrac{B}{2}\left[K_{l+1}(qr)e^{j(l+1)\phi} + K_{l-1}(qr)e^{j(l-1)\phi}\right]e^{j(\omega t-\beta z)} & , \quad r > a \end{cases} \tag{10.30b}$$

$$H_y = \begin{cases} \dfrac{\beta}{\omega\mu}AJ_l(hr)e^{jl\phi}e^{j(\omega t-\beta z)} & , \quad r < a \\ \dfrac{\beta}{\omega\mu}BK_l(qr)e^{jl\phi}e^{j(\omega t-\beta z)} & , \quad r > a \end{cases} \tag{10.30c}$$

$$H_z = \begin{cases} \dfrac{h}{\omega\mu}\dfrac{A}{2}\left[J_{l+1}(hr)e^{j(l+1)\phi} + J_{l-1}(hr)e^{j(l-1)\phi}\right]e^{j(\omega t-\beta z)} & , \quad r < a \\ \dfrac{q}{\omega\mu}\dfrac{B}{2}\left[K_{l+1}(qr)e^{j(l+1)\phi} - K_{l-1}(qr)e^{j(l-1)\phi}\right]e^{j(\omega t-\beta z)} & , \quad r > a \end{cases} \tag{10.30d}$$

$$E_y = H_x \approx 0, \tag{10.30e}$$

Since the waves within the fiber were considered to be transverse [1–4], the solution can be resolved conveniently into two linearly polarized components, just as for free-space propagation. The modes are called *linearly polarized* (LP) modes [1–4, 6]. All solutions obtained above relate directly to the optical fiber guiding structures. The latter has just the cylindrical geometry shown in Figure 10.8 for a typical fiber with the core radius $r = a$, for which it is supposed that a system of equations (10.30) is valid accounting for a "weakly guiding" approximation according to Refs. [1, 2]. The corresponding system (10.30) has two solutions for regular and modified Bessel functions of l-kind (see Figure 10.9):

$$h\frac{J_{l+1}(ha)}{J_l(ha)} = q\frac{K_{l+1}(qa)}{K_l(qa)} \qquad h\frac{J_{l-1}(ha)}{J_l(ha)} = -q\frac{K_{l-1}(qa)}{K_l(qa)} \tag{10.31}$$

Let us introduce two variables:

$$X = ha \qquad Y = qa = V - X^2 \tag{10.32}$$

In this case from the left side equation of (10.31) for $l = 0$, we get:

$$ha\frac{J_1(ha)}{J_0(ha)} = qa\frac{K_1(qa)}{K_0(qa)} \tag{10.33}$$

Generally, for each $i = 0,1,2,\ldots$ yields

$$. X_i = h_i a = a\left(n_1^2 k_0^2 - \beta_i^2\right) \tag{10.34}$$

Or accounting for definition (10.16) of the normalized frequency V, and converting variables X_i via V, taking V = 10, the following dependence can be presented by Figure 10.11.

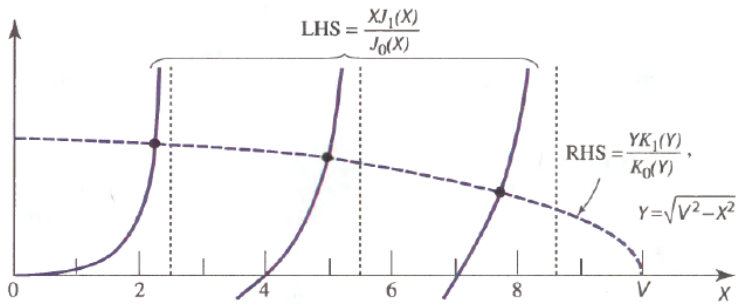


Figure 10.11. Solutions of the left and the right sides of Eq. (10.33) vs. *X* for *V*=10.

The abbreviations "LHS" and "RHS" , as above in Figure 10.7, indicate the left-hand side (LHS) and the right-hand side (RHS) of Eq. (10.33), respectively, where the solution of the LHS is presented by a set of the exponential functions (bold curves) and the solution of the RHS is presented by a dashed curve. Their crossings give points where, for given

parameters of X (or V) along the horizontal axis and of Y along the vertical axis, the corresponding guiding modes of specific linear polarization can propagate inside a core in the fiber optic structure. We should notice that we obtained the same behavior of optical modes propagation inside a slab, as the 2-D guiding structure (see Figure 10.7), and inside the 3-D fiber optic cable shown schematically in Figure 10.8.

As examples, we present in Figure 10.12 a view of four main LP optical fiber modes: LP_{01} ($m = 0$, $l = 0$) and LP_{11} ($m = 1$, $l = 1$), LP_{20} ($m = 0$, $l = 0$) and LP_{21} ($m = 1$, $l = 1$) [1–3].

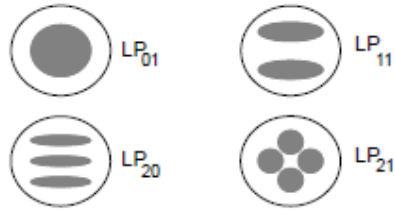


Figure 10.12. The mode LP_{01} presents a single-mode propagation; the mode LP_{11} consists of two modes, the mode LP_{20} consists of three propagating modes, and the mode LP_{21} consists of four propagating modes.

To understand more precisely mode LP_{11}, we present it separately in Figure 10.13 for horizontal and vertical linear polarization, respectively. Thus, on the left side of Figure 10.13, which corresponds to right top panel in Figure 10.12, both the electric field components E_y and E_x and the corresponding components H_x and H_y are directed oppositely, generating the two-wave form of LP_{11} mode.

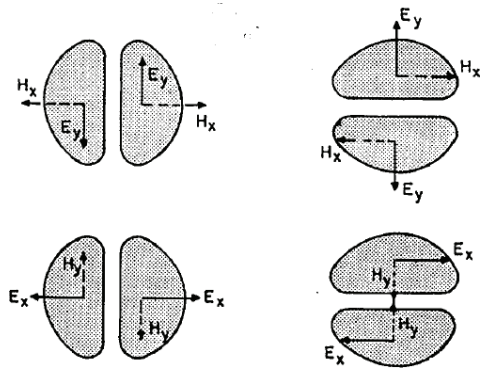


Figure 10.12. The mode LP_{11} for both kinds of polarization: 2 vertical (or *TE*), which changes the modal form of the wave (left side); 2 horizontal (or *HE* or *EH*), which does not change its form.

At the right side of Figure 10.13, the 2 rays of the horizontal polarization are presented. Here, components H_x and H_y are coincidently directed and therefore form a single-mode shape of mode LP_{11}.

We also notice that each mode propagates inside the core according to the value of the corresponding propagation parameter b defined by Eq. (10.28) and the normalized frequency V, introduced above, but now presented in another form via the parameter Δ:

$$V = \sqrt{(ha)^2 + (qa)^2} = k_0 a \sqrt{n_1^2 - n_2^2} = \frac{2\pi a}{\lambda_0} NA = \frac{2\pi a}{\lambda_0} n_1 \sqrt{2\Delta} \qquad (10.35)$$

As an example, the dependence of several from TE, TM, EH and HE polarized modes in the b-V plane on the refractive indexes [via Eqs. (10.28) and (10.35)] are presented in Figure. 10.14.

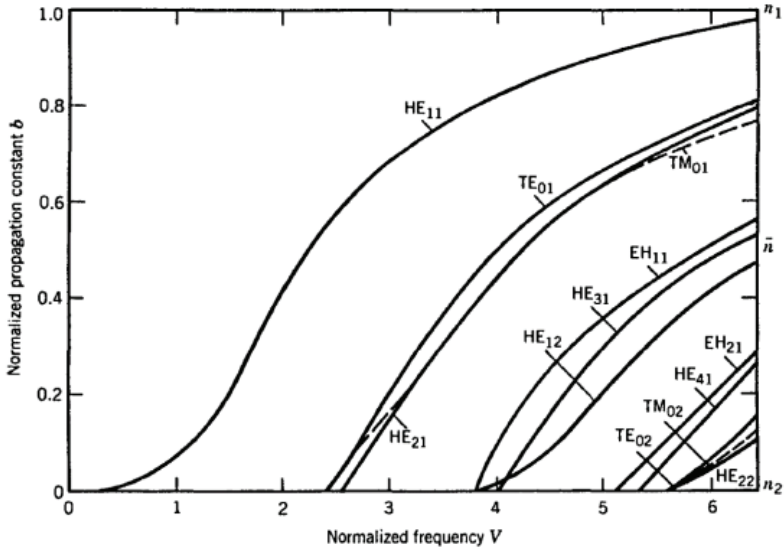


Figure 10.14. Presentation of TE and TM modes in the *b*-*V* plane for various refractive indexes n_1 and n_2.

For cylindrical geometry the *single-mode condition* is [1–3]:

$$\frac{2\pi a}{\lambda}(n_1^2 - n_2^2)^{1/2} < 2.404 \tag{10.36}$$

or more simply: $V < 2.404$. From Eq. (10.35) and constraint (10.36) for given refractive indexes of the core and the cladding, the optimal radius of the core, a, can be found, which allows a single-mode propagation regime inside the fiber (see the corresponding exercises below).

As was shown in [1–6], depending on the shape of the intrinsic refractive index distribution, step-index or graded-index, the corresponding LP-modes can propagate asymmetrically and non-homogeneously. This phenomenon is called the *modal dispersion* [1–6]. Before entering into this important subject, let us introduce the main important operational parameters of fiber optic guiding structures and the corresponding problems of optical propagation via geometrical optics presentation of rays within such structures.

10.5. Dispersion of Signals in Fiber Optic Links

A problem of transmission of pulses via fiber optic structures occurs because of two factors. One is that the source of light is not emitted at a single wavelength but exists over a range of wavelengths called the source spectral width [1–6]. The second factor is that the index of refraction is not the same at all wavelengths. This property, when the light velocity is dependent on wavelength, is called *dispersion*.

As was discussed earlier, in fiber optic cables fading and the corresponding noise of optical signals occurs due to four factors: 1) multimode dispersion phenomena leading to inter-ray interference (IRI); 2) material dispersion; 3) waveguide dispersion; 4) polarization mode dispersion.

Dispersion of these types was discussed in detail by Refs. [4–6] and below, in our description of the subject, we will follow on from some of the discussions presented there.

Before entering into the subject, let us briefly describe the physical meaning of dispersion occurring in fiber optic structures.

As was shown in Chapter 2, an optical wave of angular frequency $\omega_0 = 2\pi f$ (f is the carrier frequency) propagates in free space with parameter β_0 and with phase velocity $v_{ph} = \omega_0 / \beta_0$. As for group velocity, which defines propagation of total optical signal, as a wavelet, energy, its definition is more complicated and can be defined as:

$$v_g = \frac{1}{\beta'} = \frac{1}{d\beta/d\omega}\bigg|_{\omega=\omega_0} = \frac{d\omega}{d\beta}\bigg|_{\omega=\omega_0} \tag{10.37}$$

The corresponding time of signal energy transfer along the z-axis along the fiber can be found as

$$\tau_g = \beta' z = \frac{z}{v_g} \tag{10.38}$$

We can rewrite Eq. (10.37) via the absolute value of the wave vector $v=|\mathbf{v}|$, as

$$v_g = \frac{d\omega}{dk} = \frac{d}{dk}\left(v_{ph}k\right) = k\frac{dv_{ph}}{dk} + v_{ph}\frac{dk}{dk} \tag{10.39}$$

Accounting for

$$k=\frac{2\pi}{\lambda} \longrightarrow \frac{dk}{d\lambda}=-\frac{2\pi}{\lambda^2} \longrightarrow dk=-\frac{2\pi}{\lambda^2}d\lambda$$

we finally get

$$v_g=v_{ph}+\frac{2\pi}{\lambda}\frac{dv_{ph}}{-\frac{2\pi}{\lambda^2}d\lambda}=v_{ph}-\lambda\frac{dv_{ph}}{d\lambda} \tag{10.40}$$

The *multimode* dispersion can be found by knowledge of the second derivative of the parameter β, that is,

$$\beta''=\frac{\partial^2\beta}{\partial\omega^2}=\frac{\partial^2}{\partial\omega^2}\left(\sqrt{k^2-h^2}\right)=\frac{\partial^2}{\partial\omega^2}\left(\sqrt{k_0^2n_1^2-h^2}\right) \tag{10.41}$$

or accounting for definition (10.35) via the normalized frequency V.

In the case of *modal dispersion* caused by multimode propagation inside the optic fiber, a spread of information pulses at the length of optical guiding structure in time can be found as [9]:

$$\Delta\tau_w=l\,n_1\Delta/c\,n_2 \tag{10.42}$$

where l is the length of the fiber optic structure; other parameters were described above.

The *material* dispersion also can be defined through the second derivative of β or by taking into account the relation between the latter and the group velocity,

$$\beta''=\frac{\partial^2\beta}{\partial\omega^2}=\frac{\partial}{\partial\omega}\left(\frac{\partial\beta}{\partial\omega}\right)=\frac{\partial}{\partial\omega}\left(\frac{1}{v_g}\right) \tag{10.43}$$

Finally, we get for *material dispersion* the definition:

$$D = \frac{\partial}{\partial \lambda}\left(\frac{1}{v_g}\right) = -\frac{2\pi c}{\lambda^2}\frac{\partial}{\partial \omega}\left(\frac{1}{v_g}\right) = -\frac{2\pi c}{\lambda^2}\beta'' \left[\frac{1}{\mathrm{m}}\frac{1}{\mathrm{m/s}} = \frac{\mathrm{s}}{\mathrm{m}^2}\right] \tag{10.44}$$

Usually, in practical applications of optical fibers, this parameter is presented in units of *ps /nm·km*.

Waveguide Dispersion depends on the material parameters (refractive indices, radius) of the fiber as well its modal parameters, such as wavelength and time of energy channeling. To analyze this more complicated phenomenon that combines both kinds of dispersion discussed above, let us return to Eq. (10.38) and find the time spread $\Delta\tau_g$ between two modes which lie in the small spectral range $\Delta\omega$ of the total signal, that is,

$$\frac{\Delta\tau_g}{\Delta\omega} = \frac{d\tau_g}{d\omega} \longrightarrow \Delta\tau_g = \frac{d\tau_g}{d\omega}\Delta\omega = \frac{d}{d\omega}\left(\frac{L}{v_g}\right)\Delta\omega = L\beta''\Delta\omega \tag{10.45}$$

Accounting for

$$\frac{\Delta\omega}{\Delta\lambda} = \frac{d\omega}{d\lambda} = \frac{d}{d\lambda}(2\pi f) = \frac{d}{d\lambda}\left(2\pi\frac{c}{\lambda}\right) = -\frac{2\pi c}{\lambda^2}$$

and Eq. (10.42), yields

$$\Delta\tau_g = L\beta''\Delta\omega = DL\Delta\lambda \tag{10.46}$$

Usually, in fiber optic cable engineering, an additional parameter is used, called the index of transmission of energy via cable, defined as:

$$n_g = \frac{c}{v_g} \tag{10.47}$$

We will find an expression for this index via the well-known and already introduced above effective index of refraction, which we will now rewrite in the following manner:

$$n_{eff} = \frac{\beta}{k_0} = \frac{\beta}{\omega/c} = \frac{\beta c}{\omega} \longrightarrow c\beta = \omega n_{eff} \tag{10.48}$$

Accounting for relation

$$\frac{d}{d\omega}(c\beta)=\frac{d}{d\omega}\left(\omega n_{eff}\right)$$

or

$$c\frac{d\beta}{d\omega}=\frac{d\omega}{d\omega}n_{eff}+\omega\frac{dn_{eff}}{d\omega}$$

and substituting in it (10.40) and (10.47), we finally get:

$$n_g=n_{eff}+\omega\frac{dn_{eff}}{d\omega} \tag{10.49}$$

If now we suppose that

$$n_{eff}=n_{ph} \tag{10.50}$$

after straightforward manipulations, we get:

$$n_g=\frac{c}{v_g}=n_{ph}+\omega\frac{dn_1}{d\omega}=n_{ph}+2\pi f\frac{dn_{ph}}{d(2\pi f)} \tag{10.51}$$

Moreover, trivial relations between the radiating frequency and the wavelength of the optical wave, propagating along the fiber optic structure

$$f=\frac{c}{\lambda}\longrightarrow\frac{df}{d\lambda}=-\frac{c}{\lambda^2}\longrightarrow df=-\frac{c}{\lambda^2}d\lambda$$

yields:

$$n_g=n_{ph}+\frac{c}{\lambda}\frac{dn_{ph}}{-\frac{c}{\lambda^2}d\lambda}=n_{ph}-\lambda\frac{dn_{ph}}{d\lambda} \tag{10.52}$$

So, we obtained two equivalent definitions of the index n_g via the frequency (10.51) and via the wavelength (10.52) of the optical wave inside the fiber optic cable.

Generally speaking, all the discussions above allow us to present the total dispersion as a cumulative effect of both *waveguide dispersion* and *material dispersion*,

$$D = D_M + D_W$$

as:

$$D = -\frac{2\pi}{\lambda^2}\frac{d}{d\omega}\left(\frac{c}{v_g}\right) = -\frac{2\pi}{\lambda^2}\frac{d}{d\omega}\left(n_{eff} + \omega\frac{dn_{eff}}{d\omega}\right) = -\frac{2\pi}{\lambda^2}\left(\frac{dn_{eff}}{d\omega} + \frac{d\omega}{d\omega}\frac{dn_{eff}}{d\omega} + \omega\frac{d^2n_{eff}}{d\omega^2}\right)$$
$$= -\frac{2\pi}{\lambda^2}\left(2\frac{dn_{eff}}{d\omega} + \omega\frac{d^2n_{eff}}{d\omega^2}\right) \tag{10.53}$$

We notice here that if we introduce, according to (10.28), the parameter b, then, instead of (10.48), we will get:

$$b = \frac{n_{eff} - n_2}{n_1 - n_2} \quad \longrightarrow \quad n_{eff} = n_2 + b(n_1 - n_2) \approx n_2(1 + b\Delta) \tag{10.54}$$

Finally, we can rewrite the expression (10.44) for the material expression via the index n_g inside the cladding layer of the optic fiber:

$$D_M = -\frac{2\pi}{\lambda^2}\frac{\partial n_{2g}}{\partial \omega} = \frac{1}{c}\frac{\partial n_{2g}}{\partial \lambda} \tag{10.55}$$

Qualitative analysis of the process of optical wave propagation inside the fiber optic for the case when the refractive index can be presented by a sum of its value in free space and a sum of effects from $i = 1, 2, \ldots, M$, harmonics propagating inside it with time dispersion (see definitions below), that is,

$$n^2(\omega) = 1 + \sum_{i=1}^{M}\frac{B_i\omega_i^2}{\omega_i^2 - \omega^2} \tag{10.56}$$

Computations of the parameters n and n_g are shown in Figure 10.15.

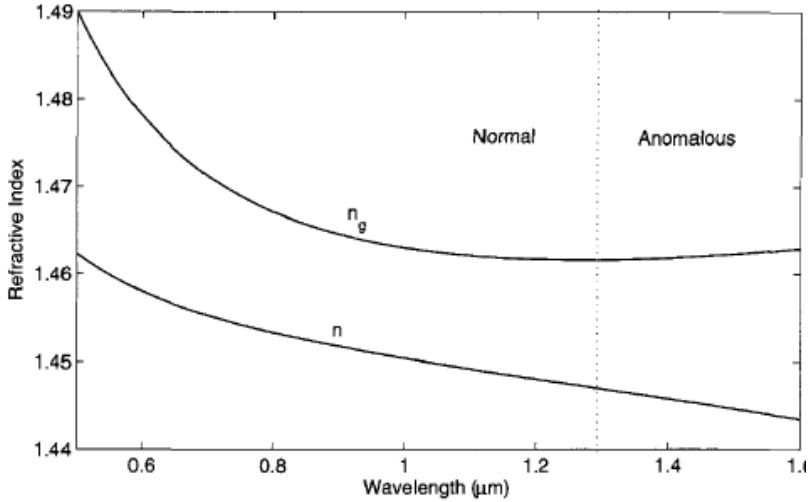


Figure 10.15. Refractive indexes vs. the wavelength (in μm).

The vertical dotted line indicates the case occurring in fiber optic based on SiO_2 semiconducting material (see Chapter 3) for $\lambda = 1.276$ μm. This line shows the boundary of the normal cable, where n and n_g have the same tendency to decrease with an increase of λ. This wavelength, called the zero-dispersion (ZD) wavelength, can be defined for $\lambda = 1.276$ μm by the following constraint:

$$dn_g / d\lambda = 0 = D_M \tag{10.57}$$

As for various wavelengths, we now get:

$$D_M(\lambda) \approx 122\left(1 - \frac{\lambda_{ZD}}{\lambda}\right) \tag{10.58}$$

General assumptions made above allow us to present the *waveguide dispersion* via parameters b from (10.28) and V from (10.35), as:

$$D_w = -\frac{2\pi\Delta}{\lambda^2}\left[\frac{n_{2g}^2 V}{n_2\omega}\frac{d^2(Vb)}{dV^2} + \frac{\partial n_{2g}}{\partial\omega}\frac{d(Vb)}{dV}\right] \tag{10.59}$$

Figure 10.16 presents the total dispersion $D = D_m + D_w$, their separate dependence on the wavelength according to Ref. [6], where $\lambda_{ZD} = 1.35$ μm indicates the case of $D = 0$, that is, $D_m = -D_w$ [6].

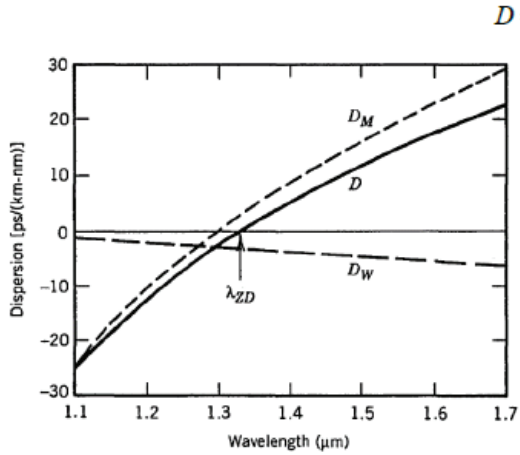


Figure 10.16. The total (continuous curve), material (dashed curve) and waveguide (dashed bold curve) dispersions [in $ps / km \cdot nm$] vs. wavelength; $\lambda_{ZD} = 1.35$μm (indicated by arrow) corresponds to $D = 0$, and $D_m = -D_w$ according to [6].

The summands of Eq. (10.59) and the parameter b according to (10.54) for fiber optic cable with diameter $d = 2a$ and refraction indexes of the core n_1 and the cladding n_2, are shown in Figure 10.17 versus the normalized frequency V, according to computations made in Ref. [6].

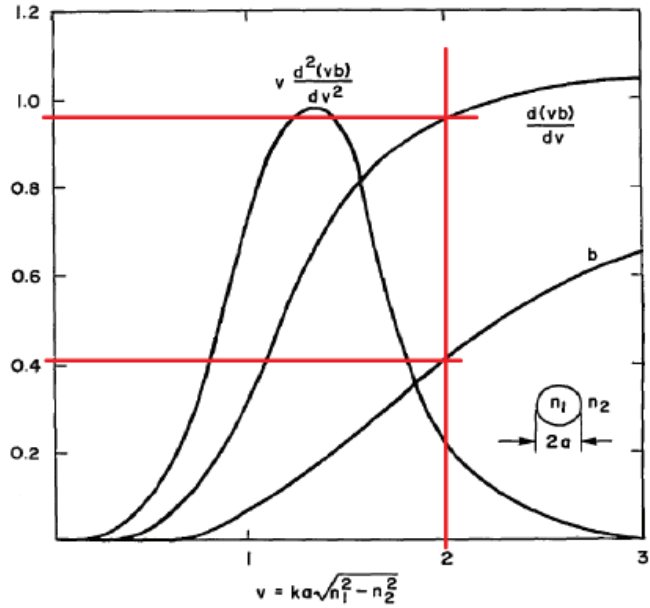


Figure 10.17. Summands and parameter *b* in Eq. (10.59) vs. dimensionless parameter *V* computed according to [6]; the left-bottom side is the description of geometry and refraction indexes of the fiber.

Using now another approach presented in [6], we can rewrite Eq. (10.59) as:

$$D_W = -\frac{2\pi\Delta}{\lambda^2}\left[\frac{n_{2g}^2 V}{n_2\omega}\frac{d^2(Vb)}{dV^2} + \frac{dn_{2g}}{d\omega}\frac{d(Vb)}{dV}\right] = -\underbrace{\frac{2\pi}{\lambda}}_{k_0}\frac{\Delta}{\lambda}\frac{n_{2g}^2}{n_2\omega}\frac{Vd^2(Vb)}{dV^2} - \frac{d(Vb)}{dV}\underbrace{\frac{2\pi}{\lambda^2}\frac{dn_{2g}}{d\omega}}_{D_M}\Delta \tag{10.60}$$

This equation, as well as the corresponding curves presented in Figure 10.17 for parameters b and V (crossing straight lines in Figure 10.17), will be used in further exercises presented below.

Polarization mode dispersion (PMD) may occur in the optical guiding structures for different forms of optical wave polarization, vertical and

horizontal. Its characteristic, called a *pulse spread* due to changes of polarization, is defined as [6]:

$$\sigma_p = D_p \cdot L^{1/2} \tag{10.61}$$

where D_p is the polarization mode dispersion (PMD) parameter, measured in picoseconds per square root of kilometer [ps/(km)$^{1/2}$]. In other words, light rays with different polarizations propagate at different speeds. For the usually used graded-index and step-index fibers, D_p is less than 0.5 ps/km$^{1/2}$, but sometimes can exceed ~10 ps/(km)$^{1/2}$. Critical limitations exist to transmit information signals of high data rates [9, 10].

10.6. Attenuation and Scattering Inside Fiber Optic Structures

Attenuation losses inside fiber optic structures are usually determined by factor α, called the *attenuation coefficient*. This coefficient was fully described in Chapter 2. We now notice that, typically, attenuation inside an optical fiber is determined in dB/km, but not per Np/m, that is,

$$dB/km = -8.685 \cdot \alpha \tag{10.43}$$

where the units of the attenuation coefficient are in km^{-1}. As an example, the cladding or core of an optical cable fabricated from silica absorbs optical waves over a wide range of wavelengths – from ultraviolet (UV), due to electronic resonances, to infra-red (IR), due to vibrational resonances [4, 5].

Scattering phenomena can be characterized by several types of scattering which occur inside fiber optic links, most of which are Rayleigh and Raman (Stocks and anti-Stocks) [1–6]. Generally, the *Rayleigh scattering* approach, which is ~λ^{-4} [6] and takes place for roughness and defects of the inner (core) fiber surface, is valid when the dimensions of which are much less than the wavelength of light ($l << \lambda$). Impurities or defects that play major roles are defects with l = 0.6–1.6 μm and which satisfy the constraint $l < \lambda$ or $l \sim \lambda$, then the *Mie scattering* phenomenon is valid (see Refs. [1–6]).

The total loss, obtained experimentally, as well as the Rayleigh scattering effects, with limits on UV and IR absorption, and the effects of different waveguide imperfections, are summarized in Figure 10.18.

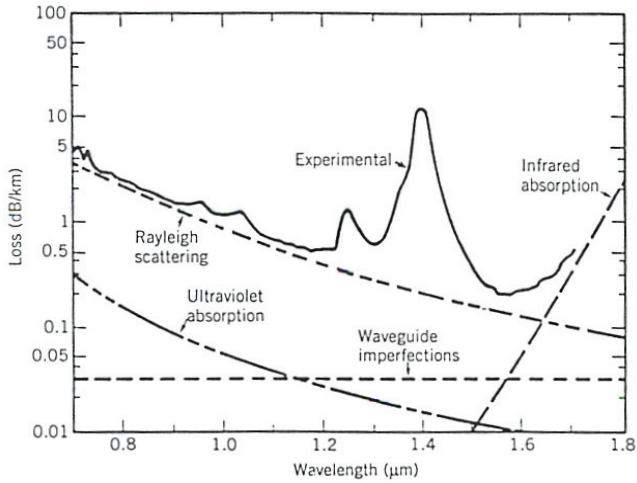


Figure 10.18. The total experimental attenuation for silica fiber optic structure, Rayleigh scattering, UV and IR absorption curves, and Mie scattering for waveguide imperfections vs. wavelength ranging from 0.6 to 1.8 μm.

As follows from Figure 10.18, the minimum loss for IR observed at $\lambda = 1.55$ μm, which coincides with the experimental curve of total loss around 0.3 dB/km. As for Mie scattering, it shows independence of wavelength over the total waveband. The Rayleigh scattering law is fully approximated by the experimentally obtained data for silica in the near-infrared band from 0.8 to 1.2 μm, and a decrease of loss is observed from 5 dB/km to 1 dB/km.

Summary

The information mentioned above allows us to outline the following:

1. Depending on the wavelength of the optical signal propagating inside the fiber optic channel, the material dispersion index decreases exponentially with an increase of the signal wavelength.
2. At the same time, with an increase of the wavelength of the optical signal passing through the fiber optic channel, the delay spread (i.e., widening) of the resulting optical signal inside the cable increases linearly, caused by the modal dispersion.
3. The material time dispersion parameter along the fiber optic cable

increases linearly with an increase of the difference between the refraction indexes of the inner and outer parts of the fiber, called fractional refractive indexes difference (FRID), which has a tendency to decrease exponentially the optical signals passing such a cable.

4. The multimode time dispersion depends significantly on the difference between the refraction indexes of the inner (core) and outer (cladding) parts of the fiber, and with an increase of FRID, it increases linearly.
5. For all types of optical digital signals, multimode dispersion depends on the difference between the refraction indexes of the inner and outer parts of the fiber and on the increase of the length of the fiber.

Exercises

Exercise 1.

Let us consider that $n_1 = 1.45$ and $\Delta = 0.02$ (2%).

Find: *N.A.* and $2 \cdot \theta_a$.

Solution

1) Accounting for (10.1) and (10.2), we get $N.A. = n_1 \cdot (\Delta)^{1/2} = \sin^{-1} \theta_a$

Then: $N.A. = n_1 \cdot (\Delta)^{1/2} = 1.45 \cdot \sqrt{2 \cdot 0.02} = 0.29$

2) $N.A. = \sin^{-1} \theta_a = \sin^{-1} 0.29 = 15.66^0$

Then: $2 \cdot \theta_a = 2 \cdot 15.66^0 = 31.33^0$.

Exercise 2.

Given: Fiber optic step-index multimode cable with the following parameters: $\Delta = 0.01$, $n_1 = 1.455$, $d = 20\mu m$, and $\lambda_0 = 780$ nm.

Find: Normalized frequency *V*.

Solution

1) Taking into account Eq. (10.2) we get:

$$n_2 = n_1 \cdot (1-2\Delta)^2 = 1.455 \cdot 0.3 = 0.4365$$

2) Taking into account Eq. (10.1) we get:

$$N.A. = [(1.455)^2 + (0.4365)^2]^{1/2} = 1.4$$

Taking Eq. (10.29), we get: $V = (\pi d/\lambda_0) \cdot NA$ and taking into account Eq. (10.1), we get:

$$V = [3.14 \cdot 20 / 0.78] \cdot 1.4 = 5.6$$

Exercise 3.

Given: Radius of the core $d = 20$ μm, $\lambda_0 = 1300$ nm, $n_1 = 1.467$, $\Delta = 1.36\%$.

Find: How many mode solutions for *N* from Eq. (10.12) can be obtained (see also Fig. 10.7).

Solution

1) Accounting for Eq. (10.2), we get

$$n_2 = n_1\sqrt{1-2\Delta} = 1.467\sqrt{1-2\frac{1.36\%}{100\%}} = 1.447$$

2) Accounting for Eq. (10.2), we get

$$NA = \sqrt{n_1^2 - n_2^2} = \sqrt{1.467^2 - 1.447^2} = 0.242$$

3) Accounting for Eq. (10.29) we get

$$V = \frac{2\pi}{\lambda_0}\frac{d}{2}\sqrt{n_1^2 - n_2^2} = \frac{\pi d}{\lambda_0}NA = \frac{\pi \cdot 20 \cdot 10^{-6}}{1.3 \cdot 10^{-6}}0.242 = 11.69$$

4) Accounting for Eq. (10.17) we get

$$N = \left\lfloor \frac{2V}{\pi} + 1 \right\rfloor = \left\lfloor \frac{2 \cdot 11.668}{\pi} + 1 \right\rfloor = \lfloor 8.446 \rfloor = 8$$

It also can be found by use of dependence of N with NA according to Eq. (10.17):

$$N = \left\lfloor \frac{2d}{\lambda} NA + 1 \right\rfloor = \left\lfloor \frac{2 \cdot 20 \cdot 10^{-6}}{1.3 \cdot 10^{-6}} \cdot 0.242 + 1 \right\rfloor = \lfloor 8.446 + 1 \rfloor = 8$$

Conclusion: In such a fiber optic structure, only 7 modes can propagate according to Figure 10.7.

Exercise 4.

Given: Radius of the core d = 20 μm, λ_0 = 1300 nm, n_1 = 1.467, Δ=1.36%.

Find:

1) Angle of full transfer of main mode along the fiber (according to (10.1)) $\theta_{max} = \alpha_{\max}$
2) Critical angle θ_c for the main mode and for mode 7 (according to Exercise 3).
3) Incident angle θ_0 inside the fiber (see Figure 10.3).
4) The range of n_{eff} between the main and the 7-th modes.
5) Wavelength ranged between 6-th and 4-th modes, and similarly – the normalized frequency V ranged between these modes.

Solution

1) According to (10.1)

$$\alpha_{\max} = \sin^{-1}\frac{NA}{n_0} = \sin^{-1} 0.242 = 14°$$

2) According to Exercise 3:

$$NA = \sqrt{n_1^2 - n_2^2} = \sqrt{1.467^2 - 1.447^2} = 0.242$$

and following (10.1), we get

$$\theta_c = \sin^{-1}\frac{NA}{n_1} = 9.495°$$

3) The incident angle θ_0 can be found as:

$$\sin\theta_0 = 0.9\frac{\lambda}{2d}.$$

from which follows:

$$\theta_0 = 1.676°$$

Now, accounting from Exercise 3 for $N\max = 8.446$. As follows from Figure 10.7, for all 7 bold points N=8, finally, we get:

$$\frac{\theta_7}{\theta_c} = \frac{\left\lfloor \frac{2V}{\pi} + 1 \right\rfloor}{\frac{2V}{\pi} + 1}$$

from which

$$\theta_7 = \frac{8}{8.446}\theta_c = 8.994°$$

According to Snell's law:

$$n_1 \sin\theta_7 = n_0 \sin\theta_{7,\text{air}}$$

we get:

$$\theta_{7,\text{air}} = \sin^{-1}\left(1.467 \sin\left(8.994°\right)\right) = 13.26°$$

4) As well-known from above:

$$n_{\text{eff},m} = \frac{\beta_m}{k}$$

As for β_0 and β_7, they can be found from step 3 of this Exercise:

$$\beta_0 = n_1 k \cos\theta_0 = n_1 \frac{2\pi}{\lambda}\cos\left(1.676°\right) = 7.078 \cdot 10^6 \frac{\text{rad}}{\text{m}} = 7.078 \frac{\text{rad}}{\mu\text{m}}$$

$$\beta_7 = n_1 \frac{2\pi}{\lambda}\cos\left(8.994\right) = 7.53 \cdot 10^6 \frac{\text{rad}}{\text{m}} = 7.53 \frac{\text{rad}}{\mu\text{m}}$$

If so, finally we get that

$$\frac{\beta_7}{2\pi/\lambda} < n_{\text{eff}} < \frac{\beta_0}{2\pi/\lambda}$$

or after straightforward computation we get:

$$1.449 < n_{\text{eff}} < 1.466$$

5) Accounting now for the following relation:

$$\lambda = \frac{\pi d}{V} NA$$

we get for wavelengths of waveguide modes via

$$\lambda_6 < \lambda < \lambda_4$$

or

$$\frac{\pi d}{V_6} NA < \lambda < \frac{\pi d}{V_4} NA$$

That

$$\frac{\pi \cdot 20 \cdot 10^{-6}}{16.53} 0.242 < \lambda < \frac{\pi \cdot 20 \cdot 10^{-6}}{4\pi} 0.242$$

Finally, the wavelength is varied from 4th to 6th mode at the range of

$$1.94\mu m < \lambda < 3.2\mu m$$

Similarly, we can find the normalized frequencies ranged between these modes. Thus, accounting for well-known relations (10.16), we get:

for N=4 $4 = 2V_4 / \pi + 1$, from which $V_4 = 4.71$

for N=6 $6 = 2V_6 / \pi + 1$, from which $V_6 = 7.85$

Exercise 5.

Given: 2-D guiding structure with width d = 20 μm, (Fig. 10.3), where light propagates along it with the wavelength λ_0 = 1550 nm.

Find: 1) NA and V;
2) Angle of incidence for modes of m = 0, 2 and m = 6;
3) Wave numbers for modes of m = 0, 2 and m = 6;
4) For single-mode propagation conditions find the corresponding width of the slab and the wavelength for this mode.

Solution

1) Accounting for (10.1), we get:

$$NA = \sqrt{1.465^2 - 1.445^2} = 0.241$$

In the same manner will be calculated the normalized frequency V:

$$V = \frac{\pi d}{\lambda_0} NA = \frac{\pi d}{\lambda_0} \sqrt{n_1^2 - n_2^2} = \frac{\pi \cdot 20 \cdot 10^{-6}}{1.55 \cdot 10^{-6}} 0.241 = 9.77$$

2) Following the wave mode propagation condition, following geometry presented in Fig. 10.6, yields:

$$\underbrace{d\frac{2\pi n_1}{\lambda_0}\sin\theta_m}_{k_x} = \pi m$$

Then

for $m = 0$

$$\theta_0 = \sin^{-1}\left(\frac{m\pi\lambda_0}{2\pi n_1 d}\right)\Bigg|_{m=0} = 0°$$

for $m = 2$

$$\theta_2 = \sin^{-1}\left(\frac{m\pi\lambda_0}{2\pi n_1 d}\right)\Bigg|_{m=2} = \sin^{-1}\left(\frac{1.55\cdot 10^{-6}}{1.465\cdot 20\cdot 10^{-6}}\right) = 3.032°$$

for $m = 6$

$$\theta_6 = \sin^{-1}\left(\frac{m\pi\lambda_0}{2\pi n_1 d}\right)\Bigg|_{m=6} = \sin^{-1}\left(\frac{6\pi\cdot 1.55\cdot 10^{-6}}{2\pi\cdot 1.465\cdot 20\cdot 10^{-6}}\right) = 9.13°$$

3) The corresponding wave numbers can be found in the same manner:

for $m = 0$

$$\beta_0 = k\cos\theta_0 = \frac{2\pi n_1}{\lambda_0} = 5.94\frac{\text{rad}}{\mu\text{m}}$$

for $m = 2$

$$\beta_2 = k\cos\theta_2 = \frac{2\pi n_1}{\lambda_0}\cos\theta_2 = 5.927\frac{\text{rad}}{\mu\text{m}}$$

for $m = 6$

$$\beta_6 = k\cos\theta_6 = \frac{2\pi n_1}{\lambda_0}\cos\theta_6 = 5.86\frac{\text{rad}}{\mu\text{m}}$$

4) For single-mode propagation inside the slab, as following from (10.30) we get:

$$N = \left\lfloor \frac{2V}{\pi} + 1 \right\rfloor < 2 \quad \longrightarrow \quad V \le \frac{\pi}{2}$$

or using the definition of V and the above conditions, we can find the conditions for the width d of the slab that guides only the main mode (with $m = 0$)

$$V = \frac{\pi d}{\lambda_0} NA \le \frac{\pi}{2} \quad \longrightarrow \quad d \le \frac{\lambda_0}{2NA} = \frac{1.55 \cdot 10^{-6}}{2 \cdot 0.241} = 3.21\mu m$$

with the corresponding wavelength

$$\lambda_0 = \frac{\pi d}{V} NA = \frac{\pi dNA}{9\pi/2} = \frac{2 \cdot 20 \cdot 10^{-6} \cdot 0.241}{9} = 1.07\mu m$$

Exercise 6.

Given: 2-D guiding optical structure (see Figure 10.19), $n_1 = 1.48$ and $n_2 = 1.46$.

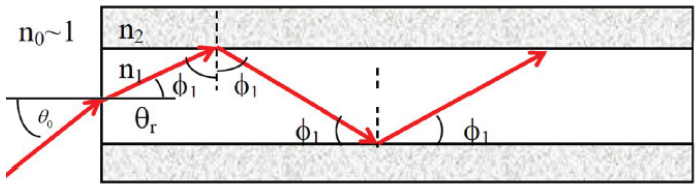


Figure 10.19. Mode propagation inside the 2-D slab.

Find: 1) *NA* and Δ.
2) Maximum incident angle θ_{0max} for $n_0 \sim 1.0$.

Solution

1) Accounting for (10.1), we get:

$$NA = \sqrt{1.48^2 - 1.46^2} = 0.242$$

From (10.3), the approximate formula for Δ can be obtained

$$\Delta = \frac{(n_1 - n_2)(n_1 + n_2)}{2n_1^2} \approx \frac{(n_1 - n_2)2n_1}{2n_1^2} = \frac{n_1 - n_2}{n_1}$$

Then, we finally get:

$$\Delta = \frac{1.48 - 1.46}{1.48} = 0.0135$$

2) Accounting for Snell's law (see geometry in Fig. 10.19) yields:

$$n_1 \sin\phi_1 = n_2 \sin\phi_2 \longrightarrow \sin\phi_1 = \frac{n_2}{n_1}\sin\phi_2$$

At the same time, as follows from geometry presented in Figure10.7

$$\theta_r + \phi_1 = \frac{\pi}{2} \longrightarrow \sin\theta_r = \cos\phi_1$$

Finally, we get:

$$\sin\theta_r = \cos\phi_1 = \sqrt{1 - \sin^2\phi_1} = \sqrt{1 - \frac{n_2^2}{n_1^2}\sin^2\phi_2} = \frac{1}{n_1}\sqrt{n_1^2 - n_2^2\sin^2\phi_2}$$

Moreover, according to Snell's law

$$n_0 \sin\theta_i = n_1 \sin\theta_r$$

If now inside core is some material, let us say, with n= 1.33, then we get:

$$\phi_1 > \phi_c \longrightarrow \phi_2 = \frac{\pi}{2}$$

and the corresponding $\theta_{0,\max}$ for such conditions will equal:

$$\theta_{0,\max} = \sin^{-1}\left(\frac{NA}{n_0}\right) = \sin^{-1}(0.242) = 14°$$

Exercise 7.

Given: Multimode step-index fiber with parameters NA = 0.3, V = 75, n_1 = 1.458, λ_0 = 820 nm.

Find:
1) n_2 .
2) Radius of core d.

Solution

1) from (10.1) we can find n_2 as

$$n_2 = (n_1^{\,2} - NA^{\,2})^{1/2} = (1.458^2 - 0.3^2)^{\,1/2} = 1.427$$

2) from (10.16) radius d equals

$$d = V \cdot \lambda_0 / 2\pi \cdot NA = 75 \cdot 820 \cdot 10^{-9} / 6.28 \cdot 0.3 = 32.63 \text{ μm}$$

Exercise 8.

Given: Multimode step-index fiber with parameters d = 25μm, n_1 = 1.48, λ_0 = 820 nm.

Find: 1) Normalized frequency V for Δ = 0.01;
2) The fiber mode parameter M for Δ = 0.003.

Solution

1) According to (10.16), we get

$$V = (2\Delta)^{1/2}\, 2\pi \cdot d \cdot n_1 / \lambda_0 = (0.02)^{1/2} \cdot 6.28 \cdot 25 \cdot 10^{-6} \cdot 1.48 / 820 \cdot 10^{-9} = 24.9$$

2) the fiber mode parameter M can also be found via (10.16) as

$$M_1 = V^2/2 = 0.5\cdot(2\pi\cdot d\cdot n_1\cdot(2\Delta)^{1/2}/\lambda_0)^2 =$$
$$= 4\cdot 0.003\cdot(3.14\cdot 25\cdot 10^{-6}\cdot 1.48/820\cdot 10^{-9})^2 = 241$$

Exercise 9.

Given: Multimode step-index fiber with the parameters M = 100 and *NA* = 0.2; λ_0 = 850 nm.

Find: 1) Diameter of core *D*.
2) Number M for λ_1 = 1320 nm and λ_2 = 1550 nm.

Solution

1) It is known that

$$M = V^2/2, \text{ so } V = (2\cdot 1000)^{1/2} = (2000)^{1/2}$$

At the same time, according to (10.16)

$$d = V\cdot\lambda_0/2\pi\cdot NA = (2000)^{1/2}\cdot 850\cdot 10^{-9}/6.28\cdot 0.2 = 30.25\mu m$$

Then the diameter of the core equals

$$D = 2d = 60.5\ \mu m$$

2) For λ_1 = 1320 nm

$$M_1 = V^2/2 = 2\cdot(\pi\cdot d\cdot NA/\lambda_1)^2 = 2\cdot(6.28\cdot 30.25\cdot 10^{-6}\cdot 0.2/1320\cdot 10^{-9})^2 = 414$$

For λ_2 = 1550 nm

$$M_2 = V^2/2 = 2\cdot(\pi\cdot d\cdot NA/\lambda_2)^2 = 2\cdot(6.28\cdot 30.25\cdot 10^{-6}\cdot 0.2/1550\cdot 10^{-9})^2 = 300$$

Exercise 10.

Given: Step-index fiber of the length L, $n_1 = 1.5$.

Find:
1) Maximum bit rate for a) $\Delta = 1/3$, and b) $\Delta = 2 \cdot 10^{-3}$.
2) Time dispersion along the length of the fiber, ΔT.

Solution

1) If, due to time spread, overlapping between bits occurs, called *inter-symbol interference* (ISI), and its period T_B exceeds ΔT, accounting that the bit rate $B \sim T_B^{-1}$, we finally get the following constraint:

$$\Delta T \cdot B < 1$$

Accounting for relations between the ΔT, the refraction indexes, parameters of fiber, L and Δ, we can rewrite the above constrain as following

$$B < \frac{n_2 c}{L n_1^2 \Delta}$$

a) For $\Delta = 1/3$ and $n_1 = 1.5$, we get

$$n_2 = n_1(1-\Delta) = 1.5(1-1/3) = 1$$

and

$$B_{max} = \frac{n_2 c}{L n_1^2 \Delta} = \frac{3 \cdot 10^8}{1 \cdot 10^3 \cdot 1.5 \cdot \frac{1}{3}} = 0.4\,\text{Mbits/s}$$

b) For $\Delta = 2 \cdot 10^{-3}$, and $n_1 = 1.5$, we get

$$n_2 = 1.497$$

$$B_{max} = 100\,\text{Mbits/s}$$

2) According to geometry presented in Figure 10.20,

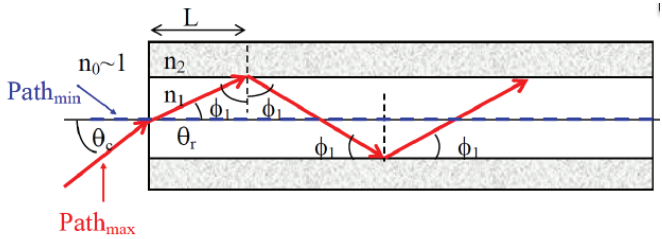


Figure 10.20. Geometrical presentation of the minimum path (straight without reflections – dashed blue line in the middle of the fiber) and of the maximum path (after reflections inside the core) during mode propagation inside the 2-D slab.

we find the maximum path of the optical mode in the fiber:

$$\text{Path}_{\max} = \frac{L}{\sin\phi_1} = \frac{L}{\sin\phi_c} = \frac{L}{\sin\left(\sin^{-1}\frac{n_2}{n_1}\right)} = \frac{n_1}{n_2}L$$

where the critical angle of the total inner reflection inside the core corresponds to (see discussions in Paragraph 10.2)

$$\phi_c = \sin^{-1}\frac{n_2}{n_1}$$

Finally, we get the total time dispersion along the cable length L:

$$\Delta T = \frac{\text{Path}_{\max} - \text{Path}_{\min}}{v} = \frac{\frac{n_1}{n_2}L - L}{\frac{c}{n_1}} = \frac{n_1^2 L - n_1 n_2 L}{n_2 c} = \frac{n_1^2 L}{c n_2}\left(\frac{n_1 - n_2}{n_1}\right) = \frac{n_1^2 L}{c n_2}\Delta$$

or for

$$\Delta = 1/3/\ \Delta T / L = 2.5\ (\mu s \cdot km^{-1})$$

and for

$$\Delta = 2 \cdot 10^{-3}/\ \Delta T / L = 7\ (\mu s \cdot km^{-1})$$

Conclusion: Less difference between n_1 and n_2, and smaller parameter Δ, means the weaker the time spread along the cable takes place.

Exercise 11.

Given: The Step-index fiber with the following parameter

$$2a = 6.2\mu\text{m}, \quad n_1 = 1.451, \quad n_2 = 1.442, \quad n_{2g} = 1.457, \quad \lambda_{ZD} = 1.276\mu\text{m}$$

Find: The total dispersion coefficient for $\lambda_{ZD} = 1.55\mu\text{m}$

Solution

For λ = 1.276 μm according to (10.58) we get:

$$D_M = 122\left(1 - \frac{\lambda_{ZD}}{\lambda}\right) = 122\left(1 - \frac{1.276}{1.55}\right) = 21\frac{\text{ps}}{\text{nm}\cdot\text{km}}$$

and

$$V = k_0 a\sqrt{n_1^2 - n_2^2} \approx 2$$

The summands of Eq. (10.59) shown in Figure 10.21 (we present it again for the readers' convenience) can be computed based on the above parameters.

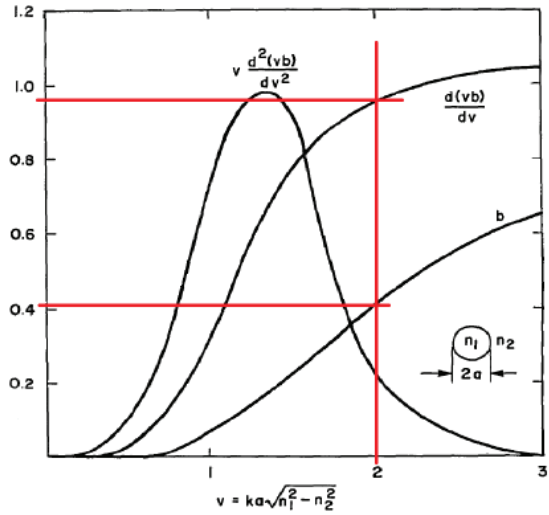


Figure 10.21. Summands and parameter b in Eq. (10.59) vs. dimensionless parameter V computed according to [6]; the left-bottom side is the description of geometry and refraction indexes of the fiber.

Finally, from Figure 10.21, we get for $V = 2$ (see crossing straight lines):

$$V\frac{d^2(Vb)}{dV^2}\bigg|_{V=2} = 0.22$$

$$\frac{d(Vb)}{dV}\bigg|_{V=2} = 0.975$$

Returning now to Eq. (10.60) and rearranging it, we get:

$$D_w = -\frac{2\pi\Delta}{\lambda^2}\left[\frac{n_{2g}^2 V}{n_2\omega}\frac{d^2(Vb)}{dV^2} + \frac{dn_{2g}}{d\omega}\frac{d(Vb)}{dV}\right]$$

$$= -\underbrace{\frac{2\pi}{\lambda}}_{k_0}\frac{\Delta}{\lambda}\frac{n_{2g}^2}{n_2\omega}\frac{Vd^2(Vb)}{dV^2} - \frac{d(Vb)}{dV}\underbrace{\frac{2\pi}{\lambda^2}\frac{dn_{2g}}{d\omega}}_{D_M}\Delta$$

Accounting for

$$\Delta = \frac{n_1 - n_2}{n_1} = \frac{0.009}{1.451} = 6.2 \cdot 10^{-3}$$

we get for waveguide dispersion coefficient:

$$D_W = -\frac{6.2 \cdot 10^{-3}}{3\lambda_{ZD} = 1.55\mu m} \frac{(1.457)^2}{0^{-6}m} \frac{}{1.442} \cdot 0.22 + 6.2 \cdot 10^{-3} \cdot 21 \frac{ps}{nm \cdot km} \cdot 0.975$$

$$= -4.3183 \cdot 10^{-6} \frac{s}{m^2} + 0.1269 \frac{ps}{nm \cdot km}$$

$$= -4.3183 \frac{ps}{nm \cdot km} + 0.1269 \frac{ps}{nm \cdot km}$$

$$= 4.1914 \frac{ps}{nm \cdot km}$$

Accounting for the obtained above material dispersion coefficients, we finally get:

$$\lambda_{ZD} = 1.55\mu m$$

$$D = D_M + D_W = 21 - 4.1914 = 16.8086 \frac{ps}{nm \cdot km}$$

Conclusion: The material dispersion, occurring in the fiber optic structure caused by impurities and defects (called irregularities or roughness structures) is a more essential factor with respect to the waveguide dispersion caused by multimode propagation inside the core of the optical cable.

Bibliography

[1] Adams, M. J. 1981. *An Introduction to Optical Waveguides*. New York: Wiley.
[2] Elliott, R. S. 1993. *An Introduction to Guided Waves and Microwave Circuits*. New Jersey: Prentice-Hall.
[3] Palais, J. C. 1998. *Fiber Optic Communications*, 4th Ed. New Jersey: Prentice-Hall.
[4] Dakin, J., and B. Culshaw. 1988. *Optical Fiber Sensors: Principles and Components*. Boston-London: Artech House.
[5] Palais, J. C. 2006. "Optical Communications." In Handbook: *Engineering Electromagnetics Applications*, edited by R. Bansal. New York: Taylor and Frances.
[6] Yariv, A., and P. Yeh. 2006. *Photonics*, 6th Ed. Oxford: Oxford University Press.
[7] Korn, G, and T. Korn. 1961. *Mathematical Handbook for Scientists and Engineers*. New York: McGraw-Hill.
[8] Abramowitz, M, and I. A. Stegun. 1965. *Handbook of Mathematical Functions*. New York: Dover Publications.
[9] Blaunstein, N., S. Engelberg, E. Krouk, and M. Sergeev. 2019. *Fiber Optic and Atmospheric Optical Communication*. Hoboken, NJ: Wiley.
[10] Proakis, J. G. 2001. *Digital Communication*, 4th Ed. New York: McGraw-Hill, 2001.

INDEX