Photonics

Editor: Alexei Tolstik

Alexei Tolstik

Igor Agishev

Dmitry Gorbach

Viktor Myshkovets

Aliaksandr Maksimenka

Georgy Baevich

Alena Melnikova

Alexander Lyalikov

Alexander Fedotov

Joan Peuteman

Mikhail Tivanov

Natalia Strekal

Gennady Vasilyuk

Igor Semchenko

Sergei Khakhomov

Riga, 2019

This document has been prepared by the financial support of European Union. The authors from Riga Technical University and Belarusian State University are responsible for the content of this document. This publication reflects the views only of the authors, and it cannot be regarded as the European Union's official position.

The book is developed in a frame of the project “ERASMUS+ Capacity-building in the Field of Higher Education 2015 Call for Proposals EAC/A04/2014 561525-EPP-1-2015-1-LV-EPPKA2-CBHE-JP - ERASMUS+ CBHE.

The textbook is devised for students of applied physic and electrical engineering specialties. The textbook can be useful for students and professionals focusing on photonics issues. The book gives overview of current laser physics and nonlinear optics, coherent optics and holography, optical waveguides and optoelectronics, basics of nanophotonics as well as possible areas of their application.

Key Action: KA2 - Cooperation for innovation and the exchange of good practices

Action: Capacity Building in Higher Education

Action Type: Joint Projects

Deliverable: 2.3. Five electronic courses e-Books by the declared directions

Co-funded by the
Erasmus+ Programme
of the European Union

Project Scientific Managers: Leonids Ribickis, Nadezhda Kunicina

Project Coordinator: Anatolijs Zabasta

Editor: Alexey Tolstik

Institution: Riga Technical University

ISBN - 978-9934-22-144-6 (pdf)

Contributors

Alexei Tolstik, Head of the Department of Laser Physics and Spectroscopy, professor, Doctor of Physical and Mathematical Sciences, Physics Faculty, Belarusian State University, 220030, Minsk, 4, Nezavisimosti ave., Republic of Belarus, tel. +375 17 2095441, tolstik@bsu.by

Igor Agishev, Head of educational laboratory at the Department of Laser Physics and Spectroscopy, Physics Faculty, Belarusian State University, 220030, Minsk, 4, Nezavisimosti ave., Republic of Belarus, tel. +375 17 2095118, agishev@bsu.by

Dmitry Gorbach, Senior Lecturer in the Department of of Laser Physics and Spectroscopy, Candidate of Physical and Mathematical Sciences, Physics Faculty, Belarusian State University, 220030, Minsk, 4, Nezavisimosti ave., Republic of Belarus, tel. +375 17 2095118, tolstik@bsu.by

Victor Myshkovets, Head of the Department of radio physics and electronics, Candidate of Physical and Mathematical Sciences, associate professor, Francisk Skorina Gomel State University, 102, Sovetskaya Str., Gomel 246019, Belarus, tel. +375 232 578854, myshkovets@gsu.by

Aliaksandr Maksimenka, associate professor in the Department of radio physics and electronics, Candidate of Technical Sciences, Francisk Skorina Gomel State University 102, Sovetskaya Str., Gomel 246019, Belarus, tel. +375 232 578854, maximenko@gsu.by

Georgy Baevich, Senior Lecturer in the Department of radio physics and electronics Francisk Skorina Gomel State University 102, Sovetskaya Str., Gomel 246019, Belarus, tel. +375 232 578854, baevich@gsu.by

Alena Melnikova, associate professor in the Department of Laser Physics and Spectroscopy, Candidate of Physical and Mathematical Sciences, Physics Faculty, Belarusian State University, 220030, Minsk, 4, Nezavisimosti ave., Republic of Belarus, tel. +375 17 2095120, tolstik@bsu.by

Alexander Lyalikov, professor in the Department of Information Systems and Technologies at YK State University of Grodno, Physics and Technical Faculty.

Researcher: YKSUG Laboratory Campus, BLK-5, 209, Grodno 230009, Belarus, tel. +375 152431279, amlialikov@grsu.by

Alexander Fedotov, professor in the Department of Energy Physics at Belarusian State University, Physics Faculty, Belarusian State University, 220030, Minsk, 4, Nezavisimosti ave., Republic of Belarus, tel. +375 17 2095425, fedotov@bsu.by

Joan Peuteman, teaching professor in the M-Group (Mechatronics), KU Leuven, Campus Bruges, Spoorwegstraat 12, B-8200 Brugge, Belgium, joan.peuteman@kuleuven.be

Mikhail Tivanov, Head of the Department of Energy Physics, Belarusian State University, Candidate of Physical and Mathematical Sciences, associate professor, Physics Faculty, Belarusian State University, 220030, Minsk, 4, Nezavisimosti ave., Republic of Belarus, +375 17 209-54-51 , tivanov@bsu.by

Natalia Strekal, professor of General Physics Department at Yanka Kupala State University of Grodno, Physics and Technical Faculty, Grodno, 230023, Ozheshko str. 22, Belarus, +375 152 743414, nat@grsu.by.

Gennady Vasilyuk, associate professor of General Physics Department at Yanka Kupala Grodno State University, Physics and Technical Faculty, Grodno, 230023, Ozheshko str. 22, Belarus, tel. +375 29 7849749, vasilyuk@grsu.by

Igor Semchenko, Vice-Rector of Francisk Skorina Gomel State University, Doctor of Sciences, Professor of Physics, Francisk Skorina Gomel State University, 104, Sovetskaya St., Gomel, 246019, Belarus, tel.+375 293 489801, isemchenko@gsu.by

Sergei Khakhomov, Rector of Francisk Skorina Gomel State University, Doctor of Sciences, Docent of Physics, Francisk Skorina Gomel State University 104, Sovetskaya St., Gomel, 246019, Belarus, tel.+375 296 913316, khakh@gsu.by

Contents

INTRODUCTION

Photonics is a field of science and engineering associated with light radiation (photons) used in optical and optoelectronic elements and systems. Photonics covers the principles of designing, operation, and implementation of the devices, where optical signals of the ultraviolet, visible, and infrared regions (terahertz range including) are generated, transformed, propagating, and detected. At the present time all the latest information technologies are based on the principles of photonics. Photonics that has occurred at the junction of laser physics, optics, and quantum radio physics is oriented to solving the problems of the classical electronics when using optical radiation (photons) instead of the electric current (electrons). Photonics is called the electronics of the XXI century.

The development of photonics in the Republic of Belarus is based on such classical courses of scientific research as laser physics and nonlinear optics successfully realized since the beginning of the 60-ies of the last century due to the advent of lasers. Now photonics is recognized as one of the research trends of high priority. Because of this, training of the specialists in the field of photonics at Master's level is an important task of our University and of all other universities in the Republic.

Photonics involves the materials, devices, techniques, and technologies which are intended for transmission, recording, processing, mapping, and storing of information on the basis of material carriers – photons. Presently, the principal problem of photonics is miniaturization and integration of optical elements and devices, creation of multipurpose optical materials and systems, conversion of analog devices to the digital ones, development of new-generation computer technique, etc. Photonics is understood as a filed embracing laser physics, nonlinear optics, optical holography, fiber optics, integrated optics, optoelectronics. The related fields (optoinformatics and optical data processing) also become more and more important.

In modern laser systems radiation frequency doubles and cascaded multipliers for the third, fourth and higher harmonics are widely used now. The parametric optical generators operating on the basis of the principles of nonlinear optics are created. High-power laser systems are impossible without the elements of adaptive optics using the wavefront conjugation effect. At the junction of waveguide optics and laser physics a new trend has been developed that is associated with design of high-power fiber lasers. Due to the nonlinear and optical methods in spectroscopy,

the potentialities of spectroscopic studies in material science have been greatly improved.

Most extensively used in different fields of science and technology are holographic systems for recording, storage, and processing of information; holographic interferometry, diffraction optical elements; holographic technologies for protection of documents and securities. Holographic principles are used to solve the problems of adaptive optics including systems for the formation of light fields with the desired spatial structure.

Modern communication systems are based on fiber-optical devices enabling telephone communication and Internet services. Waveguide systems are effectively used in science, engineering, and medicine: high-power beams of laser radiation are implemented for laser welding and cutting, in laser surgery and cosmetology, etc.

Studies in the field of optical data processing are also in progress. Apart from the classical analog techniques based on Fourier transforms of images, optical bistable elements have been created to offer digital processing of optical signals. With the development of such systems, optical processors have been designed. A great interest to the indicated studies is associated with the possibility to use in data processing the advantages of optical methods: parallel processing of signals and commutation of numerous channels, direct storage of images, such integral transformations as correlation and convolution. Optical frequency range offers a wider transmission band and hence enables extremely fast response as compared to the radio-frequency bandwidth.

Quantum optics as a combination of a quantum field theory and physical optics is developing at a great pace, we can name studies of squeezed states in a light field; atomic coherence; development of a laser without the population inversion, of nanosized laser systems and even single-atomic lasers.

In nanophotonics special attention is given to the processes of propagation, transformation, and generation of optical radiation by the nanostructures; to the development of nanostructured optical devices from lasers to biochips. These studies involve the nanostructured optical fibers, light-emitting diodes based on heterostructures, and photonic crystals enabling the control of light beams at a microlevel.

Continuously developing photonics necessitates the use of innovative technologies and advanced technologies from other fields. The photonic devices have found application in all spheres of our everyday life including optical communications, visualization, data processing and storage, energy-saving

technologies for illumination equipment and material processing, biophotonics, laser medicine, etc. The creation of new systems for data processing and communication requires the creation of innovative materials and technologies. Nanophotonics embracing photonics and nanotechnologies is associated with the architecture development and with the production technologies of nanostructured devices for the generation, amplification, modulation, transmission and detection of electromagnetic radiation. Besides, nanophotonics is effective in studies of the physical phenomena determining the operation of nanostructured devices, which are proceeding when photons are interacting with the nanosized objects. The developments in this field are of great importance for the dynamically progressing integrated optics. At the same time, the innovative approaches to the interactions between light and materials are influencing the latest technologies in laser material processing, laser medicine, superhigh-resolution microscopy. The development of innovative methods for the light beam transformation makes it possible to design advanced devices for diagnostics of different media and objects, for communication and data storage. Considering significant advances in the development of electronic computer systems, it seems logical to study closer the possibilities of designing the electronic-optical computer with the balance of its electronic and optical parts dependent on the problem at hand – such an approach will contribute to optimal realization of the advantages inherent in both parts. A significant role of the innovative technologies in the field of photonics may be demonstrated by examination of the data for the world market, where the share of photonic technologies comes to several dozens of billion dollars a year, and is still growing.

As any other research field, photonics is in need of highly qualified specialists capable of working in the related fields. Material studies, development of new approaches for the creation of photonic devices are impossible without competent people who have adequate knowledge in the field of material structure, interaction between electromagnetic radiation and material, characteristics of electromagnetic radiation, quantum effects. Training of the specialists in the field of photonics, apart from the courses of general physics, necessitates profound knowledge in spectroscopy, material science, optics of condensed media and nanostructures, quantum mechanics, nonlinear optics, laser physics, laser-material interactions, optical data processing, and some other courses depending on the specialization. The presented book «Photonics» is devoted to all these leads, covering the fundamentals of photonics which are considered in the

following sections: «Laser physics», «Nonlinear optics», «Coherent optics and holography», «Optoelectronics», «Fiber optics», and «Nanophotonics».

Chapter 1. Laser physics

Introduction

Great interest to lasers and laser systems for a period of several decades is associated with the ever growing sphere of their applications. The use of laser systems in research, engineering, medicine, military department necessitates designing of compact semiconductor lasers and laser chips with the dimensions from a few microns to several millimeters as well as high-power laser systems requiring hundreds of square meters for their arrangement. For example, laser systems used in nuclear fusion offer the production of nanosecond single laser pulses with the energy amounting to hundreds of kilojoules at a peak power of ~ 1014 W. Focusing of such a pulse on the target a few tens of microns in size results in the power density about 1020 W/cm2 for which the intensity of the light wave field is greater than the intraatomic field strength by several orders of magnitude. These high-intensity fields enable the development of principally new nonlinear processes of the interaction between light and matter. Various nonlinear optical effects make it possible to control the temporal and spectral characteristics of laser radiation. Owing to the methods of nonlinear optics, the transfer to the femtosecond-range laser pulses becomes possible when a pulse represents only a single period of the light wave.

The word «laser» is formed from the first letters of the English expression «Light Amplification by Stimulated Emission of Radiation» – most important feature characteristic for all the laser devices.

The notion of stimulated emission was first introduced by A. Einstein in 1916. In his work «Emission and Absorption in a Quantum Theory» A. Einstein has postulated that the radiative transition from upper to lower molecular energy levels may be of two types: spontaneous and stimulated. Spontaneous transitions are realized without the external effects, whereas stimulated transitions are caused by exciting radiation.

This postulate was supported in 1927 – 1930 by P. Dirac who advanced a quantum electrodynamic theory of radiation and calculated the probability of systems's transition from upper to lower level with the emission of photons having the given energy, propagation direction, and polarization. This probability has two components: one of them is independent of an external field and the other is directly proportional to the photon density. By Dirac's theory, the photons originating as a result of stimulated transitions are physically identical to those of exciting radiation. This means that there is no way to distinguish between the

initial and the emitted photons. Their characteristics (propagation direction, polarization, frequency, and phase of a light wave) are congruent.

The next step in the development of laser physics is associated with the idea, suggested by the Soviet physicist V.A. Fabrikant in 1939, that electromagnetic radiation is amplified on its transmission through a material by means of stimulated emission. According to V.A. Fabrikant, due to the introduction of impurities into a gas discharge, the number of particles at a lower level may be decreased compared to that at a higher level leading to the creation of the inverse population. When such a medium is subjected to the effect of a luminous flux at the frequency associated with a resonance transition between the levels, the intensity of exiting radiation is higher than that of incident radiation.

In 1951 V. A. Fabrikant, M. M. Vudynski, and F. A. Butaeva developed and registered the method to amplify electromagnetic radiation using media with the inverse population. Their claim is very informative: "The method to amplify electromagnetic emissions (ultraviolet, visible, infrared, and radio frequency bands of waves) distinguished by transmission of the amplified emission through a medium, where, by means of an additional emission or in some other way, the concentration of atoms, other particles or of their systems at the upper energy levels associated with the excited state is excessive compared to the equilibrium concentration". In fact, these researchers have suggested the idea how to create an active laser medium. In 1964 they registered the discovery with the priority from 18 of June 1951.

Unfortunately, the publication describing Fabrikant's invention was published only in 1959 and had no marked influence on the origination and progress of laser physics. In 1952 the Soviet physicists N. G. Basov, A. M. Prokhorov (P. N. Lebedev Physical Institute of the USSR Academy of Sciences, Moscow) and independently the American physicist Ch. Townes (Columbia University, New-York City) presented their reports about the possibility to amplify emission in the microwave range . In 1954 the first quantum oscillators were created on the basis of the beam of ammonia molecules (masers) producing coherent radiation at the wavelength 1.26 cm. For their fundamental works in the field of quantum electronics, which have resulted in the development of lasers and masers, N.G. Basov, A. M. Prokhorov, and Ch. Townes were awarded the Nobel Prize in 1964.

After realization of lasing in the radio frequency band, the researchers have directed their efforts to the optical range. In his lecture at the Nobel Prize ceremony A. M. Prokhorov has noted that, despite expectations, quantum

oscillators for the optical range did not followed soon after the creation of masers in the radio frequency band, they have appeared only five-six years later because of two problems: (1) unavailable cavities for the optical wavelength range; (2) inexistent systems and methods to produce the inverse population in the optical range.

The indicated problems were solved in 1960 by T. Maiman who has developed the first laser (optical quantum oscillator) based on ruby crystals to generate pulses of monochromatic radiation at the wavelength 694 nm. The inverse population was produced with the help of a helical flash lamp, and a plane-parallel Fabry-Perot interferometer was used as a cavity.

The «laser» period in optics begins since the advent of this first ruby laser. In the end of the sixties Ali Javan, W. Bennett, and J. Harriot have developed the first continuous-wave gas laser (633 nm) based on the mixture of helium and neon. The electric charge in the low-pressure gas mixture was used as a pumping source for the active medium. In 1961 E. Snitzer suggested to use glass activated by trivalent neodymium ions as a laser medium at the generation wavelength 1.06 μm. At the end of 1962 – beginning of 1963 the generation of infrared radiation on the current injection through the p–n junction in gallium arsenide was reported at once in several research centers of the USSR and USA facilitating the development of different-type semiconductor lasers.

The first Nd:YAG (neodymium-doped yttrium aluminate) laser with the generation wavelength 1.064 μm and the first CO2 laser with the wavelength 10.6 μm were created in 1964. At the present time Nd:YAG lasers are extensively used in the pulsed (peak power coming to hundreds of megawatts) and continuous (kilowatt powers) modes of operation. Owing to the use of laser systems on neodymium glass with a cascaded amplification, one can produce nanosecond single pulses with the energy approximating one hundred kilojoules at the peak power exceeding a hundred of terawatts. Gas-dynamic CO2 lasers offer generation of continuous radiation with the power of about one hundred kilowatts.

In 1966 the Belarusian physicists B. I. Stepanov, A. N. Rubinov, V. A. Mostovnikov succeeded in producing the generation based on the dye solutions excited by the pulses of a ruby laser. In the same year such generation was independently realized in the USA by P. P. Sorokin, D. R. Lankard in the process of studies of stimulated Raman scattering and in Germany by F. P. Schäfer, W. Schmidt who have studied saturation of spontaneous fluorescence. In1967 the scientists of Belarus and of the USA produced generation on pumping of dyes by emission of flash lamps to make dye lasers

independent sources of coherent radiation. A great choice of the dyes has enabled covering of the whole spectral range from the near UV to the near infrared region. The principal advantage of dye lasers is the possibility to produce radiation smoothly tunable over a wide range of the wavelengths with a narrow spectral line.

To conclude, in the period of about ten years the researchers have created different laser systems operating in the pulsed and continuous-wave modes with the use of different active media (solid-state, gas, liquid) and various pumping sources (optical, electrical, ionization, and so on). These achievements have contributed to rapid progress in laser physics, offered new possibilities in finding the pattern of interaction between light and matter, enabled one to establish new nonlinear-optical phenomena extensively used in different spheres including research, engineering and technology, medicine.

1.1. Principles of laser operation and characteristics of laser radiation. Methods of active medium pumping. Optical resonators.

The operation of lasers is based on the fundamental processes of interaction between an electromagnetic field and a material. Let us consider a medium composed of identical particles (atoms, ions, molecules), each particle having a pair of coupled states with the energies E_1 and E_2 (Fig. 1.1.1.). If we neglect all other energy levels, for a two-level resonant medium the energy-conservation processes are possible: photons with the frequency $\nu = (E_2 - E_1)/h$ (where h – Planck constant) are emitted or absorbed by the medium particles which are involved into the transitions between the energy states E_1 and E_2. When an electromagnetic wave with the resonance transition frequency ν is incident on the medium, there exists the possibility that a particle goes to the upper energy level E_2. This process is known as absorption and its probability is proportional to the volume energy density of an electromagnetic wave $U(\nu)$. The proportionality factor is determined by the Einstein coefficient for stimulated transitions with absorption, B_{12}, that characterizes the spectral properties of a resonant medium. The product $B_{12}U$ has the dimensions of (time)$^{-1}$ and determines the resonance transition probability.

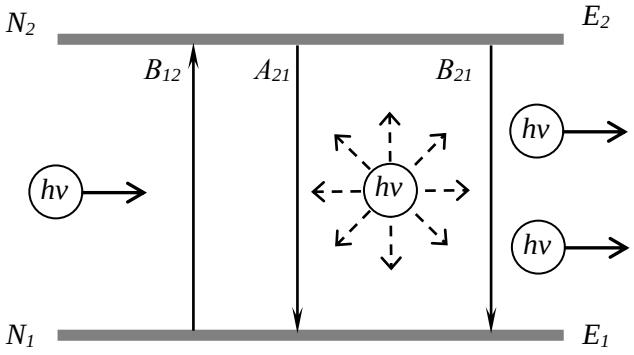


Fig. 1.1.1. – Schematic of the energy levels of a resonant medium including the spontaneous A_{21} and stimulated B_{12}, B_{21} transitions

Because of this, a dynamics of the absorption process may be described by the following kinetic equation:

$$\frac{dN_1}{dt} = -B_{12}U\ N_1, \qquad (1.1.1)$$

where N_1 – population of the level E_1 (number of the particles within the unit volume which are in the ground state).

Now we consider a behavior of the particles in the excited state E_2 with the population N_2. For these particles there is the probability A_{21} of spontaneous transitions to the lower state E_1 with emission of the photons possessing the energy $h\nu$:

$$\frac{dN_2}{dt} = -A_{21}N_2. \qquad (1.1.2)$$

Besides, there is the probability $B_{21}U$ of stimulated transition with photon emission in presence of radiation having the energy density U:

$$\frac{dN_2}{dt} = -B_{21}UN_2. \qquad (1.1.3)$$

The Einstein coefficients for the spontaneous A_{21} and for the stimulated B_{12}, B_{21} transitions are related as

$$A_{21} = \left(\frac{8\pi h\nu^3}{c^3}\right)B_{21}, \qquad B_{12}g_1 = B_{21}g_2, \qquad (1.1.4)$$

where c – speed of light in a medium; g_1 and g_2 – degeneracy degree for the corresponding energy levels.

Resonance transitions with emission and absorption of photons lead to variations in the intensity of a light beam when it propagates within the medium volume. Considering the features of stimulated emission, when the emitted photons are not different from those stimulating them (their propagations directions, polarizations, frequencies and phases are coincident), we can write an expression for the power absorbed within the unit volume as follows:

$$W_{\text{погл}} = \left(B_{12}N_1 - B_{21}N_2\right)Uh\nu\,. \qquad (1.1.5)$$

Taking into account the relationship between the light field intensity I and the volume energy density U ($I = U\,c$), from formulas (1.1.5) it follows that for the absorption factor we have

$$K_{\text{погл}} = \frac{h\nu}{c}\left(B_{12}N_1 - B_{21}N_2\right). \qquad (1.1.6)$$

As seen from equation (1.1.6), on condition that $B_{12}N_1 < B_{21}N_2$, i.e. $N_1 < (g_1/g_2)N_2$, the absorption factor is negative. This state of the medium is known as an inverse population and the medium itself is termed as the amplifying or active medium − the light beam transmitted through this medium is amplified.

1.1.1. Methods of the active medium pumping

At the present time the creation of inverse population, or pumping of a medium, is realized by different methods. Selection of the pumping method is dictated by the medium itself, by its spectral and luminescent properties. Optical pumping is usually used for solid-state and dye lasers. A source of pumping light should provide a high energy density in the spectral region that is responsible for population of the upper energy level. This is realized with the use of flash lamps, various constant radiating lamps, light-emitting diodes or of radiation emitted by other lasers.

Gas lasers are pumped with the use of different gas discharge mechanisms. Due to electron or atomic collisions, the upper energy levels of a medium become populated. When only one kind of gas is used (as in noble gas ion lasers), excitation is caused directly by electron collisions. Electric pumping and resonant energy transfer between different atoms is used in a helium-neon laser that was the first gas laser.

Semiconductor lasers are electric current pumped, and electrons are injected into the upper region acting as the upper laser level.

With the pumping based on chemical reactions, the molecules of a material are formed already in the excited state as a result of the reaction. Also, an inverse

population is brought about by nonstationary gas-dynamic processes: at the first stage a mixture of gases is heated to 2000 K and then drastic expansion leads to quicker depletion of the lower energy levels.

1.1.2. Resonators

To produce generation, an active medium is placed into a resonator offering a positive feedback. A laser resonator comprises two mirrors, with the reflection factors R_1 and R_2, positioned in parallel to each other. Spontaneously emitted photons participate in initiation of the generation. Being reflected from mirrors of the resonator, these photons initiate stimulated emission in an active medium. The spontaneously emitted photons which are propagating along the resonator axis and have maximal path in the active medium are predominantly amplified due to the induced radiation.

The generation mode is realized on condition that amplification is higher than the resonator loss associated with a partial reflection from the mirrors, with scattering from inhomogeneities of the active medium, and diffraction from the resonator mirrors (aperture of the active element). If the loss is determined only by transmission of the mirrors, the generation threshold is attained when the following condition is met:

$$R_1 R_2 \exp\left(2K_{\text{ус}} l\right) = 1, \qquad (1.1.7)$$

where R_1 and R_2 – reflection factors of the resonator mirrors; $K_{\text{ус}} = -K_{\text{погл}}$ – amplification factor [given by expression (1.1.6) for the established inverse population]; l – length of the active element. The condition of (1.1.7) points to the fact that threshold is reached when the loss for the round trip of the resonator (reflection from both mirrors) is compensated by amplification at double pass of the active element.

An optical resonator in the simplest case represents a pair of mirrors at the common optical axis (e.g., Fabry-Perot interferometer). A distance between the mirrors is much greater than the wavelength ($L >> \lambda$). Such resonators are called the open resonators because they have no lateral surface. The main types of resonators are shown in Figure 1.1.2.

For symmetric confocal resonator *1* a distance between the mirrors is equal to the curvature radius of the resonator mirrors; a radius of the mirrors for spherical resonator *2* equals a half of the distance between them. For hemispherical resonator *3* a distance between the mirrors equals a half of the

spherical mirror radius. The confocal, spherical, and plane-parallel *4* resonators are at the stability threshold. In other words, with the introduction, e.g., of diffraction losses into a theoretical study, radiation passes beyond the resonator. Because of this, most common are the resonators, where one mirror is plane, and the other is spherical, its curvature radius greatly exceeding the resonator base. In this case diffraction losses in the resonator are compensated for and the output laser beam has a minimal divergence.

Due to multiple reflection from the resonator mirrors, the standing waves are formed, the constant phase and amplitude relations of which connect the waves propagating in opposite directions. The stable spatial configurations of an electromagnetic field in a resonator, repeated on multiple roundtrips of the resonator, are called the modes or oscillation types. Every mode is characterized by the corresponding field configuration on the mirror surface and the number of half-waves for the resonator length. The modes are designated as TEM_{mnq} (abbreviation for *transverse electric and magnetic*), where m and n are integers, the so-called transverse or angular mode numbers determining the field minima in the mirror plane; q – longitudinal (axial) mode number giving the field maxima along the resonator axis. The resonator modes differ in the amplitude, phase, and frequency distributions, and also in magnitude of diffraction losses.

Apart from the resonator characteristics, a frequency spectrum of laser generation is associated with positions of the energy levels (sublevels) of an active medium. Laser generation takes place for the resonator modes whose amplification factor is higher than the loss. The spectral dependence of the amplification factor K_{yc} is determined by different factors responsible for the spectral line broadening. We distinguish three broadening types: natural, collisional, and Doppler. Natural broadening is associated with a finite lifetime of the excited states; collisional – with a reduced lifetime of the excited state and with an abrupt change of the radiation phase at the instant of collisions of radiating

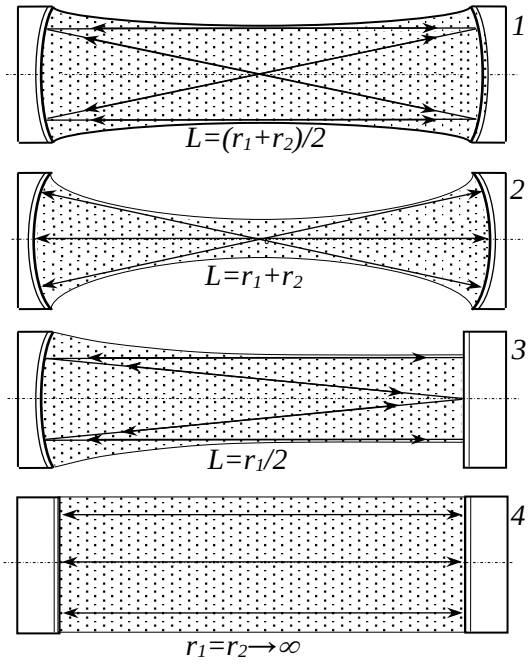


Fig. 1.1.2. Resonator types

particles; Doppler broadening – with atomic motion in various directions at different frequencies. Based on Doppler effect, instead of one frequency, the atoms emit at several frequencies in a particular interval. For Maxwell distribution of the atoms in rates, a width of the Doppler profile is given by

$$\Delta\nu_D = \frac{2}{\lambda}\sqrt{\frac{2RT}{M}\ln 2}, \qquad (1.1.8)$$

where R = 8.31 J/(K×mol) – universal gas constant; T – temperature of atoms; M – molar mass. In the case of neon atoms (M = 20 g/mol) at T = 400 K for λ = 632.8 nm we have $\Delta\nu_D$ = 1.5 GHz. For the same spectral line the natural width $\Delta\nu_e$ = 1.9 MHz, whereas the collisional width comes to $\Delta\nu_{\text{ст}}$ = 0.6 MHz (at the pressure p = 0.5 in millimeters of mercury). All the broadening mechanisms of a spectral line act simultaneously but, as $\Delta\nu_D >> \Delta\nu_e > \Delta\nu_{\text{ст}}$, for a helium-neon laser the amplification factor profile is virtually Doppler. This means that the generation is possible simultaneously at different frequencies associated with the resonator modes in the spectral range, where the amplification factor is higher than the loss.

1.1.3. Longitudinal resonator modes

To analyze the longitudinal mode structure of radiation, let us consider a resonator with two plane mirrors (Fabry-Perot interferometer) having the reflection factors for the laser radiation intensity R (reflection factor for the light field amplitude is $\sqrt{R}$). Let a plane wave with the amplitude A_0 be propagating along the optical axis. After multiple reflections from the mirrors, the field strength in the resonator is given as

$$E = A_0\left(1 + R\exp(-i\varphi) + R^2\exp(-2i\varphi) + ...\right), \qquad (9)$$

where $\phi = 4\pi L/\lambda$ – phase incursion of a light wave for a double path in the resonator. Using a sum of the geometric progression, we can write an expression for the light intensity in a resonator as follows:

$$I \sim \frac{1}{1 + F\sin^2(\varphi/2)}, \qquad (10)$$

where $F = 4R/(1-R)^2$. The intensity in a resonator is at maximum when the condition $\varphi/2 = q\pi$ (q – integer) is met, that may be transformed to

$$L = q\lambda/2. \qquad (11)$$

Fig. 1.1.3 shows the spectral dependences for the superthreshold amplification factor (see Fig. 1.1.3, *a*), interferometer transmission (Fig. 1.1.3, *b*), and generation power (Fig. 1.1.3, *c*. As seen, a spectrum of generation represents a set of the equidistant longitudinal modes the frequencies of which are coincident with the resonator modes.

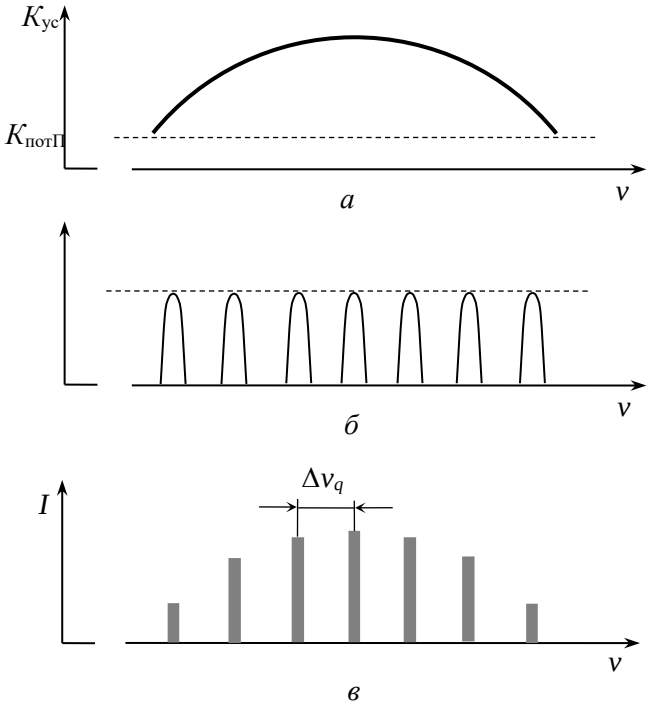


Fig. 1.1.3 The effect of resonator parameters
on the generation frequency spectrum

Expression (1.1.11) means that along the resonator length an integer number of half-waves should be present, i.e. a standing wave should be formed. The states of a light field for different values of q differ by the number of nodes along the resonator axis and hence are associated with different longitudinal modes.

As follows from (1.1.11), the generation is possible at the frequencies

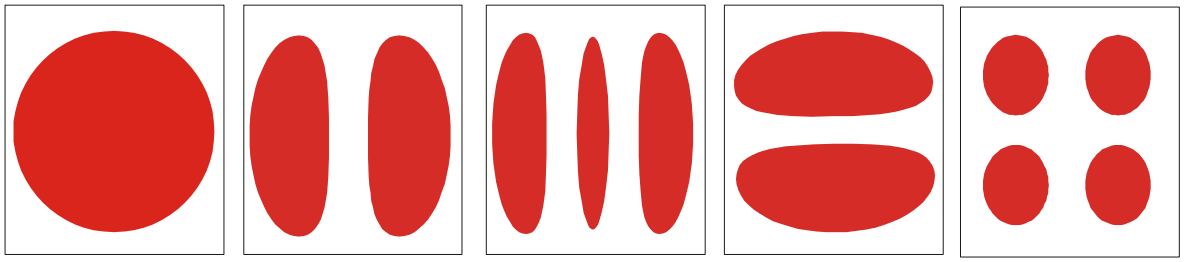
$$\nu_q = qc/2L\,. \qquad (1.1.12)$$

It is easy to determine the intermode interval (difference between the adjacent modes)

$$\Delta\nu_q = c/2L\,. \qquad (1.1.13)$$

1.1.4. Transverse resonator modes

Transverse resonator modes are formed by the waves propagating at a certain angle to the resonator axis; these modes are associated with stable light-field configurations reproducing themselves in the process of the resonator round-trip. The simplest transverse mode of a confocal resonator presents the beams propagating along the axis. This mode tem_{00} (Fig. 1.1.4, *a*) is characterized by the Gaussian intensity distribution. The following modes at the output of the resonator represent a set of several light spots (Fig. 1.1.4, *b*, *c*, *d*, *e*). These modes necessitate several round-trips of the resonator to present a closed trajectory. The mode subscripts indicate numbers of minima on horizontal (first subscript) and vertical (second subscript) scanning of the intensity. Practically, the higher-order modes have higher diffraction losses than TEM_{00} mode. Because of this, lasers are ordinary designed so that their generation be realized at zero transverse mode.



а *б* *в* *г* *д*

Figure 1.1.4. Transverse resonator modes: *a* – TEM_{00}; *b* – TEM_{10}; *c* – TEM_{20}; *d* – TEM_{01}; *e* – TEM_{11}

1.1.5. Properties of laser radiation

The laser generated radiation is characterized by principally new properties as compared with the previously known radiation sources. First of all, we should name such a property as coherence. The notion of *coherence* (from Latin cohaerens – being in coordination with something) that stems from a vibration theory defines coordination of several vibrational or wave processes in time. Coherence of a laser is predetermined by a nature of the stimulated transitions in an active medium, which are accompanied by emission of the photons absolutely identical to those stimulating them. Coherence of a beam, temporal intensity fluctuations, frequency spectrum are manifestations of the same physical properties of radiating particles. As a consequence, laser sources are remarkable

for their high monochromaticity, high spectral power, directivity, and small divergence.

Stimulated emission representing a resonance process is related to the resonance transition frequency. However, there are some factors leading to (natural, Doppler, collisional) broadening of a spectral line of the generation. The spectral line width $\Delta\nu$ determines the coherent properties of radiation; these properties may be characterized by the coherence time $\tau_{\text{ког}} \sim 1/\Delta\nu$. The precise relationship between the line width and the coherence time is dependent on the form of a spectral line. For the Lorentzian profile form that describes natural or collisional broadening we have $\tau_{\text{ког}} = 1/\pi\Delta\nu$. For the Gaussian profile (Doppler broadening) we have $\tau_{\text{ког}} = \sqrt{(2\ln 2)/\pi}/\Delta\nu$. Time coherence governs the correlation of a light field at one point of the space separated by the time interval τ. Within the time interval $\tau < \tau_{\text{ког}}$ the phase difference of electromagnetic oscillations emitted by a source is invariable.

The generation line width for typical laser sources is from several thousands of nanometer (narrow band gas lasers) to several nanometers (dye lasers). By the use of selective resonators we can further narrow the generation line and attain the coherence hundreds of meters or even dozens of kilometers long. It should be noted that ordinary narrow-band incoherent light sources, e.g. consider some spectral lines of a sodium lamp, are associated with coherence time about $\sim 10^{-10}$ s at the coherence length coming to several centimeters. For thermal sources ($\tau_{\text{ког}} \sim 10^{-12\,-13}$ c) a coherence length is on the order of one hundred microns and higher.

Apart from time coherence, that determines a degree of monochromaticity for a laser source, we distinguish the notion of «spatial coherence». Spatial coherence determines the correlation of a light field at different spatial points and at the same instant of time. For example, in analogy with Young's experiment, relating the source size a (distance between the slits), angular aperture φ, radiation wavelength λ, we can determine the limiting ratio for which the interference pattern is still observed as $\varphi \approx \lambda/a$. This ratio is coincident with an expression for the diffraction divergence angle.

The induced radiation is accompanied by emission of a photon in the same direction the initial stimulating photon had. In this way, provided the fundamental transverse mode is excited (when in a resonator the electromagnetic waves propagating along the resonator optical axis are maintained and all the generated photons are absolutely coherent), we have fully spatially-coherent light. When

higher transverse modes are excited, the spatial coherence decreases leading to the increased radiation divergence. Typical divergences of laser sources are ~ 10^{-3}–10^{-4} rad offering the possibility to transmit signals to large distances (at a distance of 1 km the light spot diameter may be smaller than 1 m). Besides, small divergence of laser radiation makes it possible to have a micrometer focusing region and to provide millions as high concentration of the spatial energy. Owing to small divergence in combination with a narrow spectral line for generation, an extremely high radiation brightness is possible. To illustrate, even low-powered semiconductor lasers or helium-neon lasers with a power of several milliwatts is higher than the spectral brightness of the Sun. With the generation duration reduced to the nano- or picosecond range, the power of light pulses of the standard laser systems may be as high as several megawatts or even gigawatts, exceeding the power of large atomic power stations. The concentration of electromagnetic radiation energy in space, time or in a narrow spectral range offers new potentialities for the propagating light radiation and for its interactions with various media. One can observe nonlinear effects which result in transformations of the radiation frequency (harmonic generation, frequency summation and subtraction, induced scattering), formation of the light fields with «unusual» properties (wavefront conjugation, squeezed states, spatial-temporal solitons, optical vortexes, etc.). All these possibilities became practicable only with the advent and fast progress of lasers.

1.2. Continuous mode of laser operation. Power generation. The lasing threshold. Free-running mode.

Fig. 1.2.1 shows a schematic diagram of a laser operating in the free-running mode. The basic physical process underlying the operation of lasers is stimulated emission by the excited atoms or molecules. The photons, having absolutely identical characteristics (propagation direction, frequency, polarization, and phase), are generated in the process of stimulated emission determining such unique properties of laser radiation as directivity, monochromaticity, coherence. At the initial stage of generation, due to spontaneous emission, photons are formed in various directions enabling one to find the positions of mirrors in an optical resonator. The photons propagating in the directions which are not coincident with the resonator axis escape from the resonator, whereas those propagating along its axis are reflected from the mirrors passing multiply through the active medium. The number of photons is growing with every trip in the active

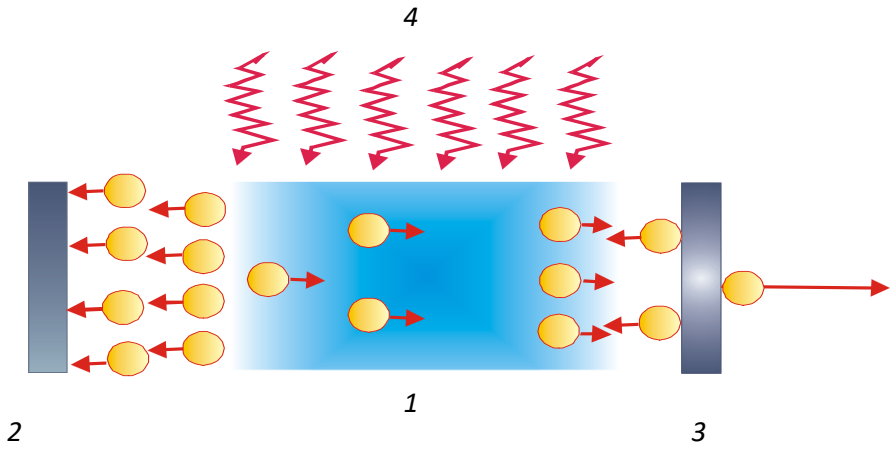


Fig.1.2.1. Basic schematic of laser operation: *1* – active medium, *2*, *3* – resonator mirrors, *4* – optical pumping system

medium, this growth being avalanche in character (Fig. 1.2.1.). To illustrate, when for one trip the number of photons is increased by a factor of M, after N trips the number of photons increases by a factor of M^N.

Generation of laser radiation is possible at frequencies associated with the resonator modes (stable field configurations within the resonator repeated in the process of multiple propagation of waves between the mirrors) on condition that a power of the induced transitions is higher than a power of the energy loss in the resonator. In a resonator most considerable are useful losses associated with partial mirror transmission (offering radiation coupling from the resonator) and harmful losses due to radiation absorption and scattering by the mirrors and by optical inhomogeneities of the active medium.

Spike structure of free-running generation

Solid-state lasers (yttrium aluminum garnet laser, neodymium laser, ruby laser, etc.) having wide absorption and emission bands are characterized by a great number of the modes within the limits of the amplification line width for the active medium. Every mode has its spatial distribution of the nodes and antinodes of a standing wave within the medium volume and this distribution influences the spatially inhomogeneous removal of inversion. As a particular mode is generated, in its nodes some regions of the active medium have a high level of inversion leading to generation of a new mode with other positions of the nodes and antinodes. Laser radiation represents a train of short (about 1 μs) pulses irregular

in time with random amplitudes. The generated mode varies on going from one antinode to another.

A typical oscillogram for the radiation pulse of a laser operating in the free-running mode is given in Fig. 1.2.2. This is the so-called *spike* structure of a laser pulse. Spikes are characterized by short duration < 1 μs by different amplitudes, and their duration is chaotic in time. Chaoticity of oscillations is associated with the multimode structure of laser radiation, inhomogeneity of the active medium, and irregularity of pumping.

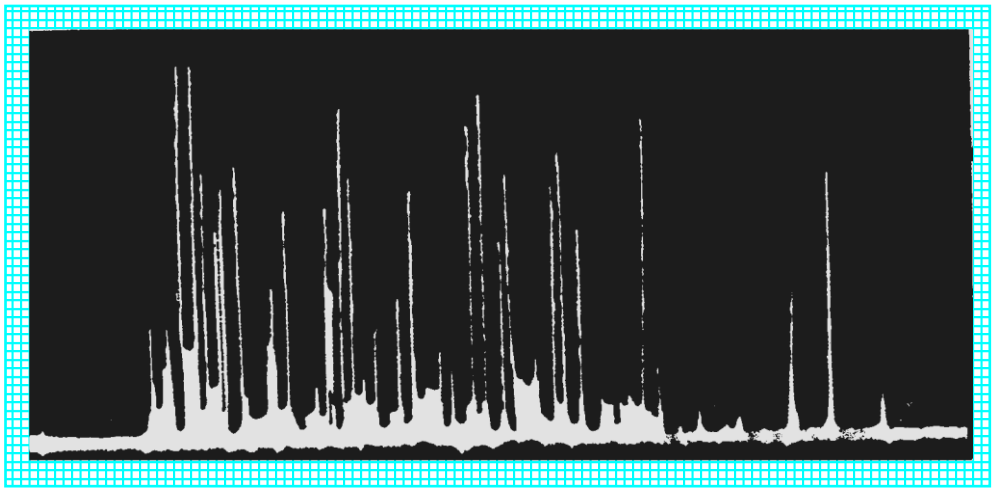


Fig. 1.2.2. Spike structure of laser radiation in the free-running mode

The generation process in solid-state pulsed lasers with lamp pumping is as follows. A high-voltage condenser is discharged through the lamp, its light is focused at the active element. The pulse length is dependent on the lamp type coming to 0.1 – 10 ms. The upper curve in Fig. 1.2.3 gives the lamp power as a function of excitation time.

The curve in the middle demonstrates a dynamics of the amplification factor that is growing until its value approaches the loss factor. With initiation of generation, the inverse population is decreased and the amplification factor is lowered below the loss. A laser (lower curve) produces a single pulse and the generation is terminated. But, as optical pumping of the active medium still continuous, the amplification factor begins to grow and a pulse of laser radiation is generated again. The process is repeated till the pumping power is sufficient to provide the amplification factor that approaches the loss. In this way a pulse of a solid-state laser represents a set of spikes.

Among high-power solid-state lasers, most widespread are lasers, where the active medium is represented by glass or yttrium-aluminum garnet crystal

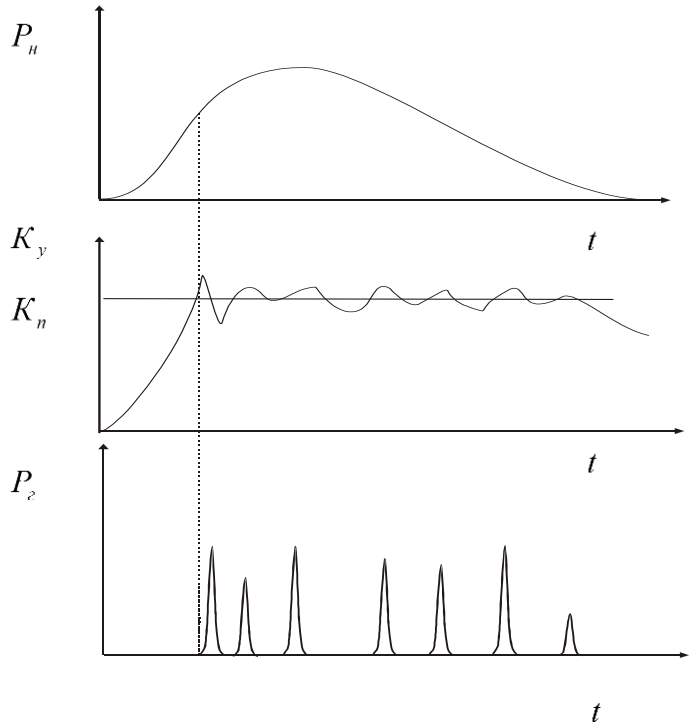


Fig. 1.2.3. Origination of the spike structure of laser radiation

($Y_3Al_5O_{12}$) with the addition of the trivalent neodymium ion Nd^{3+} (Nd^{3+}:YAG). Neodymium glass is characterized by a lower amplification fcator and a greater spectral-line width; the mode structure of neodymium glass lasers is high. The glass technology is well-developed offering the production of components of any shape or size: from fibers several micron in diameter to disks with the diameter of one meter. As early as the 80-ies of the XX century a laser system based on neodymium glass was designed to generate pulses with the peak power 100 TW and the total energy $\cong 100$ kJ («Nova»-laser). This laser has several amplifiers based on neodymium glass, the greatest of them representing a disk 4 cm thick and 75cm in diameter. Unfortunately, its repetition rate was strongly limited due to very low thermal conductivity of glass (one tenth as large as that of the crystal). The limitation is removed in yttrium aluminum garnet lasers (Nd:YAG) operating both in the continuous-wave and pulsed modes. The output parameters of this laser are as follows:

- lasing power in the continuous-wave mode up to 1 kW;
- average radiation power in a pulsed laser with a high repetition rate (≥ 50 Hz) up to 1 kW;
- peak power in the Q-switching mode (pulse length ~ 10^{-8} s) ~ 10^8–10^9 W;
- peak power in the mode-locking regime (pulse length ~ 10^{-11} s) ~ 10^{10}–10^{11} W.

Energy characteristics of free-running generation

As noted, the generation occurs when amplification of an electromagnetic wave for a single round-trip of the resonator is higher than its attenuation due to losses (useful losses, K_r, associated with the radiation output through a semitransparent mirror and harmful intracavity losses – ρ).

We describe the free-running mode taking a four-level scheme of the active medium as an example. This scheme is characteristic for neodymium lasers and yttrium aluminum garnet lasers (Fig. 1.2.4). Optical pumping results in occupation of the upper energy level (transition 1 – 4). Then, due to radiationless transition, a metastable level acting as an upper laser level is occupied (transition 4 – 3). As a lower laser level, we consider level 2 with a minor population because of radiationless transitions 2 – 1.

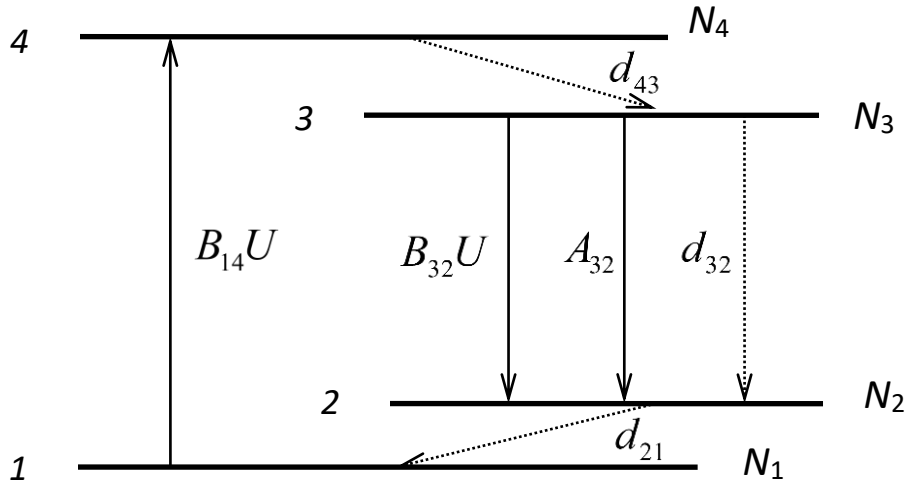


Fig. 1.2.4. Four-level scheme of energy states for a resonant medium including the stimulated $B_{ij}U$, spontaneous radiative A_{ij}, and radiationless d_{ij} transitions

Note that the first – ruby – laser was operated on the basis of a three-level scheme, the ground energy level playing a role of the lower laser level (levels 1 and 2 in Fig. 1.2.4 are coincident). A limitation of such a scheme was the requirement to activate more than a half of atoms to the excited state for the realization of population inversion that necessitated the use of high-power pumping. Active media operated according to a four-level scheme enable much easier achievement of population inversion due to the underpopulated lower level. The population inversion is attained when a few percent of the atoms are activated to the excited state.

Knowing a magnitude of the inverse population, we can find the active-medium amplification factor as

$$K_{\text{am}} = \frac{h\nu}{c} B_{32} \Delta N, \qquad (1.2.1)$$

where $h\nu$ – energy of one quantum; c – speed of light in a medium; B_{32} –Einstein coefficient for the stimulated transition in the generation channel; $\Delta N = N_3 - N_2$ – difference in populations of the upper and lower laser levels.

For a four-level scheme of laser generation in the approximation of the underpopulated lower laser level ($N_2 \approx 0$), we have

$$K_{\text{am}} = \frac{h\nu}{c} B_{32} N_3. \qquad (1.2.2)$$

In the stationary mode the number of particles activated by optical pumping to level *3*should be equal to that of the particles leaving that level. Without generation in channel *3–2*, we have

$$\eta B_{14} U_{\text{pump}} = N_3 P_{32}, \qquad (1.2.3)$$

where $\eta = P_{43} / \left(P_{41} + P_{42} + P_{43} \right)$ – quantum yield; P_{32}, P_{41}, P_{42}, P_{43} – total probability of spontaneous and radiationless transitions in the corresponding spectral channel; $\eta B_{14} U_{\text{pump}}$ – pumping rate of the upper laser level determining the number of active centers excited to level *4* and going to level *3*; U_{pump} – volume density of the pumping energy in a resonator.

Considering formulae (1.2.2) and (1.2.3) , we can get

$$K_{\text{am}} = \frac{h\nu B_{32} \eta B_{14} U_{\text{pump}}}{c P_{32}}. \qquad (1.2.4)$$

Laser generation is possible when the amplification factor is in excess of the total loss $K_{\text{am}} > K_r + \rho$ – this results in the threshold generation condition for the upper level pumping rate

$$\eta B_{14} U_{\text{pump}}^{\text{th}} = \frac{(K_r + \rho) c P_{32}}{h\nu B_{32}}. \qquad (1.2.5)$$

The threshold pumping rate $\eta B_{14} U_{\text{pump}}^{\text{th}}$ is a minimal pumping rate in excess of which the generation is initiated. Considering the afore-said, a power of the free-running generation is determined by the following expression:

$$W_{\text{gen}} = lSNh\nu_{\text{gen}} (\eta B_{14} U_{\text{pump}} - \eta B_{14} U_{pump}^{th}), \qquad (1.2.6)$$

where l and S – length and area of the generated volume cross-section; N – number of active centers within the unit volume; $h\nu_{\text{gen}}$ – energy of one quantum.

The generated radiation compensates for losses: harmful losses, associated with absorption and scattering by inhomogeneities of the medium, mirrors, and useful losses associated with the output of radiation, through a semitransparent mirror, beyond the resonator. The power of the radiant flux leaving the resonator is given by

$$S_{gen} = W_{\text{gen}} \frac{K_r}{K_r + \rho}. \qquad (1.2.7)$$

The useful loss factor $K_r = \frac{1}{2l} \cdot \ln \frac{1}{r_1 \times r_2}$,

where r_1 and r_2 – reduced reflection factors including the total reflection from the mirror r_M and from the active element face r_{T}, and we have

$$r = \left(\frac{\sqrt{r_{\text{T}}} + \sqrt{r_M}}{1 + \sqrt{r_{\text{T}} \times r_M}} \right)^2. \qquad (1.2.8)$$

According to Fresnel formula for the normal incidence, $r_{\text{T}} = \left((n-1)/(n+1)\right)^2$ (the refractive index of neodymium-doped yttrium aluminum garnet at the wavelength 1.064 μm $n = 1{,}816$). Based on formulae (1.2.4) – (1.2.7), the power of a light flux leaving the resonator may be represented in the following form:

$$S_{\text{gen}} = \frac{lSNcP_{32}(K_{\text{am}} - K_r - \rho) \cdot K_r}{B_{32}(K_r + \rho)}. \qquad (1.2.9)$$

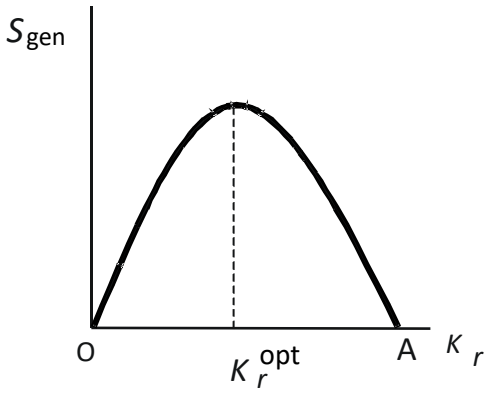


Fig. 1.2.5 Generation power as a function of the useful loss factor K_r

Fig. 1.2.5. shows the generation power as a fucntioon of the useful loss factor K_r. The point O is associated with a hundred-percent reflection from the mirrors (radiation remains within the resonator); at the point A the amplification factor is compared to the total loss ($K_{am} = K_r + \rho$ – threshold generation condition). An optimum of the useful loss factor of a laser is attained for $K_r^{\text{opt}} = \sqrt{K_{\text{am}} \cdot \rho} - \rho$. The existence of a maximum is due to the fact that, on the one hand, an increase in useful losses contributes to the increased output of the generated flux beyond the resonator and, on the other hand, a density of the radiation retained within the resonator is lowered.

The energy generated for a single pulse of the pumping lamp emission is found by integration of expression (1.2.7) with respect to the generation time t_{gen} that is not the same as the lamp emission time t_{pump}. The generation begins some time after the moment of pumping initiation – during this time the excited particles accumulate at metastable level *3* and the required population inversion of the levels is established to provide excess of amplification over the total loss in the resonator. By integration of expression (1.2.7), taking into account the relatiion of (1.2.6), we get

$$E_{gen} = lSNh\nu^{2}\eta\left(\bar{B}_{14}U_{pump} - B_{14}U_{pump}^{th}\right)\ t_{gen}\frac{K_r}{K_r+\rho}, \qquad (1.2.10)$$

where

$$\bar{B}_{14}U_{pump} = \frac{W_{\text{abs}}}{lSNh\bar{\nu}_{pump}}\ - \qquad (1.2.11)$$

average value of the absorption probability in the pumping channel; $\bar{\nu}_{pump}$ – spectrally-averaged pumping frequency. In this case the absorbed power is determined by the electric power W_{pump} of a capacitor bank

$$W_{abs} = \chi\ W_{pump}, \qquad (1.2.12)$$

where χ – coefficient characterizing the efficiency of the electric to optical power conversion with the use of pumping radiation. Proceeding from expressions (1.2.11) and (1.2.12), the average probability of absorption in the pumping channel takes the form

$$\bar{B}_{14}U_{pump} = \frac{\chi}{lSNh\bar{\nu}_{pump}}\bar{W}_{pump}, \qquad (1.2.13)$$

where $\bar{W}_{pump} = E_{pump}/t_{pump}$ – average pumping power. In a similar way, we can express the threshold pumping as

$$B_{14}U_{pump}^{th} = \frac{\chi}{lSNh\bar{\nu}_{pump}}W_{pump}^{th}, \qquad (1.2.14)$$

where W_{pump}^{th} – electric power of pumping associated with the generation threshold.

Using expressions (1.2.13) and (1.2.14), equation (1.2.10) for the laser generation energy can take the form

$$E_{gen} = \frac{\nu_{gen}}{\bar{\nu}_{pump}}\eta\chi(\bar{W}_{pump} - W_{pump}^{th})\ t_{gen}\frac{K_r}{K_r+\rho}. \qquad (1.2.15)$$

From this it follows that the average (for one pulse) power of the generated output flux $\bar{S}_{gen} = E_{\text{gen}} / t_{gen}$ is given by the relation

$$\bar{S}_{gen} = \frac{\nu_{gen}}{\bar{\nu}_{pump}} \eta\chi (\bar{W}_{pump} - W^{th}_{pump}) \frac{K_r}{K_r + \rho}. \qquad (1.2.16)$$

As seen from relation (1.2.16), the power of the output flux is linearly dependent on the pumping power.

The experimental function $\bar{S}_{gen}\left(\bar{W}_{pump}\right)$ is close to the theoretical one. Using this function, we can measure the generation threshold W^{th}_{gen}. Measuring the slope $\bar{S}_{gen}\left(\bar{W}_{pump}\right)$ with respect to the axis $\bar{W}_{pump}$, we can estimate the harmful loss factor ρ that is greatly dependent on the tuning quality of a laser resonator

$$\operatorname{tg}\alpha = \frac{\nu_{gen}}{\bar{\nu}_{pump}} \eta\chi \frac{K_r}{K_r + \rho}. \qquad (1.2.17)$$

The efficiency of a laser is found from the ratio between the generation energy (1.2.15) and the electric energy $E_{pump} = CU^2 / 2$ supplied to the lamps during one pulse (C – capacitance of capacitor bank, U – voltage):

$$\Gamma = \frac{\nu_{gen}}{\bar{\nu}_{pump}} \eta\chi \left(1 - \frac{W^{th}_{pump}}{\bar{W}_{pump}}\right) \frac{t_{gen}}{t_{pump}} \frac{K_r}{K_r + \rho}. \qquad (1.2.18)$$

The ratio t_{gen} / t_{pump} takes into account that a part of the pumping energy is expended in the absence of geneartion. As follows from (1.2.18), the efficiency of a laser is growing with an increase in the pumping power. The ultimate efficiency is given by the following expression:

$$\Gamma = \frac{\nu_{gen}}{\bar{\nu}_{pump}} \eta\chi. \qquad (1.2.19)$$

For ordinary lamps and standard illuminators the ultimate efficiency of a neodymium laser, similar to that of an yttrium aluminum garnet laser, comes to a few percent. This value is by an order of magnitude greater than the efficiencies of lasers with the operation based on a three-level scheme when the lower laser level is the ground energy state.

At the same time, note that flash lamps as pumps are most often replaced by light-emitting diodes having much higher efficiency of the electric-to-optical power conversion (χ factor). Owing to LED pumping, the efficiency of lasers has been improved by more than an order coming to dozens of percent.

1.3. Active and passive Q-switched modes. Power, energy, and length of laser pulse. Modulation methods for resonators of solid-state lasers.

The cavity *Q*-switching method (giant-pulse generation method) was proposed in 1962 to improve the peak power of pulsed laser radiation (Mc-Clang and Hellworth). In conditions of free-running lasing a laser subjected to pulse pumping produces an irregular light pulse train, the total length of which is determined by the pump duration (normally on the order of several hundred microseconds). Owing to the cavity *Q*-switching, one is enabled to gain a single or giant pulse, a few dozens of nanoseconds in length and with a peak power from one to several decades of megawatt, rather than a series of low-power pulses. The quality (*Q*-)factor of a resonant system is understood as a ratio between the energy accumulated within the cavity and the energy lost per period of light oscillations ($Q = 2\pi W_{\text{acc}}/W_{\text{period}}$). For an optical cavity we have $Q = 2\pi\nu\tau_{ph}$, where ν is the radiation frequency, τ_{ph} – photon lifetime in the cavity. Control of the *Q*-factor is realized with the help of the control element introducing additional losses. According to the effect exerted on the intracavity control element, the *Q*-switching methods for an optical cavity may be subdivided into passive and active. In case of passive *Q*-switching an absorbing medium (dye solution or crystals with color centers) introduced into the cavity is bleaching as the intensity of the propagating radiation is increased. For active *Q*-switching of the optical cavity electro-optical and magneto-optical gates (Pockels cell, Kerr cell, Faraday cell) are most common. These gates are based on the effect of the varying optical anisotropy of a crystal under the effect of an electric or magnetic field. Owing to the control element positioned within the cavity, loss at the initial stage of pulsed pumping is increased without the development of the induced avalanche transitions. Pumping of an active laser medium takes place when the *Q*-factor of an optical cavity is low and the loss factor is in excess of the amplification factor. No lasing occurs, and a high degree of inversion is attained in the active medium for the generated pair of levels. Subsequently, the cavity *Q*-factor is drastically growing due to decrease in the insertion loss. As the amplification factor is considerably greater than the loss factor, the energy accumulated in the active element is emitted as a high-power light pulse during a period of ten nanoseconds.

A typical pattern of lasing in case of active *Q*-switching is given in Fig. 1.3.1 To realize a maximum power of lasing P_{las} on pulsed pumping (pump power P_{pump}), the *Q*-factor of the cavity (Fig. 1.3.1, *a*) is increased in time t_1 when the amplification factor is at maximum. In the process the amplification factor K_{ampl} (Fig.1.3.1, *b*) is considerably higher than the loss factor K_{loss} , offering the generation of a high-power laser pulse (Fig. 1.3.1, *c*).

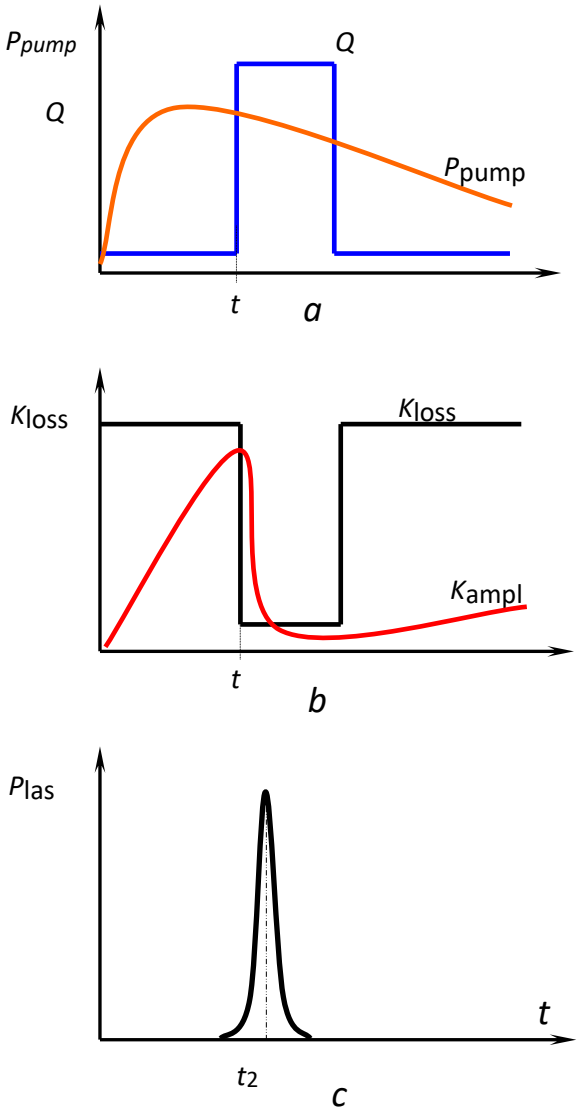


Fig. 1.3.1. Evolution of lasing for the laser pulse on active Q-switching.

Electro-optical Q-switching

The electro-optical *Q*-switching method for the optical cavity may be considered on the basis of a Pockels cell (Fig. 1.3.2). The cell comprises a crystal exhibiting the linear electro-optical effect: its refractive index is changing proportionally with the strength of the applied electric field. One can use electro-optical uniaxial crystals [ammonium dihydrophosphate – *ADP* ($NH_4H_2PO_4$), potassium dihydrophosphate – *KDP* (KH_2PO_4), deuterated potassium dihydrophosphate – *DKDP* (KD_2PO_4)] which, being subjected to the effect of an electric field, become biaxial. Most extensively used are *DKDP* crystals characterized by the electro-optical coefficient that is three-four times higher than that of other crystals, whereas their operating voltage is lower compared to other crystals.

Application of an electric field leads to deformation of the refractive index ellipsoid and to its rotation by 45°, the rotation angle being independent of the field magnitude. In the absence of the field, the refractive-index ellipsoid section

by the plane perpendicular to the optical axis (Z) is circular. At the same time, with the field applied, this cross-section is deformed as ellipse, and the principal dielectric axes make an angle of 45° with the initial ones.

Provided the light incident on the crystal is propagating along the axis Z and its polarization vector makes an angle of 45° with the set axes, the light wave may be subdivided into two waves with equal amplitudes and orthogonal polarization directions. Between these polarization components coming out of the crystal the phase difference is as follows:

$$\Phi = \frac{2\pi}{\lambda} n_0^3 r_{63} U , \qquad (1.3.1)$$

where λ – wavelength of light incident on the crystal; n_0 – refractive index for an ordinary wave; r_{63} – electro-optical coefficient; U – voltage applied to the crystal. Depending on the value of Φ, the light linearly polarized when entering the crystal is, in the general case, elliptically polarized at the exit. And, for $\Phi = \pi/2$, this light has circular polarization. Fig. 1.3.2 presents a schematic diagram of a laser used to perform this laboratory work. Its *Q*-switching, realized with the help of the gate including a polarizer and an electro-optical *DKDP* crystal, is due to the application of the quarter-wave voltage $U_{\lambda/4}$ associated with

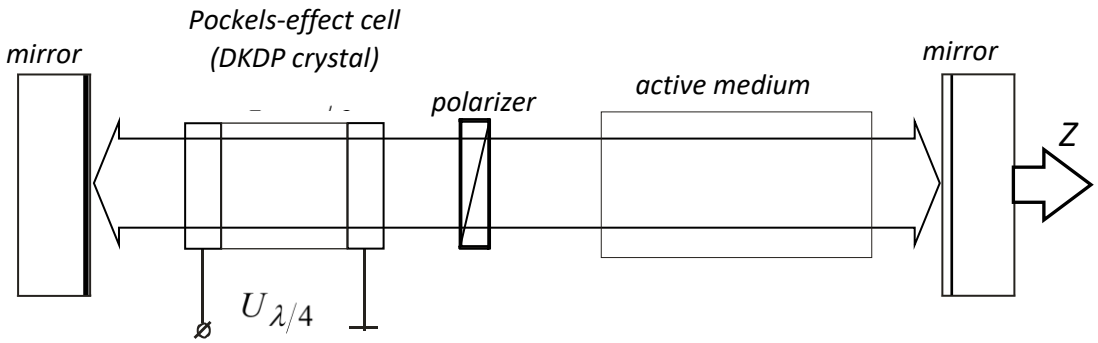


Fig. 1.3.2. Q-switching of the cavity with the use of a Pockels-effect cell.

the phase difference $\pi/2$.

The light transmitted from the active element through the polarizer acquires a circular polarization when leaving the crystal. Due to reflection from the totally transmitting mirror, the circularly polarized light is again transmitted through the electro-optical crystal. In the process the phase difference between the polarization components of a light wave goes equal to π due to rotation of the light polarization plane by 90°. Because of this, on its way back, radiation is stopped by the polarizer that prevents its return into the active laser medium.

Application of the quarter-wave voltage to the electro-optical gate results in its closing. To open the gate, this voltage should be released as in this case there is no birefringence and the incident light is transmitted without changes in its polarization. Owing to variations in the voltage across the electro-optical crystal, one is enabled to control the gate transmission and hence the *Q*-factor of the cavity.

Theoretical model for lasing in the active Q-switching mode

The active *Q*-switching method for an optical cavity may be considered theoretically in the approximation of instantaneous *Q*-switching. In this case losses of the cavity are changing over the period of time considerably less than the temporal evolution of lasing; at t<0 the loss is so great that the operation conditions of the laser are below the threshold, whereas at t>0 the loss is significantly decreased to facilitate the lasing process. When constructing a theoretical model, we take into account that the evolution time of a laser pulse ($\sim 10^{-7} \div 10^{-8}$ s) is much shorter than the relaxation time of the inverted population due to spontaneous and nonradiative transitions. Therefore, the effect of optical pump on changes in the inverted population during this period may be neglected. In this approximation, a system of equations describing the lasing evolution upon *Q*-switching may be represented as

$$\frac{\partial K_{\text{ampl}}}{\partial t} = -B_{32} U_{\text{las}} K_{\text{ampl}}, \tag{1.3.2}$$

$$\frac{\partial U_{\text{las}}}{\partial t} = c\left(K_{\text{ampl}} - K_{\text{loss}}\right) U_{\text{las}}, \tag{1.3.3}$$

where K_{ampl} – amplification factor of the active medium; $B_{32}U_{\text{las}}$ – probability induced transitions in the lasing channel; c – speed of light in the medium, K_{loss} – loss factor of the cavity.

It was taken into consideration that the amplification factor is decreased due to stimulated emission in the lasing channel, and lasing is developing when the amplification factor K_{ampl} is over the loss factor K_{loss}. It should be noted that elimination of the loss due to the intracavity gate results in K_{ampl} greatly exceeding K_{loss} and hence in rapid evolution of the lasing process. As this takes place, the amplification factor is decreasing. It follows from equation (1.3.3) that a maximum intensity of the laser pulse is attained at $K_{\text{ampl}} = K_{\text{loss}}$ (time t_2 in Fig. 1.3.1).

To find an analytical solution for a system of equations (1.3.2) and (1.3.3), we divide the second equation by the first one

$$\frac{\partial U_{\text{las}}}{\partial K_{\text{ampl}}} = \frac{c}{B_{32}}\left(\frac{K_{\text{loss}}}{K_{\text{ampl}}} - 1\right). \tag{1.3.4}$$

Then the volume energy density of the generated radiation in the cavity may be represented as

$$\begin{aligned} U_{\text{las}} &= \int_0^{U_{\text{las}}} dU_{\text{las}} = \frac{c}{B_{32}} \int_{K_{\text{ampl}}^0}^{K_{\text{ampl}}} \left(\frac{K_{\text{loss}}}{K_{\text{ampl}}} - 1\right) dK_{\text{ampl}} = \\ &= \frac{c}{B_{32}}\left(K_{\text{ampl}}^0 - K_{\text{ampl}} - K_{\text{loss}} \ln \frac{K_{\text{ampl}}^0}{K_{\text{ampl}}}\right) \quad . \end{aligned} \tag{1.3.5}$$

The generated radiation compensates for the loss: unfavorable losses (ρ) due to absorption and scattering from the medium inhomogeneities, mirrors; and favorable losses (K_r) associated with the radiation going out of the cavity through a semitransparent mirror. Taking into account the relationship between the radiation intensity and volume energy density $I = Uc$ and the total loss factor $K_{\text{loss}} = K_r + \rho$, the output power of the radiation flow from the cavity may be given by

$$P_{\text{las}} = U_{\text{las}} cS \frac{K_r}{K_{\text{loss}}}, \tag{1.3.6}$$

where S – beam area.

Lasing is at maximum when $K_{\text{ampl}} = K_{\text{loss}}$. Then, considering expressions (1.3.5) and (1.3.6), we have

$$P_{\text{las}}^{\max} = \frac{c^2 S K_r}{B_{32}}\left(\frac{K_{\text{ampl}}^0}{K_{\text{loss}}} - 1 - \ln \frac{K_{\text{ampl}}^0}{K_{\text{loss}}}\right). \tag{1.3.7}$$

In case the initial amplification factor is significantly greater than the loss factor ($K_{\text{ampl}}^0 >> K_{\text{loss}}$) from (1.3.7) it follows that

$$P_{\text{las}}^{\max} \approx \frac{c^2 S K_r K_{\text{ampl}}^0}{B_{32} K_{\text{loss}}}. \tag{1.3.8}$$

To find the energy of the generated pulse, it is necessary to determine the termination time of lasing. From equation (1.3.5) it is inferred that lasing is

terminated for the value of amplification factor $K_{\text{ampl}}^{\text{end}}$ that is associated with U_{las} = 0, i.e.

$$K_{\text{ampl}}^{0} = K_{\text{ampl}}^{end} + K_{\text{loss}} \ln \frac{K_{\text{ampl}}^{0}}{K_{\text{ampl}}^{\text{end}}}. \qquad (1.3.9)$$

Then the energy of lasing is as follows:

$$E_{\text{las}} = \int_{0}^{t_{end}} P_{\text{las}}(t)\,dt = \frac{cS\,K_r}{K_{\text{loss}}} \int_{0}^{t_{end}} U_{\text{las}}(t)\,dt. \qquad (1.3.10)$$

We arrive at the last integral by using equation (1.3.3), whose integration over the time interval from 0 to t_{end} gives

$$\int_{0}^{t_{\text{end}}} \frac{\partial U_{\text{las}}}{\partial t} dt = c \int_{0}^{t_{\text{end}}} K_{ampl} U_{\text{las}}\,dt - cK_{\text{loss}} \int_{0}^{t_{\text{end}}} U_{\text{las}}\,dt. \qquad (1.3.11)$$

Considering that $\int_{0}^{t_{\text{end}}} \frac{\partial U_{\text{las}}}{\partial t} dt = U_{\text{las}}(t_{\text{end}}) - U_{\text{las}}(0) = 0$ (there is no lasing both at $t = 0$ and $t = t_{end}$), with the use of equation (1.3.2) we can derive

$$\int_{0}^{t_{\text{end}}} U_{\text{las}}\,dt = \frac{1}{K_{\text{loss}}} \int_{0}^{t_{\text{end}}} K_{\text{ampl}} U_{\text{las}}\,dt = -\frac{1}{K_{\text{loss}} B_{32}} \int_{0}^{t_{\text{end}}} \frac{\partial K_{\text{ampl}}}{\partial t} dt = \frac{K^{0}{}_{ampl} - K^{\text{end}}{}_{\text{ampl}}}{K_{\text{loss}} B_{32}}. \qquad (1.3.12)$$

In this way from equations (1.3.10) and (1.3.12) we obtain the following expression for the lasing energy:

$$E_{\text{las}} = \frac{cS\,K_r}{K_{\text{loss}}^{2} B_{32}} \left(K_{\text{ampl}}^{0} - K_{\text{ampl}}^{\text{end}} \right). \qquad (1.3.13)$$

After the calculation of the lasing energy, using the maximum lasing power (1.3.7), we can write an approximate expression for the pulse length as

$$\Delta\tau = \frac{E_{\text{las}}}{P_{\text{las}}^{\max}} = \frac{K_{\text{ampl}}^{0} - K_{\text{ampl}}^{\text{end}}}{cK_{\text{loss}} \left(K_{\text{ampl}}^{0} - K_{\text{loss}} - K_{\text{loss}} \ln\left(K_{\text{ampl}}^{0} / K_{\text{loss}} \right) \right)}. \qquad (1.3.14)$$

For numerical analysis of the obtained results, it is expedient to introduce the inversion utilization factor $\eta = \left(K_{\text{ampl}}^{0} - K_{\text{ampl}}^{\text{end}} \right) / K_{\text{ampl}}^{0}$ and also the normalized amplification factor $x = K_{\text{ampl}}^{0} / K_{\text{loss}}$ that is actually responsible for the pump energy in excess of the threshold value (in case of the high-power pulse pump and excitation time that is below the relaxation time from the excited level). Then equations (1.3.9) and (1.3.14) may be rewritten as

$$\eta x = -\ln(1-\eta), \tag{1.3.15}$$

$$\Delta\tau = \tau_{\Phi}\frac{\eta x}{x-1-\ln x}, \tag{1.3.16}$$

where τ_{Φ} – lifetime of a photon within the cavity.

The solution for equation (1.3.15) is presented in Fig. 1.3.3, showing that for the ratio between the initial amplification factor and loss factor $x = K^{0}_{\text{ampl}} \big/ K_{\text{loss}} > 3$ the inversion utilization factor $\eta > 90$ %, the pulse length being $\Delta\tau \le 3\tau_{\phi}$.

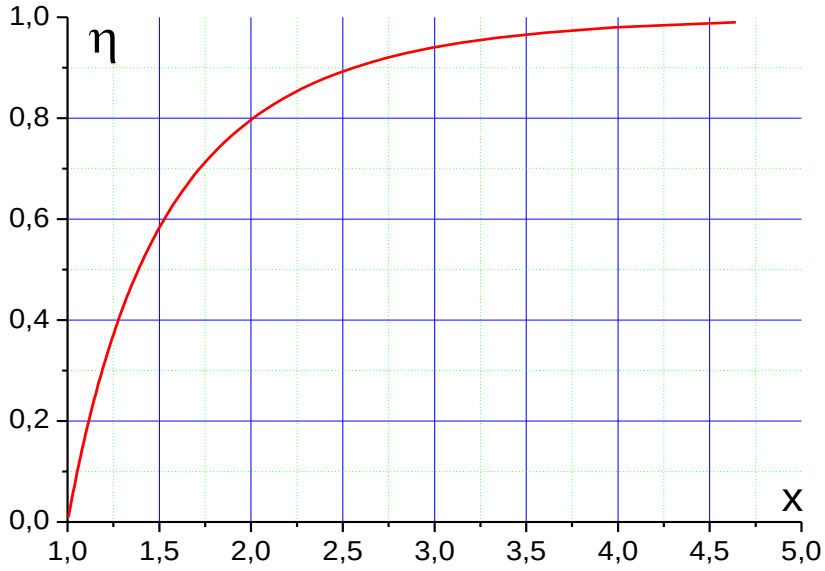


Fig. 1.3.3. The inversion utilization factor η as a function of the normalized amplification factor $x = K^{0}_{ampl} / K_{loss}$

When approaching the lasing threshold $K^{0}_{\text{ampl}} \to K_{\text{loss}}$, the pulse length is considerably increased.

Passive Q-switching of optical resonator

Passive Q-switching of optical resonator is associated with bleachable media whose transmission factor is varying under the effect of a luminous flux. Such media are represented by solutions of different dyes and crystals with color centers. The devices with bleachable media placed into laser resonator are called the passive laser shutters or Q-switches.

The principal requirements to the substance used for a passive Q-switch are as follows: great nonlinearity at the lasing frequency and photostability. To illustrate, in case of ruby lasers such properties are exhibited by the solutions of polymethine dyes, phthalo- and cryptocyanines. LiF crystals with color centers are widely used, along with polymethine dyes, in case of neodymium-doped yttrium aluminum garnet lasers as well as neodymium glass lasers. A typical schematic diagram of the energy levels in a dye molecule (crystal with color centers) is presented in Fig.1.3.4 together with the bleaching curve. Level S_1 is populated due to the effect of radiation with frequency $\nu_{S_0 \to S_1}$, and the metastable level T_1 is also populated at a rather high probability of the nonradiative transition $S_1 - T_1$.

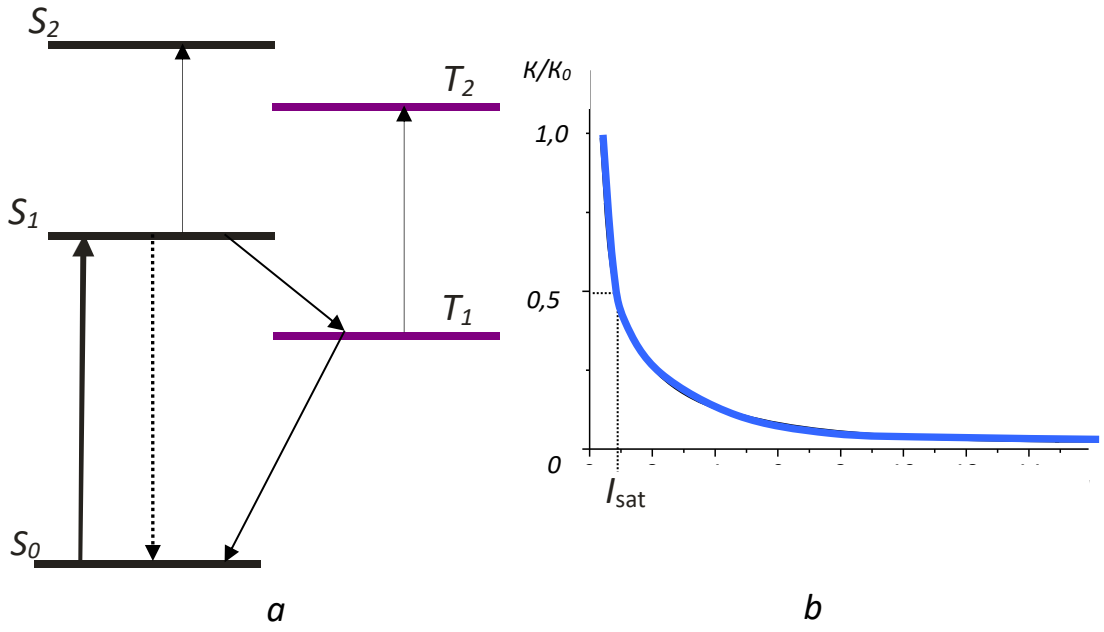


Fig. 1.3.4. Energy levels of a dye molecule (crystal with color centers) (a); absorption factor as a function of the light flux intensity (b)

As a result, the population of level S_0 is decreased and hence the medium absorption factor is decreased too. The absorption factor is halved for the intensity that is equal to the saturation intensity I_{sat}. Residual absorption at high radiation intensities $I >> I_{sat}$ is determined by absorption from levels S_1 and T_1 to higher energy levels, e.g., S_2 and T_2.

A decrease in the absorption factor of a passive *Q*-switch results in higher *Q*-factor of the cavity. As the amplification factor is considerably greater than the loss factor, the rise rate of laser radiation is increased leading to the formation of a giant pulse. The principles of laser operation in the passive *Q*-switching mode

may be considered in more detail for the case of switches based on *LiF* crystals with color centers, which are used in this laboratory work.

Passive Q-switches

In passive Q-switches the phenomenon associated with bleaching of some liquids (solutions of cryptocyanine, polymethine dyes, etc.) and solids under the effect of laser radiation is employed. Q-switches based on saturable absorbers represent a cell filled with a solution that absorbs light, the wave length of which is coincident with the laser radiation wave length. In a laser under study, where a cell is within the resonator, the generation is initiated only on condition that amplification of the active medium compensates for losses in the cell. Due to absorption, the critical population inversion in the cell is very high. When the intensity of laser radiation becomes comparable with the absorber saturation intensity, bleaching of the dye begins. This leads to a greater rate of the laser radiation intensity growth and hence to a greater rate of the dye bleaching. As the saturation intensity is relatively low, the inversion population in the active medium after bleaching remains very high. In this way, after the dye bleaching, laser amplification is much greater than the loss. As a result, a giant pulse is at the output. The generation of radiation is terminated. Molecules of the passive Q-switch relax to the lower energy level, and the loss is taking its initial value. When optical pumping continues, these stages in generation of a single pulse may be repeated. The characteristic durations of linear and nonlinear development of the generation come to ~1 μs and 10 – 100 ns, respectively. At the typical pulse energies 0.1-1 J the peak power of lasers with passive Q-switching is about 10^7 W.

In some cases passive Q-switches are advantageous over mechanical or electro-optical modulators. With their use, the device design is simplified considerably because there is no need for the modulation control units. Due to rapid bleaching of a medium, pulse lengths are minor. At the same time, bleaching of a passive Q-switch is rather random in character – there is no way to predict the instant of a single pulse generation. As the pump power greatly exceeds the threshold value, several light pulses are generated and this limits the peak power of generation.

Passive Q-switches based on $LiF:F_2^-$ ***crystals***

Crystals of lithium fluoride with color centers, featuring a good combination of spectral-luminescent and thermal properties, may be useful as a medium for the creation of passive *Q*-switches for neodymium lasers operating at 1.06 μm.

A color center is generally understood as any point defect of the lattice in a dielectric crystal: the presence of an impurity ion (atom) or absence of an ion (atom) at a particular site of the crystal lattice. Disturbance of the rigorous lattice periodicity due to the defect is responsible for the occurrence of local levels within the forbidden band of a crystal. The optical transitions involving these levels result in the additional absorption bands exhibited by a spectrum of the crystal. Because of this, a crystal, initially transparent in the visible, assumes a certain color after the introduction of point defects. In laser physics a color center is associated with a somewhat narrower class of defects due to the absence of anions at the lattice sites of an ionic crystal (anion vacancies). The optical properties of these specific color centers are very similar to the optical properties of organic dyes used for *Q*-switching of the optical cavity.

Color centers are most characteristic for alkyl-halide crystals with the M^+X^-- type structure which are composed of negative halide ions (X^-, anions), having closed shells, and positive metallic ions (M^+, cations). The crystal structure may be represented in the form of nested face-centered sublattices of cations and anions (Fig. 1.3.5).

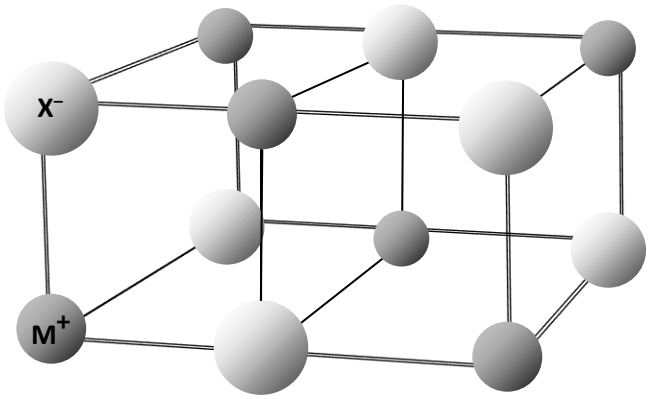


Fig. 1.3.5. Model of alkyl-halide crystals

The absence of a negatively charged ion in the neutral crystal lattice is equivalent to the positive charge distributed in the region of the anion vacancy, for which the electron capture is energy advantageous. The vacancy capturing an electron is termed as F center (from German word *Farbe* meaning color, pigment). An electron of F center, similar to that of an atom (ion), may be found both in the ground and excited state. A lower energy level of the color centers and some of the upper levels fall within the bandgap of the crystal. The transitions between the levels of F center are stimulated by optical radiation and accompanied by the light absorption or emission.

Fig. 1.3.6 shows the spectral-luminescent characteristics of F_2^- color centers in *LiF* crystal, which represent two adjacent anion vacancies capturing three electrons. A long-wavelength absorption band of F_2^- center in *LiF* crystal is homogeneously broadened and has two maxima at $\lambda = 960$ nm. A fairly long

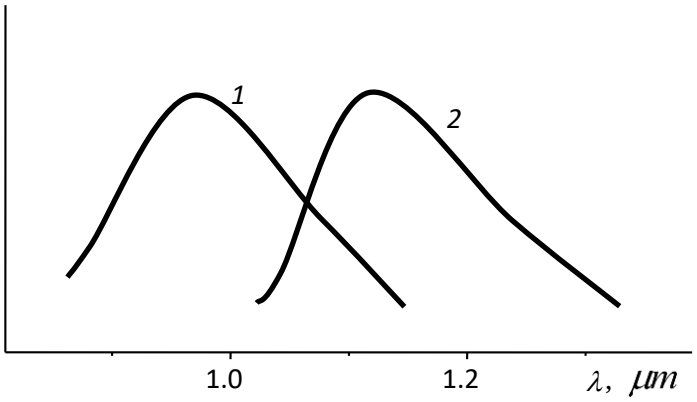


Fig. 1.3.6. Spectral-luminescent characteristics of F_2^- color centers in *LiF* crystal: *1* – absorption spectrum, *2* – emission spectrum

lifetime of the excited state (≈100 ns); large absorption cross-section, and good overlapping of the absorption band of F_2^- center in *LiF* crystal with the lasing wave length of a neodymium laser; very good bleaching of *LiF* crystals (ratio between transmission in the completely bleached state and the initial transmission ~ 20) enable one to use the components made of $LiF:F_2^-$ as passive *Q*-switches for neodymium ion lasers.

Features of lasing with passive Q-switch

Evolution of a single pulse begins at a time when the amplification factor, growing due to the pump, reaches a value of the loss factor (Fig. 1.3.7).

Even though the inversion is most pronounced, evolution of lasing is slow because of great losses introduced by a nonbleached crystal (linear stage of the pulse evolution). Spontaneously emitted photons initiate in the active medium new transitions, resultant in small avalanches of secondary photons. Some of

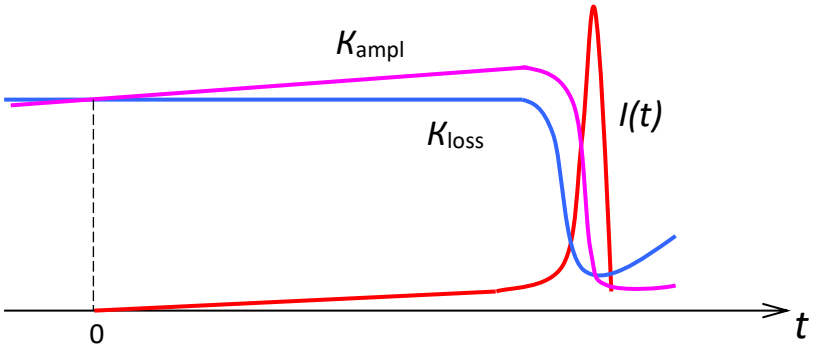


Fig. 1.3.7. Formation of a laser pulse in case of passive *Q*-switching

these photon avalanches are propagating along the cavity axis and partially absorbed by the F_2^- centers, activated to the excited energy levels, to cause a decrease in the absorption factor. Bleaching of the switch facilitates further development of photon avalanches to lower the absorption factor still further. As a result, a minimum value of the cavity loss factor is attained. Due to the amplification factor that is considerably greater than the loss factor, a high-power laser pulse is formed (nonlinear stage of lasing evolution). Practically all the energy is accumulated in the active medium of a laser in short time. As this takes place, the amplification factor decreases drastically to become lower than the loss factor, and lasing is terminated.

The characteristic duration of the linear and nonlinear stages of lasing evolution is $\sim 1\ \mu$s and 10–100 ns, respectively. In lasers with passive *Q*-switching a peak power about 10^7 W is achieved at typical pulse energies of 0.1–1 J.

When constructing a theoretical model for lasing with passive *Q*-switching, we take into account that time of the pulse evolution is much less than the inverted-population relaxation time, due to spontaneous and nonradiative transitions, and that the effect of optical pump on changes in the inverted population over this period of time can be neglected. For the specified approximations, a system of equations describing the lasing evolution after the threshold condition (amplification factor in excess of the loss factor) is reached may be represented as

$$\frac{\partial K_{\text{ampl}}}{\partial t} = -B_{32}U_{\text{las}}K_{\text{ampl}}, \tag{1.3.17}$$

$$\frac{\partial U_{\text{las}}}{\partial t} = c\left(K_{\text{ampl}} - K_{\text{loss}} - K_{\text{filt}}\frac{l_{\text{filt}}}{l} \right)U_{\text{las}}, \tag{1.3.18}$$

where K_{ampl} – amplification factor, $K_{\text{loss}} = K_r + \rho$ – total loss factor including favorable and unfavorable losses of the cavity, K_{filt} – absorption factor of a bleachable filter, l – length of an active medium, l_{filt} – thickness of a bleachable filter, $B_{32}U_{\text{las}}$ – probability of the stimulated transitions in the lasing channel, c – speed of light in a medium.

The loss factor due to a bleachable filter is dependent on the intensity and spectroscopic parameters of the medium used. In the general case it is necessary to take into account bleaching in the principal singlet channel $S_0 - S_1$ and

induced absorption from the excited levels S_1, T_1 (Fig. 1.3.4). By this approach, a solution for the equations describing the lasing process may be obtained only numerically. An analytical solution is possible in the approximation of a two-level model for the filter bleaching. Disregarding spontaneous relaxation processes, we can write the kinetic equation that describes a filter bleaching dynamics in the following form:

$$\frac{\partial K_{\text{filt}}}{\partial t} = -(B_{12}^{\text{filt}} + B_{21}^{\text{filt}})U_{\text{las}}K_{\text{filt}}. \qquad (1.3.19)$$

Using a system of equation (1.3.17) – (1.3.19), we can determine the volume energy density within the cavity as follows:

$$U_{\text{las}} = \frac{c}{B_{32}}\left(K_{\text{ampl}}^0 - K_{\text{ampl}} - K_{\text{loss}} \ln\frac{K_{\text{ampl}}^0}{K_{\text{ampl}}} - \frac{B_{32}}{B_{12}^{\text{filt}} + B_{21}^{\text{filt}}} K_{\phi}\frac{l_{\text{filt}}}{l} \right) . \qquad (1.3.20)$$

The last term in expression (1.3.20) gives the radiation loss on absorption by a passive Q-switch. Denoting the energy absorbed in the bleachable filter as E_{abs}, we can write the following expression for the energy of a generated pulse:

$$E_{\text{las}} = \frac{cS\,K_r}{K_{\text{loss}}^2 B_{32}}\left(K_{\text{ampl}}^0 - K_{\text{ampl}}^{\text{end}}\right) - \frac{K_r}{K_{\text{loss}}}E_{\text{abs}}, \qquad (1.3.21)$$

where S – beam area, K_{ampl}^0 – initial amplification factor (initial step of the pulse evolution), $K_{\text{ampl}}^{\text{end}}$ – amplification factor at the end of pulse. And the initial value of the amplification factor is depending on the density of a bleachable filter. By equation (1.3.18) we can obtain

$$K_{\text{ampl}}^0 = K_{\text{loss}} + K_{\text{filt}}^0\frac{l_{\text{filt}}}{l}, \qquad (1.3.22)$$

where K_{filt}^0 – initial absorption factor of a bleachable filter.

From the above equations it follows that the pulse energy is dependent on the initial transmission of a passive Q-switch. An increase in the initial absorption factor K_{filt}^0 results in growing of the initial amplification factor K_{ampl}^0 and energy of the generated pulse. In this case one should keep in mind that increasing optical

density of the switch leads to a higher threshold of lasing. When the pump energy is considerably higher than the threshold one, during the pump period the generation of the second and subsequent pulses is possible too. Origination of these pulses is due to relaxation of a passive Q-switch after the generation of each pulse. Provided the relaxation rate of a passive Q-switch is below the pump rate, generation of the second pulse may occur at partial bleaching of the filter. In this case the second pulse has a lower energy than the previous one. Owing to the use of fast-relaxation passive Q-switches, there is a possibility to generate pulses of practically equal energy.

1.4. Generation of mode-locked picosecond pulses

For generation of picosecond pulses the mode-locking regime is used. Generation of ultrashort pulses is due to the fact that a great number of longitudinal cavity modes may simultaneously be excited in a laser medium with a relatively broad amplification band, $\Delta\nu_{las.}$ These longitudinal modes are distinguished by the field distribution along the cavity axis and noted for the frequency division by $\delta\nu=c/2L$, where $L=\sum_{i=1}^{n} l_i n_i$ – optical path between the cavity mirrors having regard to the refractive index of each optical intracavity element n_i with the length l_i; c – speed of light (Fig. 1.4.1, *a*, *b*).

Mode-locking is a completely ordered regime of laser operation when the phase ratio between the longitudinal modes is constant. In this case the cavity modes interfere with each other, and laser radiation takes the form of a pulse train with the light pulses equally spaced in time. The point of mode-locking is that at particular instants of time all modes contribute maximally to the total intensity of an intracavity electromagnetic field. And the time interval between two neighboring maxima may be determined from the expression

$$T_{\text{cav}} = 1/\delta\nu = 2L/c \qquad (1.4.1)$$

that gives a time of the complete cavity round-trip. In this way a single pulse formed in the cavity is propagating both in the forward and backward directions.

As the phase difference established between the cavity modes is fixed, one can estimate a minimum length possible for the ultrashort laser pulse τ .This length

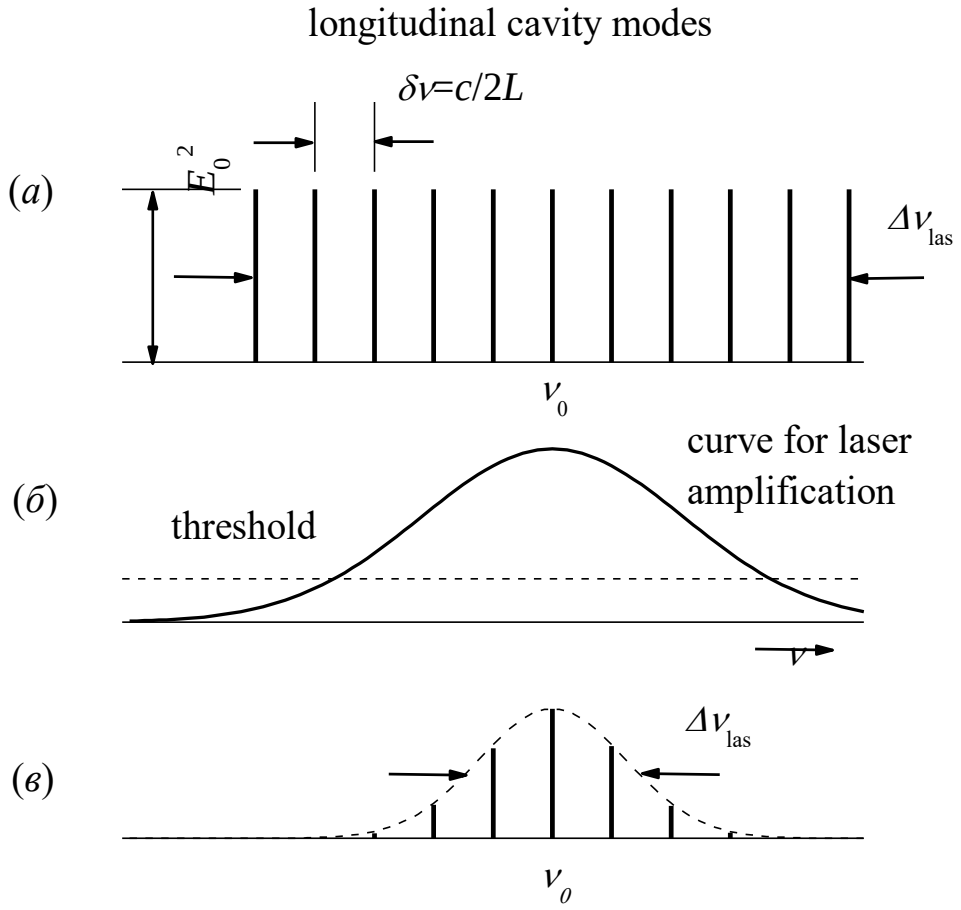


Fig. 1.4.1. Spectrum for longitudinal cavity modes (a), amplification line form (b), and lasing spectrum (c). Lasing occurs at longitudinal modes the amplification of which is over the threshold

is related to the width of a lasing spectrum in the mode-locking regime $\Delta\nu_{las}$ by

$$\tau\Delta\nu_{las} \geq C_B, \qquad (1.4.2)$$

where C_B – numerical factor in the order of unity, its value being determined by a specific form of the spectral intensity distribution for an ultrashort pulse (quantities τ and $\Delta\nu_{las}$ being determined at a half-maximum of the amplitude for the corresponding function). When a pulse is Gaussian in its spectral form, we have $C_B = 0.44$.

The passive mode-locking regime of a laser is achieved due to the use of an intracavity passive Q-switch made of the material whose absorption at the wavelength of lasing is decreased (saturated) under the effect of high-power illumination. The process of passive mode-locking in solid-sate lasers subjected

to pulsed pumping is most conveniently described by the "fluctuation model" that is based on a dynamic description of lasing. The model is developed on the assumption that in a cavity, near the lasing threshold, the initial noise fluctuations are developed, the intensity distribution of which is random being determined by a character of spontaneous emission. Because of the absorption saturation effect, the greatest fluctuation spike in the passive Q-switch is selected, whereas the others are suppressed.

Such a laser operating in the regime of simultaneous mode-locking and *Q*-switching emits a giant pulse, i.e., a train of several ultrashort pulses (Fig. 1.4.2). To select a single fluctuation spike, a relaxation time of the passive Q-switch τ_F must be considerably lower than time T_{cav}. Besides, time τ_F should be as short as possible since it limits a minimal length of ultrashort pulses attainable with this laser.

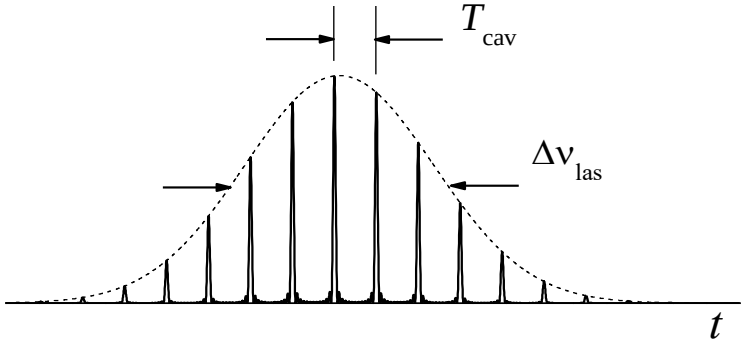


Fig. 1.4.2. Train of ultrashort pulses emitted by a solid-state laser on passive mode-locking

The operation of a laser with a passive gate may be described by a system of rate equations for the intracavity energy density U_{las} with the use of the active-medium amplification factor K_{amp} and of the passive-gate absorption factor K_F:

$$\frac{\partial U_{\text{las}}}{\partial t} + c\frac{\partial U_{\text{las}}}{\partial z} = c\left[K_{\text{amp}} - K_{\text{loss}} - K_F\, l_F / l\right]U_{\text{las}}, \tag{1.4.3}$$

$$\frac{\partial K_{\text{amp}}}{\partial t} = -B_{32}U_{\text{las}}K_{\text{amp}}, \tag{1.4.4}$$

$$\frac{\partial K_F}{\partial t} = -(B_{12}^F + B_{21}^F)U_{\text{las}}K_F\frac{S}{S_F} + \frac{K_{F0} - K_F}{\tau_F}, \tag{1.4.5}$$

where l and l_F –active element and passive gate lengths, respectively; K_{loss} – loss factor due to the transmission of mirrors and due to scattering from optical inhomogeneities of the cavity elements; S, S_F – laser beam areas in the active element and passive gate, respectively; their ratio determines a degree of the radiation focusing into the passive gate; K_{F0} – initial (nonsaturated) absorption factor of the passive gate. Here it is assumed that in the passive gate there is no absorption from the excited states and hence it may be described by a two-level scheme of the energy states (stimulated transitions B_{12}^F+B_{21}^F). But lasing in the active medium functioning according to a four-level scheme (1 - 4 – absorption channel, 3 - 2 – lasing channel) is determined by the stimulated transitions B_{32} (Fig.1.4.3).

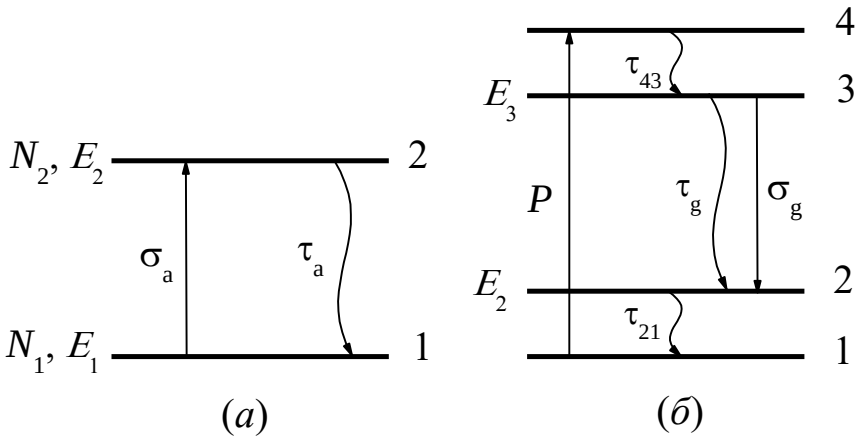


Fig 1.4.3 (*а*) – Two-level model for the passive Q-switch interacting with radiation at the frequency $\nu=(E_2-E_1)/h$. (*б*) – Schematic of energy levels for the four-level active medium [$\tau_{43}<<\tau_a$, $\tau_{21}<<\tau_a$, $\nu_{ген}=(E_3-E_2)/h$]

An analysis of the mode-locking regime on the basis of a “fluctuation model” using a system of equations (1.4.3)–(1.4.5) demonstrates that, along with the classical lasing threshold associated with the initiation of lasing (equality condition for amplification and total loss in the cavity), the “second lasing threshold“ is found, in excess of which the mode-locking regime is realized. The condition for the “second threshold”, derived in the approximation of a rapidly relaxed absorber when the passive-gate relaxation time τ_F is significantly below a length of the typical fluctuation spike, takes the form

$$c\left(K_{\text{amp0}} - K_{\text{loss}} - K_F\, l_F / l\right)\frac{K_{\text{amp0}} l}{K_{F0} l_F}\,\frac{\tau_F S}{S_F}\,\frac{B_{12}^F + B_{21}^F}{B_{32}}\,\frac{U_{\text{las}}^{\max}}{\langle U_{\text{las}}\rangle} > X_Q, \qquad (1.4.6)$$

where $U_{\text{las}}^{\max}$ – energy density of lasing in the fluctuation spike with the greatest intensity, $\langle U_{\text{las}}\rangle$ – average energy density in the cavity, X_Q – numerical factor in the order of unity that is determined experimentally. The ratio $U_{\text{las}}^{\max}/\langle U_{\text{las}}\rangle$ in the "second threshold" condition is caused by the fact that saturation of the passive gate is predominantly due to the effect of a fluctuation spike with a maximum intensity, whereas depletion of the population inversion in the active medium is governed by the average energy density in the cavity. The above-mentioned ratio is varying from flash to flash of the lamp within the limits 3 to 10.

As follows from equation (1.4.6), the "second lasing threshold" condition for the given active element and passive gate may be satisfied by changes in a degree of the radiation focusing into the gate (ratio S/S_F), that is realized with a focusing lens positioned in the cavity.

As distinct from *Q*-switching, excess of the "second threshold" (1.4.6) in the case of mode-locking must be the least possible. The physics of this situation is as follows: in the process of laser operation near the second threshold intensity of the greatest fluctuation spikes is associated with the discrimination properties of the passive gate, most marked after a considerable saturation of the active medium. A relatively minor depletion of the population inversion in the active medium combined with the absorption saturation in the passive gate results in a more effective selection of the fluctuation spike with a maximum amplitude. An effective approach to laser operation with passive mode-locking based on pulsed pumping near the second threshold is by variation of the focusing parameter S/S_F .

Passive Q-switches based on glasses with nanodimensional PbS particles

One can use glasses with the dispersed particles of semiconductor compounds as saturable absorbers in lasers generating ultrashort pulses. The form of semiconductor particles in glass is usually close to a sphere. Their radius measures from a few to tens of nanometers at a standard size spread of 5–10%. In the literature such particles are referred to as nanoparticles. The glass matrix in saturable absorbers of this type is optically inert, whereas semiconductor nanoparticles represent active centers. A spectral position of the absorption edge in similar materials is determined by the chemical composition of a semiconductor compound and by the particle size.

Nanoparticles introduced into the glass matrix belong to the so-called quasi-zero-dimensional structures, where the charge carriers (electrons, holes) and an "electron-hole molecule" (exciton) are spatially-limited by the particle bulk in three dimensions. To describe their optical properties, we use the method based on consideration of the properties exhibited by quantum-mechanical particles in a potential box, whose form corresponds to the nanoparticle form and height – to the potential barrier at the interface "matrix-nanoparticle". The spatial limitation of motion for quantum-mechanical particles, that is due to small sizes of nanoparticles at $R_p \leq a_B$ (a_B – Bohr radius of an exciton), leads to a line character of their energy spectrum, an interval between the formed levels being dependent on the particle size (Fig. 1.4.4) and to the specific selection rules for optical transitions. All this is revealed by a monotonic shifting of the absorption edge into the region of shorter wave lengths, with a decreased size of the particles, and by the appearance of line bands in an absorption spectrum (Fig. 1.4.5). Within the scope of the simplest model, a short-wavelength shift of the (primary) absorption band with a lower energy is proportional to $(1/R_p)^2$.

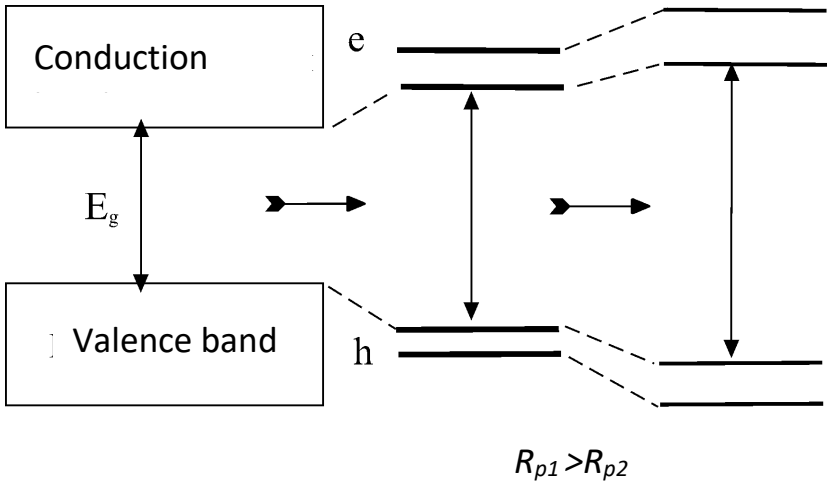


Fig. 1.4.4. Changes in the energy structure of a semiconductor with a decrease of its spatial dimensions R down to the level revealing the quantum effects due to the limited motion of charge carriers ($R_p \leq a_B$)

Being a semiconductor compound, lead sulfide PbS is characterized by a narrow forbidden band (E_g=0.4 эВ) and by the visible and near infra-red radiation absorption. However, because of the quantum-dimensional effects in

nanoparticles of PbS, at $R_p \leq a_B$ its absorption edge may be shifted to the wave lengths shorter than 1 □m. As the Bohr radius of an exciton in PbS is sufficiently large ($a_B \approx 18$ nm), its nanoparticles reveal the characteristic absorption bands at $R_p \leq 10$ nm (Fig. 1.4.5).

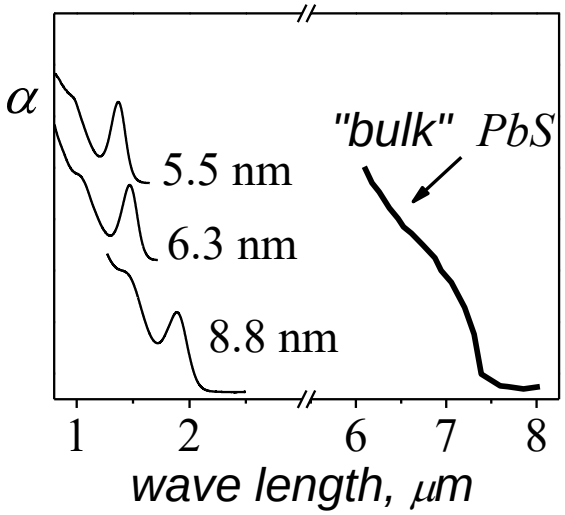


Fig. 1.4.5 Absorption spectrum for the “bulk” semiconductor and semiconductor particles of lead sulfide PbS with different sizes $R_p < a_B$ in glass

Glasses with PbS nanoparticles have a considerable width of a bleaching spectrum determined by the primary absorption-band width that may be in excess of 100 nm. For nanoparticles with the size R≈2 nm the primary absorption band is found in the region of 1 □m. Such materials may be used as passive switches for neodymium-doped lasers. The bleaching relaxation times of nanoparticles with such sizes come to tens and hundreds of picoseconds, enabling laser operation in the passive mode-locking regime.

1.5. Methods of radiation frequency tuning. Tunable lasers.

The term «tunable laser» applies to a laser with the wavelength varying over the spectral range, the width of which is much greater than that of its radiation line. Such a possibility exists not for all lasers. Tunable lasers may be of the solid-state, liquid, fiber, semiconductor, hybrid or some other types.

Most common among the currently available tunable laser systems are *dye lasers*. Using different solutions, one can generate ultraviolet (UV), visible, and infra-red (IR) radiation with the wavelengths from 0.3 to 1.5 μm. The tuning range

is approximately at a level of $\Delta\nu/\nu = 5 \ldots 15$ %. Dye lasers may be pumped by flash lamps. However, a higher quality of radiation is attainable when solid-state or gas lasers are used for pumping and hence we have the combined-type device.

Color-center lasers generating in the near infra-red region (up to 3 μm) are designed in a similar way. In this case laser media are represented by crystals of alkaline halogenides with different impurity centers. As some dyes and color centers are insufficiently stable laser media, solid-state lasers using oxide and fluoride crystals doped with ions of different metals have been developed. One of the well known lasers of this type is a Ti-sapphire ($Ti:Al_2O_3$) laser featuring a wide tuning range, from 700 to 1050 nm, and a higher efficiency.

Excimer lasers are used as tunable sources for the UV spectral region but their tuning range is relatively small – only 1 %. To obtain a wider tuning range in the UV region, the second harmonic generation of tunable visible lasers is used.

For the mid- and far-field IR region it is recommended to use molecular lasers (e.g., CO_2 – lasers) which offer tuning from line to line in the intermittent mode.

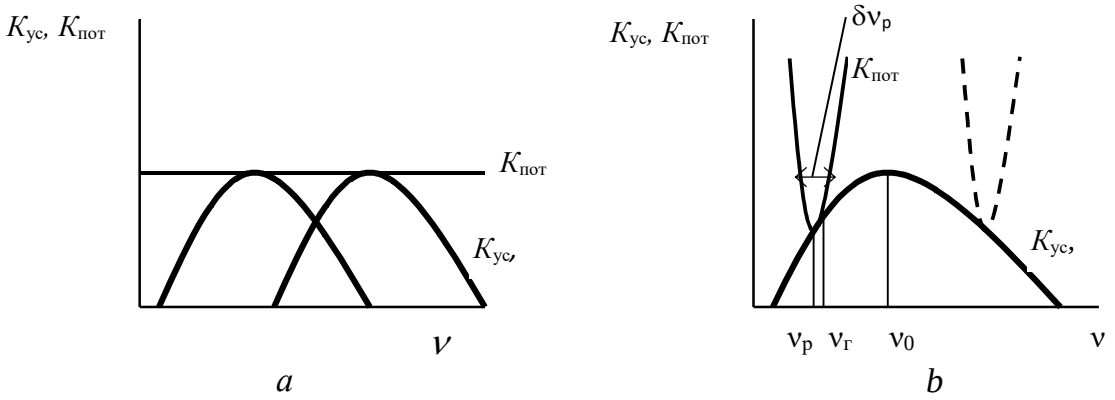
Semiconductor lasers are tunable by variations in the pump current or in temperature, their tuning range being from 0.1 to 1 %. Owing to laser diodes made of different materials or systems with impurity crystals, the range from 0.38 to 30 μm is covered.

At the present time parametric optical generators pumped by lasers with a fixed wavelength often form an alternative for tunable lasers. This is associated with a progress in nonlinear optics and with achievements in the development of highly efficient nonlinear crystals. The radiation wavelength tuning of such a generator is realized by variations in orientation or temperature of a nonlinear crystal.

1.5.1. Principles of the generated-radiation frequency tuning

By the present time, numerous methods to control the spectral parameters of laser radiation, differing both in the initial principles and techniques of their practical implementation in lasers of different types, have been developed. All these types may be subdivided into two groups:

1. Transformation of laser radiation with the help of the methods of nonlinear optics, e.g. second harmonic generation.



$\delta\nu_p$ – transmission band of dispersive resonator,
ν_0 – frequency associated with a maximum of the amplification line,
ν_p – dispersive resonator tuning frequency,
$\nu_г$ – oscillation frequency

Fig. 1.5.1 – Threshold generation conditions in nonselective (*a*) and dispersive (*b*) resonators

2. Control over the formation conditions of laser radiation due to selection or variation of the spectral characteristcis of an active medium or of a resonator. The generation condition $K_{ус} \geq K_{пот}$ (Fig. 1.5.1) will make this method clear.

The amplification line for the isolated optical transition is bell-shaped. In a laser with a nonselective resonator (resonator losses within the limits of the amplification line are independent of the frequency), generation occurs at the frequency associated with a maximum of the amplification line (Fig. 1.5.1 *a*). Tuning of radiation frequency is realized with offset of the amplification line along the frequency axis. When a resonator includes dispersive elements (prisms, diffraction gratings, Fabry-Perot interferometer), the curve for losses depending on the frequency $K_{пот}(\nu)$ is, as a rule, more abrupt than the curve for $K_{ус}(\nu)$.

A resonator of this type is referred to as dispersive (Fig. 1.5.1 *b*). The frequency $\nu_г$ at which the generation is initiated corresponds to the point of tangency of the curves $K_{ус}(\nu)$ and $K_{пот}(\nu)$. Spectral tuning of laser radiation is performed by offset of the selective loss curve with respect to the amplification line.

The relationship between oscillation frequency and tuning frequency of a resonator is called the tuning characteristic of a laser. When using a highly selective resonator ($\delta\nu_p \to 0$), the tuning characteristic is a straight line (oscillation frequency is coincident with the tuning frequency of a resonator). With due regard for a width of the selective loss profile, the oscilation frequency is shifted to the amplification line maximum relative to the tuning frequency of a

resonator $|\nu_г - \nu_0| < |\nu_p - \nu_0|$, i.e., the oscillation frequency pulling takes place (see Fig. 1.5.1 *b*). The oscillation-frequency pulling effect increases as resonator selectivity degrades (as $\delta\nu_p$ increases) and as the distance of the resonator tuning frequency from the amplification line maximum increases. In general, the tuning characteristic is nonlinear.

1.5.2.Grating resonators

Numerous variants of dispersive grating resonators may be subdivided into three main types:

- with two end mirrors and a reflective grating between them (Fig. 1.5.2 *a*);
- with an end mirror and a grating in the autocollimation position that combines the functions of a dispersive element and a resonator end mirror (Fig. 1.5.2 *b*);
- with two end mirrors and a transmission grating between them (Fig. 1.5.2 *c*).

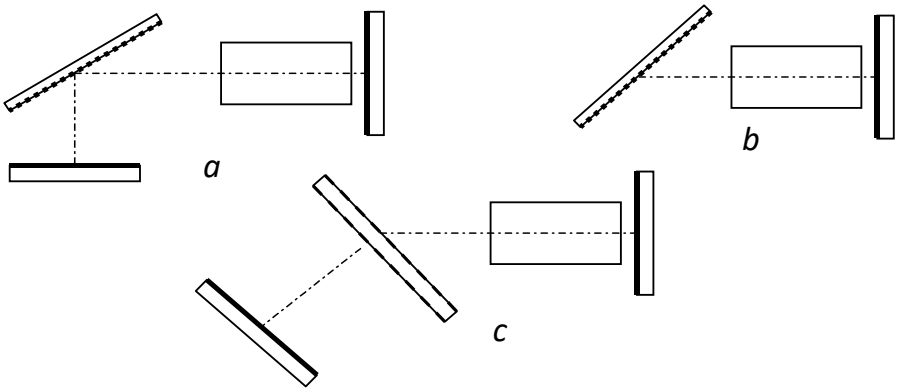


Fig. 1.5.2 – Main types of dispersive grating resonators

The radiation frequency is tuned by rotation of the mirror or of the grating. Radiation is output from resonator through the end mirror or with the use of zero-order diffraction of the grating.

The propagation direction of diffracted radiation for the given incidence angle is determined by the grating period Λ and the wavelength λ as follows:

$$\Lambda(\sin\alpha + \sin\beta) = m\lambda, \qquad (1.5.1)$$

where α – incidence angle; β – diffraction angle; m – diffraction order.

The angular dispersion of the grating is given by

$$Д = \frac{d\beta}{d\lambda} = \frac{m}{\Lambda \cos\beta} = \frac{\sin\alpha + \sin\beta}{\lambda \cos\beta}. \qquad (1.5.2)$$

In the autocollimation scheme, when the diffracted beam is coincident with the incident beam ($\alpha = \beta$), we have

$$Д = \frac{d\beta}{d\lambda} = \frac{2tg\alpha}{\lambda}. \qquad (1.5.3)$$

The generation line width of a grating laser is determined by the dispersion $Д$ and by the beam divergence $\Delta\beta$ as

$$\Delta\lambda = \Delta\beta / Д. \qquad (1.5.4)$$

A narrow generation line is ordinary attained with the use of high-dispersion gratings. As follows from equation (1.5.2), dispersion of the grating is the higher the lower is its period. But lowering of the period is restricted by the condition $\Lambda \geq \lambda/2$ as diffraction is impossible for lower diffraction period. The number of possible diffraction orders is growing with the grating period. For laser gratings of particular importance is the region with only two operating orders: zero and first. Because of this, the grating introduces minimal losses. However, it should be noted that a width of the generation line can differ from the line calculated by formula (1.5.4) due to such factors as nonstationary generation (in the process of generation the line width is narrowed on multiple round-trips of resonator), spatially-inhomogeneous amplification saturation, thermooptic distortions.

On the other hand, the selective ability of a resonator may be improved by lowering of radiation divergence and by the increased angular dispersion of a resonator. Telescopic systems (telescopes) are most successfully used to influence the radiation divergence or the resonator effective dispersion.

1.5.3.Basic characteristics of selective prism resonators

Resonators with prism dispersion elements are widely used due to the following properties: availability, simple design, versatility, multitude of the appropriate optical materials. Besides, ordinary the loss introduced by a prism with Brewster-angle orientation of the faces into a resonator is not higher than 2–3 % that is considerably lower than the loss for dispersive elements of other types. Using different optical materials, one can change the refractive index dispersion over a rather wide interval. Specifically, in the case of glass in the visible spectral region (λ= 600 nm) we have $dn/d\lambda \approx 0.03 - 1$ mrad/nm.

Dispersive resonators may be constructed with prisms of all types known in optics. Most common are symmetric prisms (Fig.1.5.3 *a*) offering both high angular dispersion and low level of losses. When a low angular dispersion of resonator is appropriate, it seems expedient to use an autocollimation prism (Littrow prism). Such a resonator is compact (Fig.1.5.3 *b*), its tuning is effected

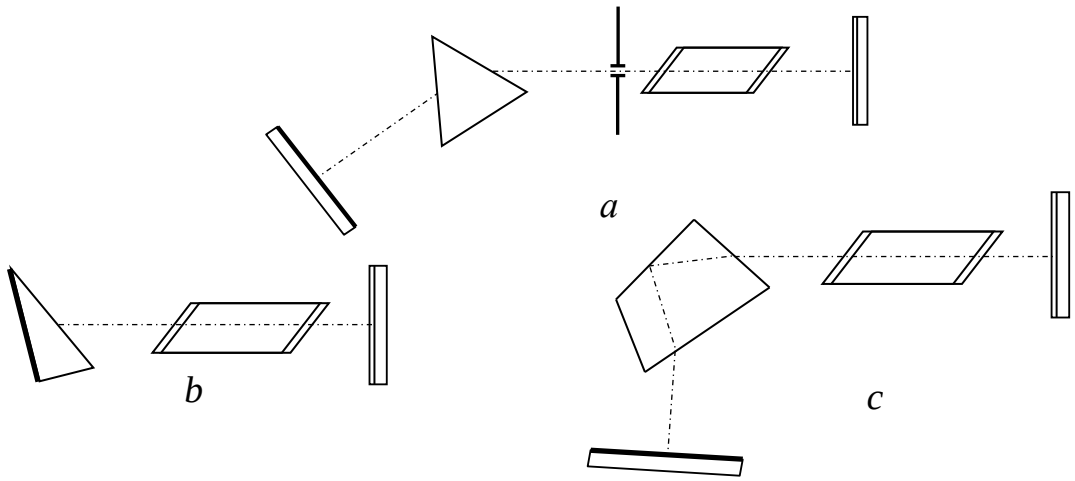


Fig. 1.5.3 – Dispersive resonators with a symmetric prism (*a*), with Littrow (*b*) and Abbe (*c*) prisms

by rotation of the prism itself, whereas the optical axis orientation is fixed. The optical axis position remains invariable in resonators with constant deflection prisms (Fig.1.5.3 *c*). These prisms (e.g., Abbe prism) are suitable for the construction of circular dispersive resonators.

Let us consider the simplest scheme of a resonator (Fig.1.5.3 *a*). The angular dispersion of this resonator is given by the following expression:

$$Д_{\text{р}} = d\alpha_1/d\lambda + d\alpha_2/d\lambda, \qquad (1.5.5)$$

where $d\alpha_1/d\lambda = (dn/d\lambda)\sin A/\cos\alpha_1 \cdot \cos\beta_2$ – angular dispersion of the prism for a light beam incident on the prism from the left (Fig.1.5.4);

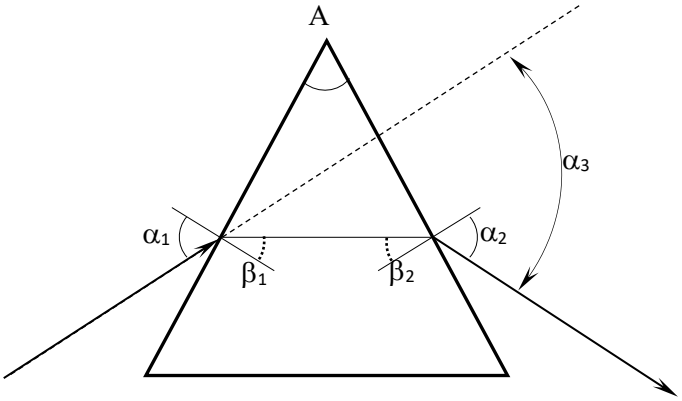


Fig. 1.5.4 – Geometry of the beam path
through a symmetric prism

$d\alpha_2/d\lambda = (dn/d\lambda)\sin A/\cos\alpha_2 \cdot \cos\beta_1$ – angular dispersion of the prism for a beam incident from the right.

In the position of minimal deflection ($\alpha_1 = \alpha_2 = \alpha$, $\beta_1 = \beta_2 = \beta$, $A = 2\beta$), formula (1.5.5) is transformed as

$$Д_р = 4n^{-1}(dn/d\lambda)\mathrm{tg}\alpha. \qquad (1.5.6)$$

If the refraction angle A of a prism meets the Brewster condition for the beam incident on the prism $(\mathrm{tg}\alpha_Б = n)$, the formula for resonator dispersion is further simplified as

$$Д_р = 4dn/d\lambda. \qquad (1.5.7)$$

It should noted that for a resonator with a symmetric prism (Fig.1.5.3 *a*) the rotation angle of the tunable (left) mirror, that is required for the transition from one wavelength to another, is specified by angular dispersion of the prism, i.e. by a half value of the resonator dispersion.

Tuning of selective resonators is a fairly complex problem, especially in the case of dispersive elements operating on transmission, when the optical axis shows a kink and changes its position with the wavelength tuning. We consider some tuning procedures for such resonators taking a resonator with a symmetric prism as an example. The resonator optical-axis position at the central wavelength is ordinary set using the generation in a nonselective resonator comprising mirror *1* and auxiliary (secondary) mirror *3* (Fig. 1.5.5,*a*). Radiation beam passes through the prism (or collection of prisms) that is set in the position of minimal deflection.

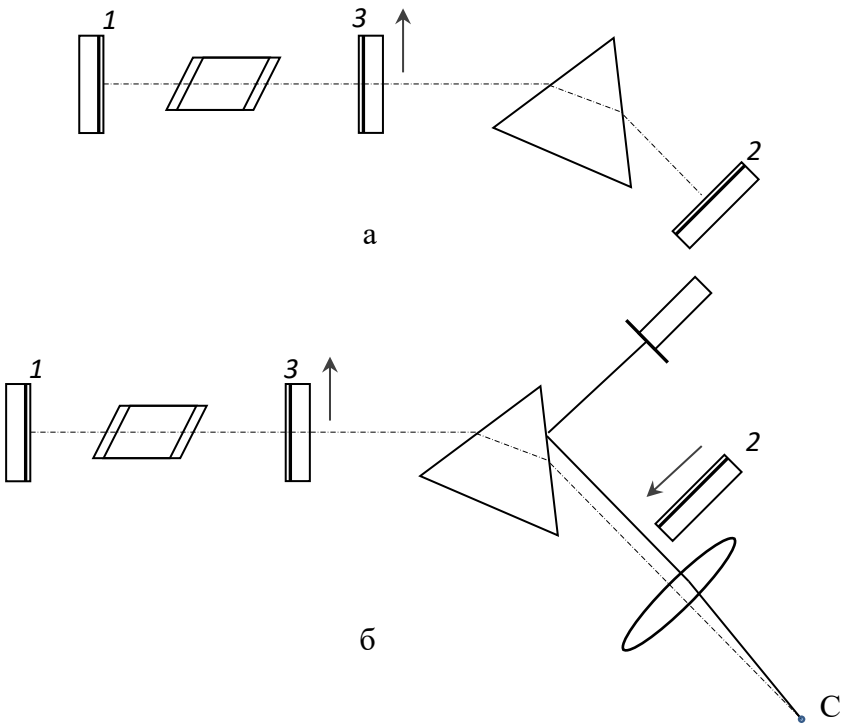


Fig. 1.5.5 – Schematic for tuning of resonators with prisms

Next end mirror *2* is tuned with the help of a diaphragm located behind mirror *3*, and then the auxiliary element is removed. The final tuning is realized by adjustment of one of the mirrors until the minimal generation threshold is reached. Note that tuning of a dispersive resonator in dye lasers may be attained without auxiliary elements due to the use of superluminescence radiation propagating along the axis of generation.

The tuning process is simplified considerably with the use of a continuous-wave laser operating in the visible. In this case a lens is positioned at the path of the beam generated by laser with nonselective resonator *1-3* (Fig.1.5.5, *b*) to bring it to the focus C. The spatial position of the focal point is recorded, for example, with the use of photographic paper. Then a beam of auxiliary laser *4* is guided so that, after its reflection from the exit face of the prism and passing through the lens, it can be also focused at the point C. The lens and mirror *3* are removed and the auxiliary beam, that is now parallel to the beam of the primary laser, is used for tuning of mirror *2*.

1.5.4. Tunable distributed-feedback lasers

Physical principles of lasing in distributed feedback lasers

As known, one of the prerequisites for generation is positive feedback realized by means of an external resonator formed by two (or more) mirrors or some other reflecting elements. Besides, there are devices without external elements of feedback and the generation is realized by means of the structures with periodic spatial inhomogeneities somehow formed in the active medium itself. In this case a positive feedback required for the generation initiation is provided due to Bragg scattering of light waves from the above-mentioned inhomogeneities and hence feedback is distributed over the active medium volume. Because of this, the devices of this type are known as distributed feedback lasers. Due to a highly selective character of Bragg reflection, generation in distributed feedback lasers is excited in a narrow spectral range, its wavelength being determined by Bragg condition as

$$\lambda_z = \frac{2dn}{m}, \tag{1.5.8}$$

where d – spatial modulation period; n – refractive index of the active medium; m – integer number giving the order of Bragg diffraction.

The optical parameters associated with modulation leading to a positive feedback are the amplification factor and the refractive index of an active medium. As applied to dyes, the main techniques used to form distributed

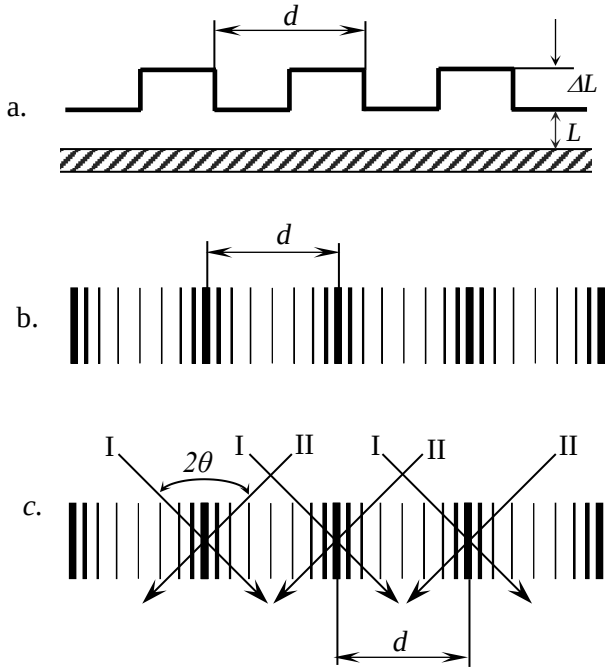


Fig. 1.5.6 – Main techniques to form distributed feedback: spatial modulation of the optical waveguide section (*a*); of the active-medium refractive index (*b*); of the active-medium refractive index and amplification factor (*c*) (d – spatial modulation period; L – optical waveguide thickness; ΔL – modulation depth of the waveguide section; 2θ - angle of interference between pump beams I and II)

feedback are demonstrated in Fig. 1.5.6. The periodic spatial structure represents a relief grating on the surface bounding a thin waveguide layer of the active medium (optical waveguide section modulation, see Fig. 1.5.6, *a*) or a volume grating (refractive index and (or) amplification factor modulation, see Fig. 1.5.6, *b*, *c*). Note that in the case shown in Fig. 1.5.6, *a* distributed feedback is always stationary and in the second case it may be both stationary (Fig. 1.5.6, *b*) and dynamic (Fig. 1.5.6, *c*) in character.

A limitation of stationary feedback lasers, where the active medium is represented by the dye-activated gelatin or polymeric films, is their low radiation power. Moreover, their useful life is short due to small volume of the working medium and molecular photobleaching of the majority of dyes in the films. Dye lasers with a dynamic feedback induced by pump radiation (so-called lasers with light-induced distributed feedback) have much greater potentialities.

The characteristic feature of such lasers is simultaneous usage of exciting radiation for the active medium pumping and for the formation of a dynamic spatial grating in this medium. Such a grating is recorded in the active medium with the help of two interfering pump beams (see Fig. 1.5.6, *c*). As this takes place, an excitation field becomes spatially modulated in the plane of an active layer of a distributed feedback laser by the sine law with the period *d* that is dependent on the pumping wavelength $\lambda_н$ and on the angle of interference 2θ between the beams

$$d = \frac{\lambda_н}{2\sin\theta}. \qquad (1.5.9)$$

In this case the formed quick-response grating is automatically "erased" after the termination of the pump beam. Because of this, in the same active medium the grating period and hence the lasing wavelength of distributed feedback lasers is easily tuned by changes in the angle of interference between the excitation beams.

Under the effect of pumping radiation, in the active medium of a distributed feedback laser in the general case the amplitude-phase grating is formed, i.e. we have modulation of both the amplification factor α and the refractive index *n* of the medium. Variations in the active-medium refractive index can result from electrostriction, Kerr effect, thermal heating due to absorption of pump radiation or due to variations in the populations of operating levels in active centers. The contribution made by specific mechanisms into the formation of phase gratings is dependent on thermal, optical, and elastic characteristics of the medium as well as on radiation intensity and duration. When the duration of exciting radiation comes to $\tau_н > 10^{-8}$ s, a phase grating is formed predominantly by thermal

mechanisms associated with radiationless and Stokes losses on optical transitions in the dye molecules. At the same time, at $\tau_н < 10^{-8}$ s the active-medium refractive index is varying predominantly by resonance mechanism. But, if at $\tau_н > 10^{-8}$ s the contributions of phase and amplitude gratings into the formation of distributed feedback are comparable, at $\tau_н < 10^{-8}$ s the amplitude grating plays a decisive role in excitation of the generation on the basis of distributed feedback.

Basic schemes of dye lasers with light-induced distributed feedback

By the present time, a series of interference schemes has been implemented

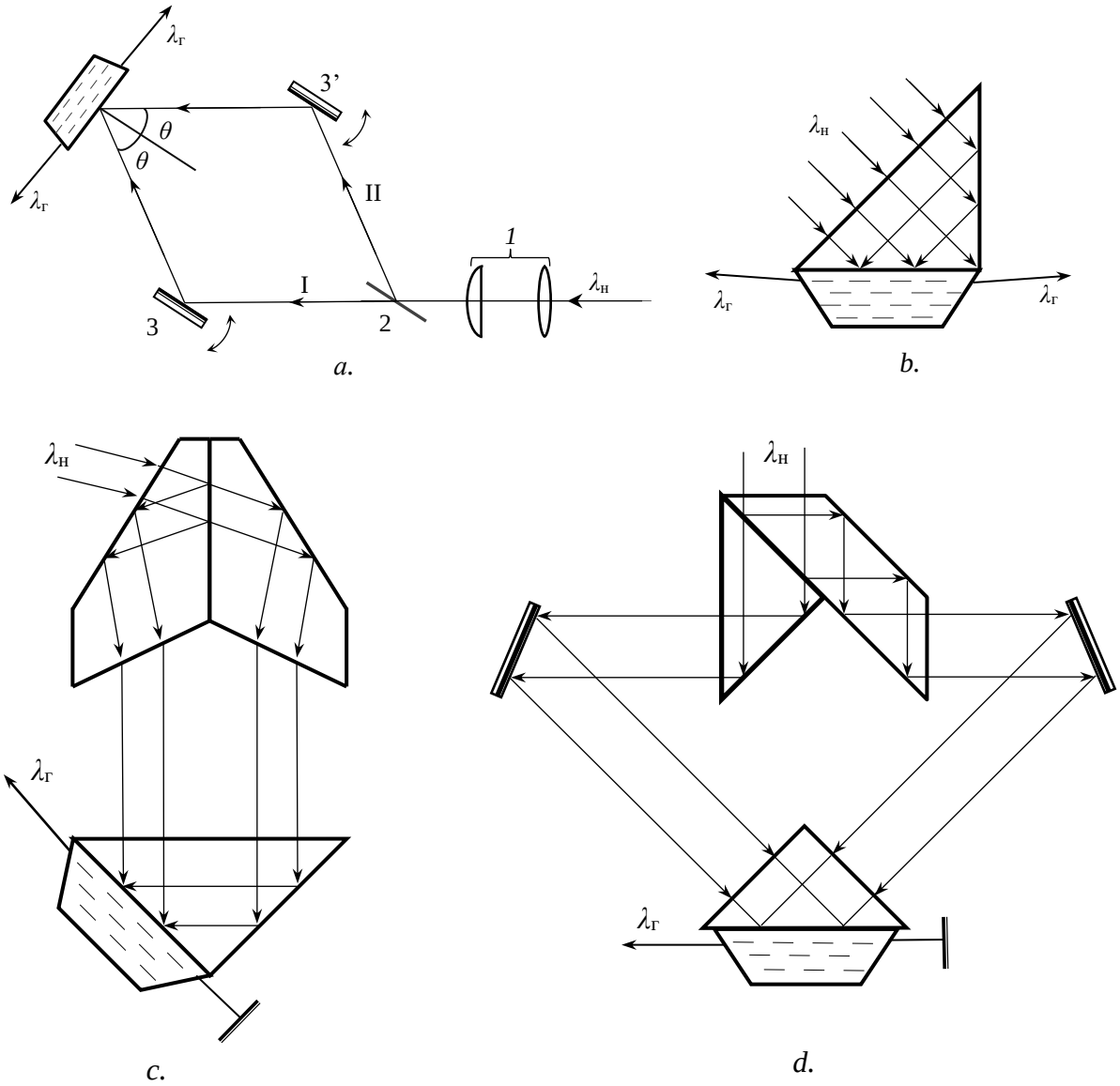


Fig. 1.5.7 – Pumping schemes of dye lasers used to realize light-induced distributed feedback

to realize spatial modulation of the optical parameters of the active laser medium. Fig. 1.5.7 presents most extensively used schemes of dye lasers with light-

induced distributed feedback.

The generation of radiation was first realized in a dye laser with light-induced distributed feedback on the basis of the scheme shown in Fig. 1.5.7, *a*.

According to this scheme, a beam of pump radiation with the wavelength $\lambda_н$ by means of spherocylindrical telescope *I* is formed as a narrow horizontal band that is subsequently split into two beams of approximately equal intensities *I* and *II* by beam splitter 2; these beams with the help of mirrors 3,3′ are convergent to the surface of cell with dye solution 4. The intersection angle of pump beams 2θ is varied by changes in the positions of mirrors: in this scheme the angle of interference is changed by synchronous turning of side mirrors and simultaneous shifting of the mirrors in the transverse direction to retain the beam intersection point invariable. As a result of beam interference, within the dye solution a spatial structure with the period *d* is formed, and we have

$$d = \frac{\lambda_н}{2\sin\theta} = \frac{\lambda_н}{2n_p \sin\theta_p}, \qquad (1.5.10)$$

where n_p – refractive index of a solution; $2\theta_p$ – angle of beam interference in a solution.

According to expression (1.5.8), in the case of the first-order Bragg diffraction (m = 1) the generation wavelength of a distributed-feedback laser is equal to

$$\lambda_г = \frac{n_p \lambda_н}{\sin\theta} = \frac{\lambda_н}{\sin\theta_p}. \qquad (1.5.11)$$

where n_p – refractive index of a dye solution at the wavelength $\lambda_г$.

In the case under study (Fig. 1.5.7, *a* – excitation beams are incident on the active medium from the air) from expressions (1.5.10) and (1.5.11) it follows: a minimal value of the period *d* and of the lasing wavelength $\lambda_г$ is limited by the highest possible angle θ_p that should not be higher than the critical angle for total internal reflection of pump radiation in a solution $\theta_{пред}$

$$\sin\theta_{пред.} = \frac{1}{n_p} \qquad (1.5.12)$$

and hence we have

$$d_{\min} = \frac{\lambda_н}{2}, \tag{1.5.13}$$

$$\lambda_{г,\min} = n_р \lambda_н. \tag{1.5.14}$$

To attain spatial modulation with a lower period, the schemes are used, in which a dye solution is in contact with a prism of glass or quartz as shown in Figs. 1.5.7, *b* - *d*. In this case the limitation on the maximal angle of interference in a medium, imposed by the critical angle of total internal reflection, is lifted. When the refractive index of a prism is $n_{пр}$, a minimal value of the period d_{min} is by a factor of $n_{пр}$ lower than in the case shown in Fig. 1.5.7, *a*. The schemes in Figs. 1.5.7, *b* - *d* enable one to obtain the interference structure with the period

$$d = \frac{\lambda_н}{2 n_{пр} \sin\theta}, \tag{1.5.15}$$

whereas the lasing wavelength is given by the following expression:

$$\lambda_г = \frac{n_р \lambda_н}{n_{пр} \sin\theta}. \tag{1.5.16}$$

A good practice is to use the scheme of a distributed feedback laser demonstrated in Fig. 1.5.7, *b*. In this laser a dye solution is in contact with one of the side faces of the 90°-prism, splitting of pump radiation being realized on reflection from the second face. Tuning of the lasing wavelength is effected by variations in the pump-beam incidence angle on the entrance face of the prism due to rotation of the prism itself or of the auxiliary reflecting mirror. The principal advantage of this scheme is simplicity of operation without the need of alignment. Its significant drawback, in analogy with the earlier considered scheme (Fig. 1.5.7, *a*), stems from the fact that different fluxes of the initial beam are spatially overlapping in the active medium and this, due to low spatial coherence of the beam, results in degraded visibility of an interference field of pumping and hence in lower efficiency of the spatial grating formed in a solution. There is no such a drawback in the optical schemes given in Figs. 1.5.7, *c*, *d* which offer in the active medium an interference of direct wave fronts of the excitation beams formed due to splitting of the initial beam as to amplitude. As a result,

visibility of an interference field of pumping and hence the efficiency of the amplitude-phase grating formed in a dye solution is fairly high, even when using the pumping sources with a low degree of spatial coherence.

Besides, some other schemes of dye lasers with the light-induced distributed feedback (e.g., distributed feedback lasers with a beamsplitter based on holographic grating) are known, each of them having particular advantages and limitations. Selection of one or another scheme depends on the specific application of distributed feedback lasers and on the target of the research, educational or applied problem at hand. Specifically, some of the earlier designed distributed feedback lasers (e.g., Gnom-2, Color, Amethyst, Amethyst-2) are intended for the formation of narrow-band radiation in the nano- and picosecond duration ranges with the use of such pulsed pumping sources as YAG:Nd and TEA N_2 lasers. The indicated lasers make it possible to attain lasing, tunable over a wide spectral range (~ 380 – 850 nm), with the ultimate characteristics.

1.6. The types of lasers and their applications

Lasers may be classified according to several features: type of active medium, excitation type, state of material aggregation, etc. For example, lasers may be subdivided into solid-state, liquid, gas or free electron types.

Among solid-state lasers, of particular importance are standard optically pumped lasers and semiconductor injection lasers. Most common lasers of the liquid type are only dye lasers.

The greatest problem of laser development is excitation of the active medium and the inverse population formation. The energy required for excitation of the active medium may be supplied in different ways. The first ruby laser was excited by light of a flash lamp and this technique was adopted for many other lasers. Lasers of this type are called the optically pumped lasers. In analogy, one can develop systems pumped by electron beams or by other corpuscular radiation. We should distinguish such lasers and free electron lasers, where electrons represent the active medium. Gases may be excited by means of the electric discharge − in this case we have a gas-discharge laser. Direct electric excitation may be realized with the use of semiconductors – in this way injection or diode lasers are designed. According to the excitation type, we can distinguish the following lasers:

– optically pumped lasers with excitation by means of a flash lamp, continuous burning lamp, radiation of other laser or light-emitting diode;

– electron-beam-pumped lasers, e.g., different gas lasers, gas-discharge lasers using glow discharge, arc or hollow electrodes;

– injection or diode lasers with excitation due to the current flowing in a semiconductor;

– gas-dynamic lasers with the population inversion on expansion of gases heated up to high temperatures;

– chemical lasers, where the formed molecules are already in the excited state due to the proceeding chemical reaction;

– nuclear-pumped lasers with excitation by gamma radiation of a nuclear reactor or as a result of nuclear explosion.

1.6.1. Ruby laser

Ruby is a material first used by T.H. Maiman to realize generation of the optical-range electromagnetic waves in June of 1960; ruby is still involved in production of lasers. Ruby that is well known as a natural precious stone represents a crystal of Al_2O_3 (corundum), where some part of AL^{3+} ions is replaced by Cr^{3+}ions. Generally, ruby crystals used as an active laser medium are grown from melt of the mixture of Al_2O_3 and Cr_2O_3. Without Cr^{3+} ions, the formed crystal (sapphire) is colorless. To obtain rosy colored ruby, it is required to add a little of Cr_2O_3 (~0.05 %). In natural gems the concentration of Cr^{3+} is higher by an order and this makes their color deep (red ruby).

The energy levels of ruby are formed due to the three electrons at 3d sheath of Cr^{3+}ion subjected to the field effect of Al_2O_3 lattice. Figure 1.6.1 shows the ground levels which are of particular interest for laser generation. These levels are

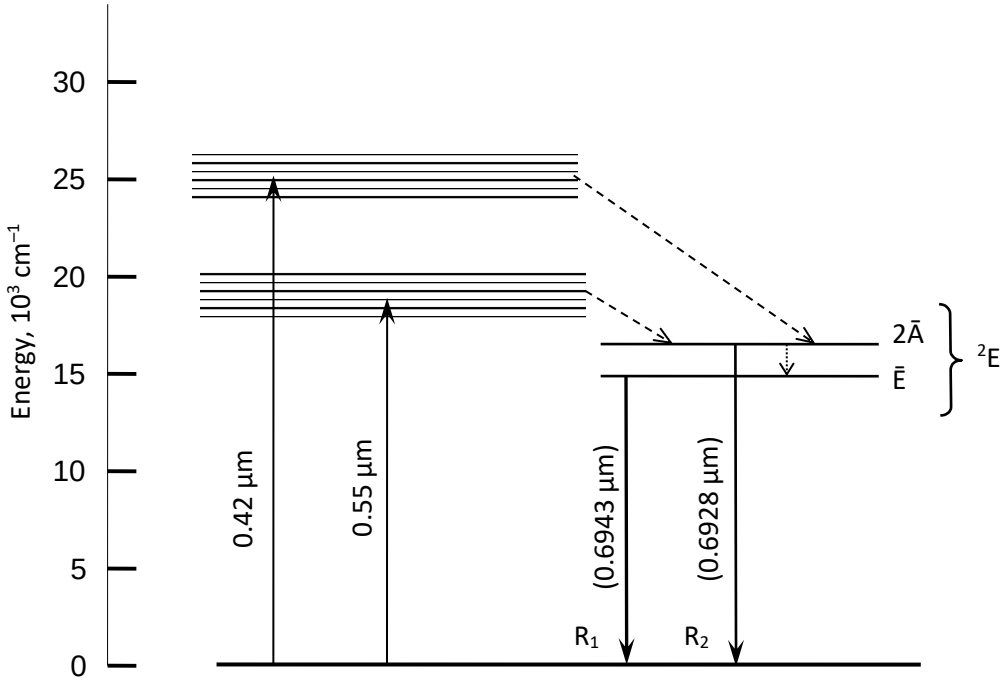


Fig. 1.6.1. Schematic of energy levels for ruby

denoted with symbols of a group theory. From the ground state $^4A_2$ there are two absorption lines: blue (maximum λ=0.42 μm) to the level $^4F_1$ and green (λ =0.55 μm) to the level $^4F_2$. By very rapid (~ 10^{-11} s) radiationless relaxation these levels are bound to the state ^2E, that, due to the spin-orbit interaction, is split into the two sublevels $2\bar{A}$ and $\bar{E}$ (splitting of 29 cm^{-1}). Fast radiationless transitions take place between these levels in time ~10^{-9} s. In accordance with Boltzmann distribution, the lower level $\bar{E}$ has a higher population.

The transition from the state ^2E to the ground state $^4A_2$ is forbidden by the spin selection rule. The chromium ion lifetime in the state ^2E is great, coming at room temperature to $3.4 \cdot 10^{-3}$s. Due to high metastability of the level, it accumulates the required number of particles to initiate the generation in the channel $^2E \rightarrow {}^4A_2$ (three-level scheme of generation).

Also, there is a possibility for generation at the transitions from both levels $2\bar{A}$ and $\bar{E}$ (so-called R_1 and R_2 lines; see Fig. 1.6.1). In ordinary conditions, because of a higher population of $\bar{E}$ level, radiation with the wavelength λ =0.6943 μm (R_1-line) is generated. As the stimulated transition cross-section

comes to $\sigma \sim 3 \cdot 10^{-20}$ cm^2, for the concentration of chromium ions $N_o \sim 10^{18}$ cm^{-3} the amplification factor in ruby may be as high as $k_o \sim 10^{-2}$ cm^{-1}.

Ruby lasers may be operated both in the continuous-wave and pulsed modes. Pumping of the active medium is realized with the help of gas-discharge lamps. Ordinary, a ruby rod is 5-10 mm in diameter at the length 10 – 20 cm. Growth of the crystals is limited by the technological difficulties associated with homogeneity of the crystals, support of homogeneous excitation. The output parameters are follows:

- in the continuous-wave mode – power about10^{-3} W (is not used in practice);
- in the free-running mode – energy up to 300 J, duration ~10^{-3}s;
- in the Q-switching mode − power ~10^{7-8} W, duration ~10^{-8} s;
- in the mode-locking regime − power ~10^9 W, duration ~10^{-11}s.

Among the advantages of ruby lasers, we should name generation in the visible wavelength range; possibility for periodic operation (frequency up to 2 Hz). *Among their disadvantages*, we should name high pump energy (threshold level ~10^{2-3} J), low efficiency (a few fractions of a percent); large divergence of laser radiation (~$10^{-2} \div 10^{-3}$ rad); technological difficulties and high price of manufacturing the active elements.

1.6.2. Neodymium-doped yttrium aluminate laser

Spectral and luminescent properties of the active laser elements based on Nd:YAG crystals are determined by the properties of the matrix (pure undoped crystal of yttrium aluminum garnet) itself and by the characteristics of the introduced neodymium ions. The matrix may have a significant effect on the spectral properties of an isolated neodymium ion, influencing the spectral-line position, intensity, and width; luminescence quantum efficiency, etc. As a rule, the inverse effect of neodymium ions on the matrix is minor due to a relatively low concentration of the ions in it. As a whole, the characteristics of Nd:YAG elements are governed by the matrix properties. Besides, the properties of neodymium ions subjected to specific changes also have their effect. Let us consider these properties.

Neodymium is a rare-earth metal belonging to the lanthanide group. The optical properties of neodymium ions are determined by electron transitions within the subshell 4 *f* of the fourth electron shell of a neodymium atom. This subshell is, to a considerable degree, shielded from the effect of external electric

fields (intracrystalline field including) by electrons of the outer subshells $5s$ and $5p$, making it possible to consider laser generation based on a model for energy levels of a free neodymium ion. The electron configuration of the neodymium ion Nd^{3+} $(4f^3 5s^2 5p^6)$ is associated with a set of doublet and triplet terms. The quartet term ${}^4I_{9/2,\,11/2,\,13/2,\,15/2}$ is primary. Note that the symbol characterizing each level takes the form ${}^{2s+1}L_J$, where s – total spin quantum number; L – orbital quantum number; J – total quantum number of the angular momentum. The allowed values of L, namely: $L = 0, 1, 2, 3, 4, 5, 6 \ldots$, are denoted with capital letters as $S, P, D, F, G, H, I \ldots$, respectively. In this way the ground state ${}^4I_{9/2}$ of Nd^{3+} ion corresponds to the state for which we have $2S+1=4$ (i.e., $S = 3/2$), $L=6$ and $J=9/2$.

Figure 1.6.2 presents a simplified scheme for energy levels of Nd^{3+} ion. These levels are caused by the transitions made by three $4f$ electrons of the inner ion shell. An absorption spectrum of a neodymium ion comprises a great number of comparatively narrow bands, most intensive of which are associated with the wavelengths 240, 350, 520, 580, 740, 800, and 900 nm. After absorption of the pump energy, a neodymium ion in time $< 10^{-8}$ s goes to the metastable level ${}^4F_{3/2}$ with the lifetime $10^{-3} - 10^{-4}$ s. Among different possible transitions from

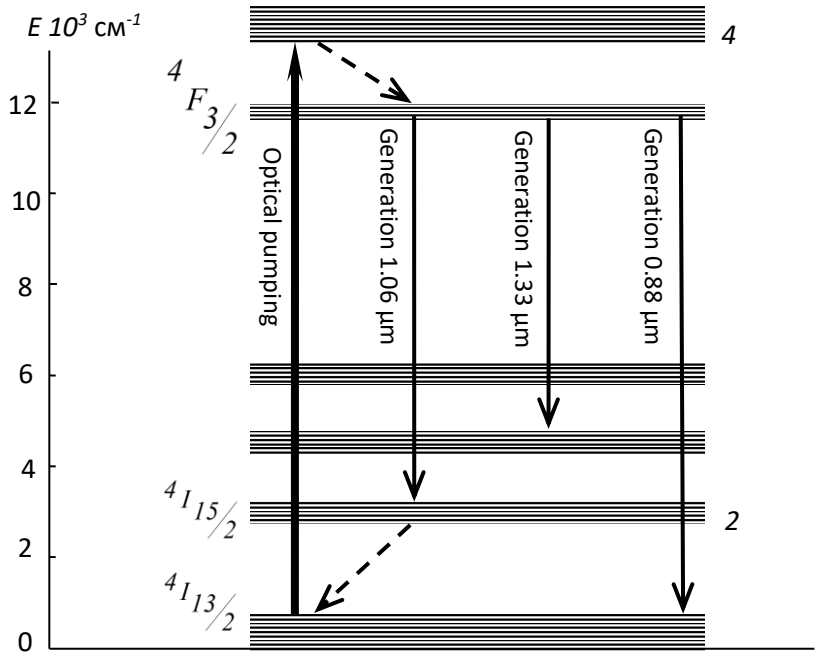


Fig. 1.6.2. Energy levels of a trivalent neodymium ion

the level ${}^4F_{3/2}$ to the lower-lying levels (${}^4I_{9/2,\,11/2,\,13/2,\,15/2}$), the transition ${}^4F_{3/2} \rightarrow {}^4I_{11/2}$ at the wavelength $\lambda = 1.06$ μm has the highest intensity. The level ${}^4I_{11/2}$ is connected to the ground state ${}^4I_{9/2}$ by the rapid (on the order of several nanoseconds) radiationless transition. The difference between the energies of ${}^4I_{11/2}$ and ${}^4I_{9/2}$ is greater than kT almost by an order of magnitude; according to the Boltzmann statistics; in a good approximation the level ${}^4I_{11/2}$ may be considered to be practically empty.

Proceeding from all the above, operation of a neodymium laser may be described by a four-level scheme: pumping is realized in channel *1–4* (transition from the ground state ${}^4I_{9/2}$ to the excited levels) and generation is realized in channel *3–2* (transition ${}^4F_{3/2} \rightarrow {}^4I_{11/2}$).This accounts for the fact that a neodymium laser has an advantage over the earlier described ruby laser that operates according to a three-level scheme (see Fig. 1.6.1). The lower generation level of a ruby laser is concurrently the lowest ground energy level, whose population may be reduced (for the inverse population) only by enhanced pumping.

Important chracteristics of laser levels are their width and broadening character. In the case of an isolated ion a width of levels is dictated by its lifetime at these levels. Ions in real laser media are subjected to the matrix effect leading

to broadening of the levels. Due to inhomogeneity of local electrctostatic fields of the immediate environment in glasses, the luminescence line at 1.06 μm is inhomogeneously broadened ($\Delta\lambda \cong 30$ nm), whereas in yttrium aluminum garnet crystals broadening is homogeneous ($\Delta\lambda \cong 0.7$ nm).

1.6.3. Helium-neon laser

Helium neon laser was developed at the end of the 1960-ies. This is the first gas laser and a laser operating in the continuous-wave mode. Its pumping is of the electric type. The active laser medium is created due to the glow discharge within a glow-discharge tube. The active material comprises a mixture of helium and neon atoms, their ions, and free electrons. Schematic of energy levels for helium and neon atoms is shown in Fig. 1.6.3. Helium levels are denoted as is conventionally accepted for LS bond: the principal quantum number on the left gives the shell, where an outer electron is positioned; the letter S denotes the total orbital momentum that is equal to zero (the values $L = 0, 1, 2, 3$ ... are denoted

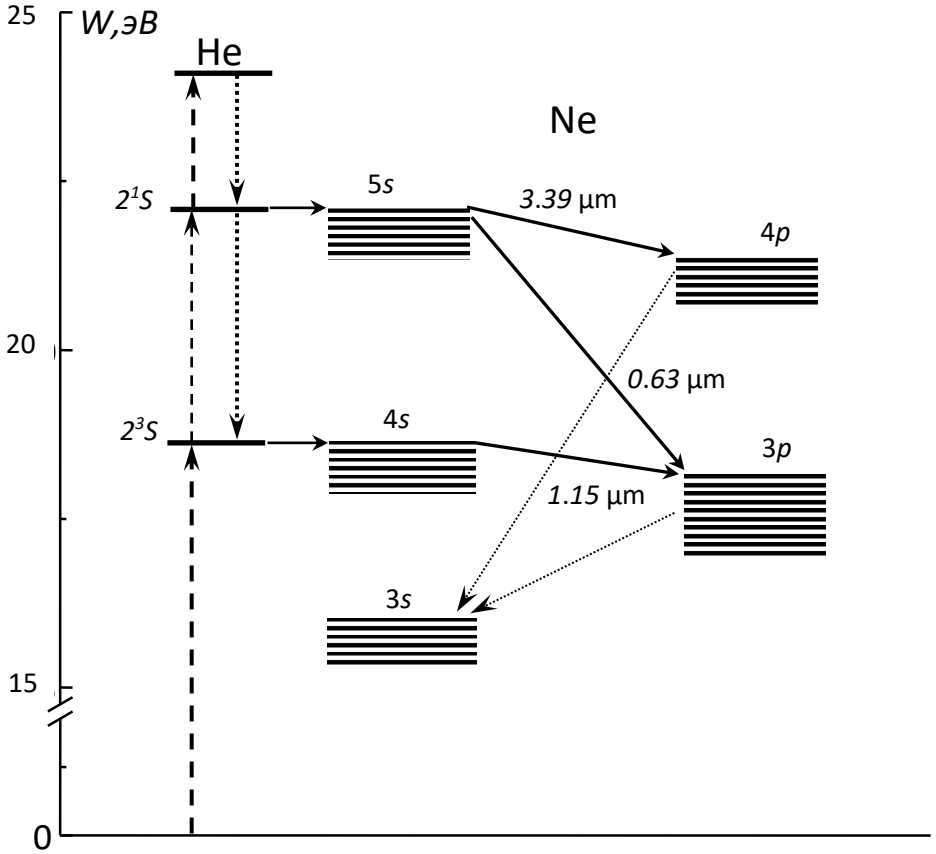


Fig. 1.6.3. Schematic of energy elevels for helium and neon atoms

by S, P, D, F ...). The superscripts are associated with multiplicity of levels: 1 denotes singlet levels, 3 – triplet levels. Ten electrons of a neon atom in the ground state form the configuration $1s^2 2s^2 2p^6$. The excited levels shown in the figure are associate with the situations when one of $2p$ electrons is activated to the excited s-state ($3s$, $4s$, $5s$) or p-state ($3p$, $4p$).

The population inversion at the levels of neon atoms is created in helium-neon lasers indirectly. Due to the gas discharge, free electrons excite helium atoms which go to higher energy levels and then are subjected to downward stepwise relaxation, slowing down at the metastable levels 2^1S and 2^3S with the lifetime on the order of 1 ms. Neon atoms, on collisions with such atoms, go to the levels $4s$ and $5s$ of the multiplets with the lifetime ~ 0.1 μs. With further transition to the state $3p$ or $4p$ the process of stimulated emission may be realized. The lifetime of the states $3p$ and $4p$ is on the order of 0.01 μs permitting the formation of the inverse population in the channels $5s \to 3p$ (generation wavelength 632.8 nm), $4s \to 3p$ (1152 nm), $5s \to 4p$ (3392 nm). Effective relaxation of the levels $3p$ and $4p$ necessitates continued depletion of $3s$-state. This process is realized only on collisions of neon atoms with the walls of a glow-discharge tube. To achieve sufficiently fast diffusion to the walls, the cross-section of the tube is made small coming to a few millimeters only.

The majority of commercial helium-neon lasers are operated at the wavelength 632.8 nm. To realize laser generation only at this wavelength, a resonator with selective dielectric mirrors is used. The typical parameters for the plasma of He-Ne lasers operating at the wavelength 632.8 nm: atomic temperature $T \approx 300$–400 °Kl; discharge tube diameter $d \approx 3$–8 mm; $pd \approx 500$ Pa · mm (p – pressure). The concentration ratio of helium and neon is 5 : 1. The current is from 5 to 50 мА. The power of these lasers is not very high (up 80 mW) but they have several advantages over other lasers: simplicity of operation; reliability; small radiation divergence ($\approx$ 1–10 angular minutes) and high monochromaticity degree (ratio between the generation line width $\Delta\nu$ and the frequency ν for commercial lasers $\approx 10^{-6}$–10^{-8}). Radiation of He-Ne lasers, where windows of the discharge tube are at the Brewster angle to lower the reflection loss, is linearly polarized.

Different processes influence the atomic distribution over the energy levels in glow-discharge lasers. The principal process leading to the population inversion

in an He-Ne laser – quasi-resonance energy exchange between atoms of helium and neon. In the process neon atoms are excited due to collisions of the second kind with the excited helium atoms

$$\mathrm{He}^{*}\left(2^{1}S\right)+\mathrm{Ne}\rightarrow\mathrm{He}+\mathrm{Ne}^{*}(5s), \qquad (1.6.1)$$

$$\mathrm{He}^{*}\left(2^{3}S\right)+\mathrm{Ne}\rightarrow\mathrm{He}+\mathrm{Ne}^{*}(4s). \qquad (1.6.2)$$

The symbol « * » denotes excited atoms. The probability of the processes described by formulae (1.6.1) and (1.6.2) is high as the energies of excited levels for He and Ne are close (Fig. 1.6.3). To attain inversion at the laser levels of neon $4s$ and $5s$, helium atoms should be excited to the corresponding levels $2^{1}S$ and $2^{3}S$. This is achieved due to the direct excitation by electron impingement

$$\mathrm{He}+\hat{e}\rightarrow\mathrm{He}^{*}+e. \qquad (1.6.3)$$

The symbol « ^ » indicates that the electron has a high kinetic energy that should be equal to or higher than that of the excited level for helium. In the case of the level $2^{1}S$ this energy is equal to 20.61 eV. The required energy is provided by electron acceleration in a glow-discharge tube.

The population inversion at the laser levels of neon $4s$ and $5s$ is lowered by the stepwise processes when an atom already positioned at the excited level is excited or ionized. This may be due to depletion of the helium level $2^{1}S$ or $2^{3}S$ by electron impact

$$\mathrm{He}^{*}\left(2^{1(3)}S\right)+\hat{e}\rightarrow\mathrm{He}^{**}+e, \qquad (1.6.4)$$

$$\mathrm{He}^{*}\left(2^{1(3)}S\right)+\hat{e}\rightarrow\mathrm{He}^{+}+e+e \qquad (1.6.5)$$

or due to an additional excitation of neon atoms, e.g., from the level $3s$

$$\mathrm{Ne}^{*}\left(3s\right)+\hat{e}\rightarrow\mathrm{Ne}^{*}(3p)+e, \qquad (1.6.6)$$

where He^{**} – helium atom that is in a higher energy state compared to $2^{1}S$; He^{+} – helium ion.

The processes described by formulae (1.6.4) and (1.6.5) lead to lower populations of the upper laser level because of the depleted helium levels $2^{1}S$ and $2^{3}S$ and the lowered population efficiency of the excited levels $4s$ and $5s$ for neon. The process described by formula (1.6.6) is responsible for an increase in the population of the lower laser level. Any of these stepwise processes results in lower density of the population inversion, this lowering being the greater the

higher is the product $N_a n_e$, where N_a – atomic concentration at the corresponding level; n_e – electron concentration. As the atomic concentration at the excited level N_a is also proportional to n_e, the probability of a stepwise process is proportional to n_e^2.

In this way, as the discharge (electron concentration n_e) is growing, the concentration of helium atoms at the levels 2^1S and 2^3S [this process is described by formula (1.6.3)] is increased together with the concentration of *Ne* atoms at the upper laser levels [these processes are described by formulae (1.6.1) and (1.6.2)]. The probability of both processes is given as $\sim n_e$ and hence the lasing power is also growing. This growth is the case until the stepwise processes (1.6.4) – (1.66) with the probability proportional to n_e^2, which cause lowering of the inverse population, become significant. Then, with the increased current, the lasing generation begins to fall. As a result, the population inversion density is maximal for a specific value of n_e. This value of the electron concentration n_e is associated with an optimum value of the electric current.

1.6.4. Dye lasers

Of great diversity of laser types, dye lasers rate high due to their ability to provide smoothly tuned coherent radiation over a wide wavelength range. The possibility to tune the lasing frequency is determined by the properties of the active medium used. Dye is a complex organic compound with a branching system of conjugated bonds that reveals high-intensity absorption bands in the near-ultraviolet (UV), visible, near-infrared (IR) spectral regions.

Currently, thousands of dyes are known, they differ in the form and position of absorption and luminescence bands; in probabilities of spontaneous or radiationless transitions for which one can realize the generation of laser radiation from 300 nm to 1 μm and more. Fig. 1.6.4 shows the chemical structure of a typical dye molecule and spectral dependences for the absorption (curve *1*) and luminescence (curve *2*) cross-section.

The luminescence line of dyes as complex organic compounds is surprisingly wide (up to 100 nm). Note that dyes are characterized by two systems of electron states (Fig.1.6.5). One of them represents the singlet levels S_i (on excitation a spin of the electron stays antiparallel to that of the remaining part of the molecule).

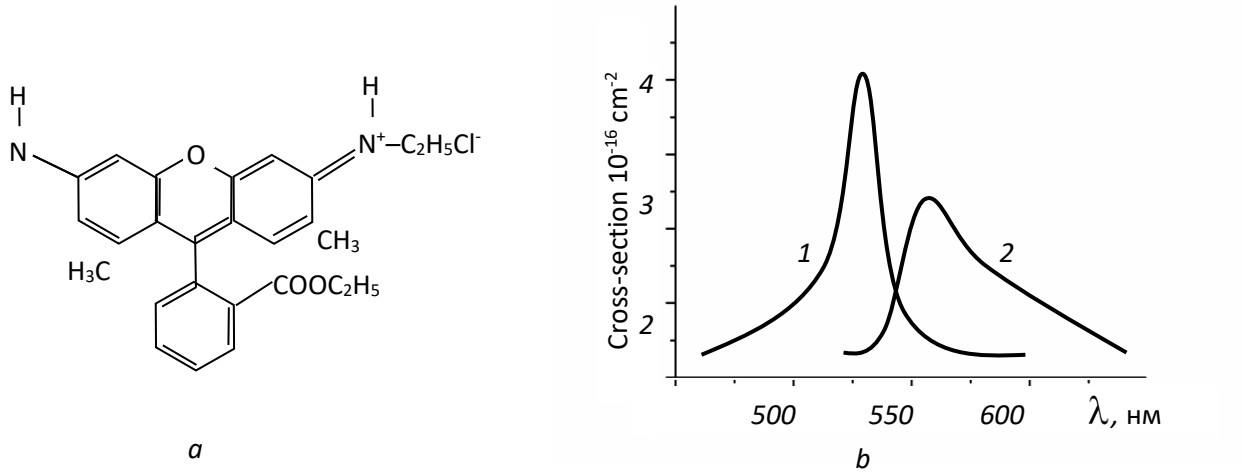


Fig. 1.6.4. – Structural formula for the Rhodamine 6J dye (*a*); absorption (*1*) and emission (*2*) cross-sections of the dye solution in ethyl alcohol (*b*)

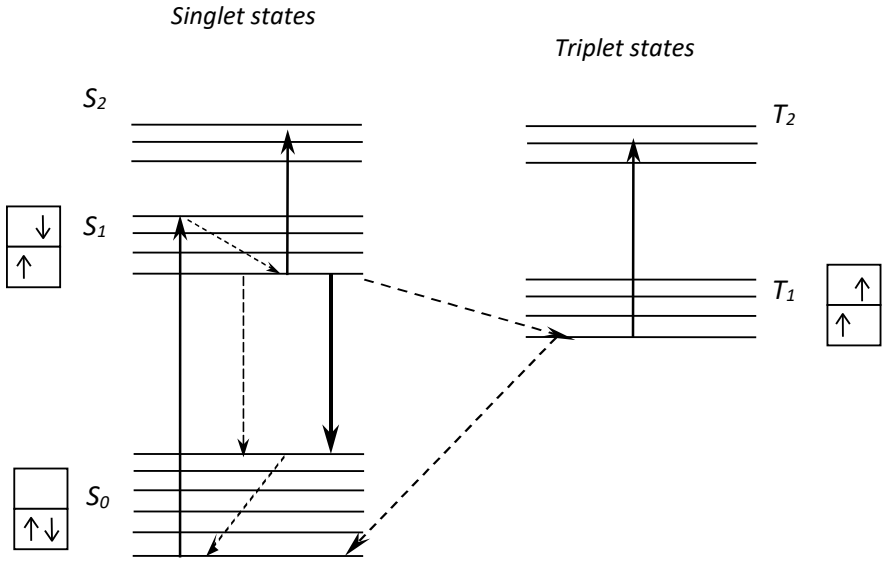


Fig. 1.6.5 – Schematic of energy levels and transitions for a dye laser

The other system consists of the triplet energy levels T_i (a spin of the electron is turned over to become parallel to the spin of the remaining molecule).

Electron states represent wide bands with a continuous set of vibrational sublevels. As a rule, time of the vibrational energy redistribution is considerably shorter than the lifetime of excited states (in dye solutions the vibrational relaxation time comes to $\sim 10^{-13}$ s). Because of this, electron levels are homogeneously broadened. To describe the processes of interaction between radiation and complex organic molecules over a wide range of intensities, we use the averaged Einstein coefficients for transition probabilities. Owing to optical

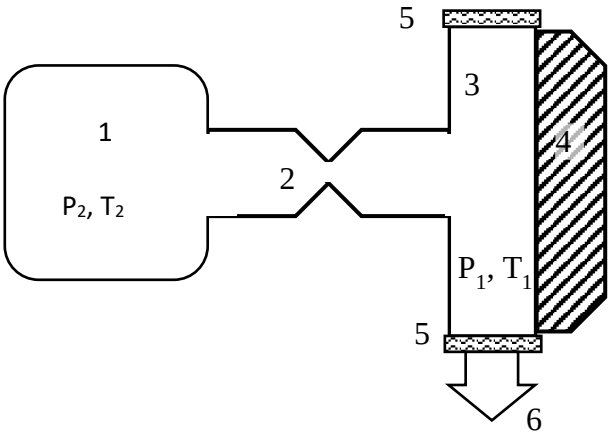
pumping, a molecule is activated from the electron state S_0 to the electron state S_1. Then the molecule can have the radiative transition $S_1 \to S_0$ (fluorescence) or it can have a radiationless transition to the state S_0 or to the triplet level T_1 with subsequent relaxation to the ground state. Overlapping of absorption bands from the ground S_0 and from the excited (S_1, T_1) levels dictates the need for multilevel models of the molecular energy states.

The first experiments with generation in solutions of dyes, when glow-discharge lamps were used as pump sources, were not successful. It has been found that this was associated with triplet states having an adverse effect on the generation process. Accumulation of the molecules in the metastable state T_1 contributes to absorption (transition $T_1 \to T_2$) of both pump and induced radiation. This problem may be solved by the use of pulse pumping, during the period of which there is no time for population of the metastable level T_1. In the case of continuous-weave lasers the triplet effect, when a dye flows through the excitation area, may be reduced as follows. At the rates from 10 до 100 m/s dyes pass through the active area for 1 μs and hence the level T_1 has no time to be populated. However, in this case the negative effect on the generation process may be exerted by induced absorption from the level S_1 (transition $S_1 \to S_2$).

At the generation of radiation the upper energy level of a laser is simultaneously the lowest vibrational level of the state S_1. The lifetime is longer, about 1 ns, than that in the lower laser state − excited vibrational level of the ground state S_0. During the period of a few picoseconds, due to relaxation, it goes to the lowest vibrational level.

1.6.5. Gas-dynamic CO2 – lasers

High radiation powers may be attained with the help of gas-dynamic lasers. A gas-dynamic CO_2 laser uses a mixture of gases (8 % CO_2, 90 % N_2, 2 % H_2O), molecules of CO_2 being the active centers. A simplified scheme of such a laser is shown in Fig. 1.6.6. A gas mixture is first kept in settling chamber 1 at high pressure (P_2=20 ÷ 30 atm) and high temperature (T_2=1400 ÷ 1600 K). With the use of a series of parallel nozzles 2, the mixture expands in effective volume 3, being cooled in the process (pressure P_2=0.05 ÷ 0.1 atm, temperature T_2=250 ÷ 300 K). The gas flow rate at the output of the nozzle unit comes to 1200 ÷ 1500 m/s. Within effective volume 3, the molecules of CO_2 are bleaching.



1 – settling chamber, 2 – nozzle unit, 3 – effective volume (optical cavity volume), 4 – diffuser, 5 – cavity mirror, 6 – laser radiation.

Fig. 1.6.6. Simplified scheme of a gas-dynamic laser

Diffuser 4 retards the gas flow and correlates its pressure with that of the outer atmosphere.

A molecule of CO_2 has four vibrational degrees of freedom corresponding to the three vibration types: symmetric, deformation (or bending), asymmetric. The frequency of these vibrations is denoted as ω_1, ω_2, and ω_3, respectively (see Fig. 1.6.7). We should note an interesting feature: $\hbar\omega_1 = 0.163$ eV, $\hbar\omega_2 = 0.078$eV, $\hbar\omega_3 = 0.2763$ eV, $\omega_1 \approx 2\omega_2$. Vibrational states of CO_2 molecule are denoted by a set of vibrational quantum numbers, v_1, v_2, v_3, which are equal to the excitation order of symmetric, deformation, and asymmetric vibrations, respectively. To

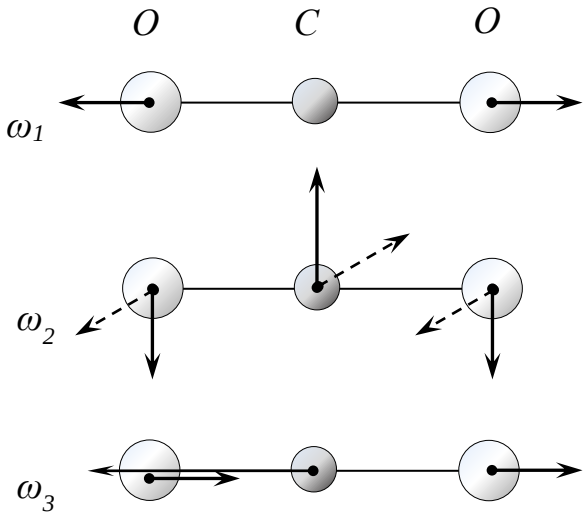


Fig. 1.6.7. Vibration types of CO_2 molecule

illustrate, the vibrational state (020) (v_1=0, v_2=2, v_3=0) is associated with double excitation of deformation vibrations and with the absence of symmetric and asymmetric vibrations. A mechanism of the inversion formation in a gas-dynamic CO_2-laser is demonstrated in Fig. 1.6.8 representing the levels associated with

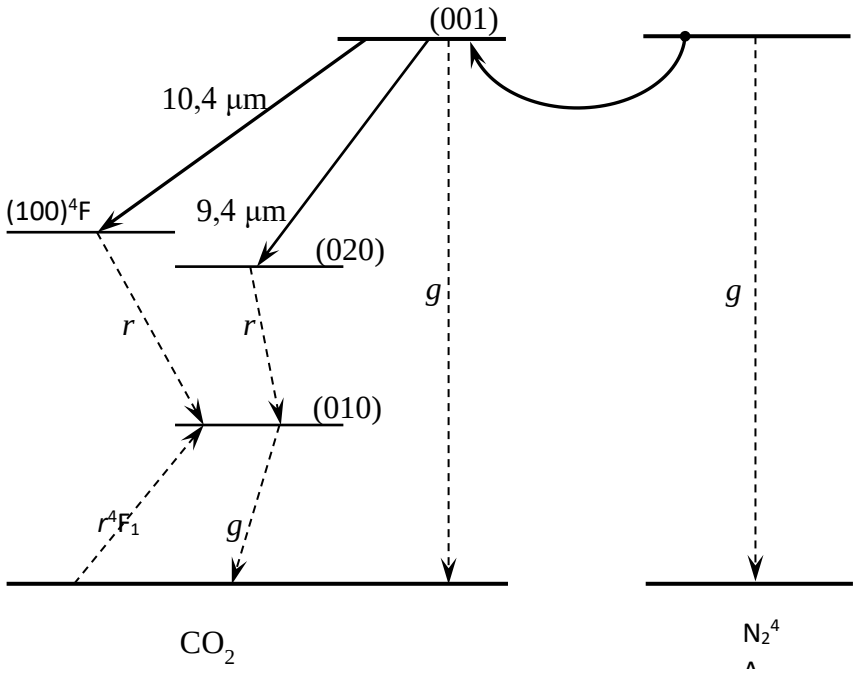


Fig. 1.6.8. Mechanism of inversion in a CO_2-laser

three different vibration types of CO_2 molecule and the first excited vibrational level of a molecule of N_2.

The population inversion is created at the transitions (001) → (100) and (001) → (020) in a molecule of CO_2. For the population of the upper effective level (001), of particular importance are the processes of the excitation energy transfer from nitrogen molecules. Note that *r* – resonance transition, i.e., energy transfer to unexcited molecules of CO_2; *g* – gas-kinetic mechanism of the energy transfer by molecules of H_2O. Relaxation of the excitation level for N_2 molecule and of the level (001) for CO_2 molecule is realized due to a gas-kinetic mechanism of the energy transfer, whereas relaxation of the levels (100) and (020) for CO_2 molecule is due to a resonance mechanism. As the resonance energy transfer rate is much higher than the gas-kinetic rate, the upper effective level of CO_2 molecule should relax more slowly than the lower effective levels. Besides, an addition of water vapors contributes to faster relaxation of the level (010). As the lifetime at the upper laser level is significantly greater than that at the lower one, depletion of the latter proceeds much faster. Because of this, the population inversion is created at the laser transition.

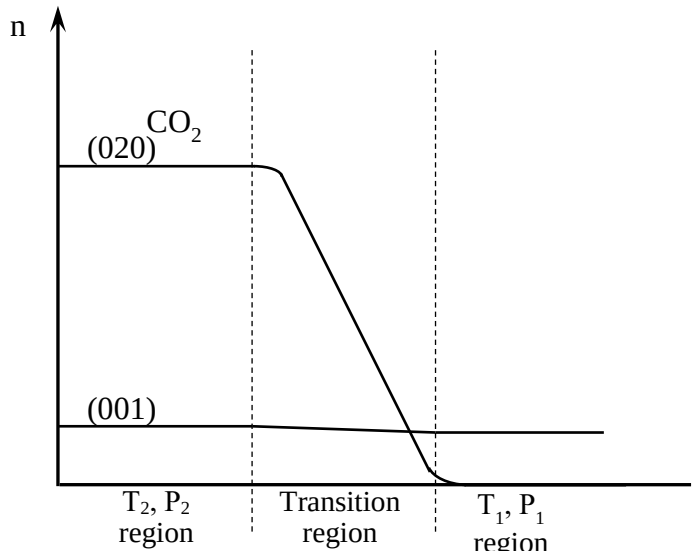


Fig. 1.6.9. Character of variations in populations of the levels (020) and (001) for a molecule of CO_2

Fig. 1.6.9 demonstrates a character of variations in populations of the levels for a molecule of CO_2 as gas mixture flows from the settling chamber to the effective volume. After the mixture passes through the nozzle, the level (020) becomes practically completely empty and the population of the level (001) decreases insignificantly. In the effective volume the upper level population actually remains as at temperature T_2. So, molecules of CO_2 enter the effective volume with inversion of the population at the levels (001) and (020). It is important to note the "frozen" population of a vibrational level of N_2 molecules which resonantly transfer the excitation energy to CO_2 molecules, maintaining a rather high population of the level (001). As nitrogen in the gas mixture is the major component (90 %), we can conclude that the energy of coherent laser radiation is mainly drawn from the vibrational energy of nitrogen molecules.

Gas-dynamic CO_2-lasers feature high radiation power, in the continuous-wave mode coming to 100 kW. Unfortunately, the efficiency of such lasers is inadequate – no higher than 1 %. Because of this, gas-dynamic lasers hardly can find extensive applications.

1.6.6.Semiconductor lasers

Semiconductor lasers with semiconductor crystals as active medium are most widely used now. By their physical nature semiconductor lasers are similar to light-emitting diodes but the generated radiation of these lasers is coherent. As distinct from other laser types, in semiconductor lasers the radiative transitions between allowed energy bands rather than between discreet energy levels are

used. It is well known that two principal conditions are required to attain the generation mode: (1) creation of the inverse population; (2) using of a resonator to afford positive feedback and quantum amplification.

The possibility to create the inverse population in a semiconductor is illustrated in Fig. 1.6.10 showing the valence band *V*, conduction band *C*, and forbidden band with the width E_g. At low temperatures and without external excitation all electrons in a semiconductor are in the valence band *V*, all the energy levels are filled, and the conduction band *C* is empty. In Fig. 1.6.10, *a* the region of the filled energy states is hatched. When a semiconductor is excited in some or other way, i.e. when electrons are transferred from the valence to the conduction band, within each band the equilibrium state is established, the electrons go to the lower energy levels of both bands leaving the upper part of the valence band, where for the electrons emerge vacancies, the so called holes. The levels E_{fV} and E_{fC} in Fig. 1.6.10, *b* are called the Fermi levels of the corresponding bands, these levels determine the boundaries between filled and empty energy regions within each band. Values of the Fermi levels for both bands depend on the number of electrons transferred into the conduction band: the greater the number of electrons the higher the Fermi level of the conduction band and the lower the Fermi level of the valence band. As a result, inverse population is formed between the lower region of the conduction band and the upper region of the valence band. Electrons from the conduction band can return back to the valence band, recombining with holes and emitting a photon in the process. Such radiation is known as

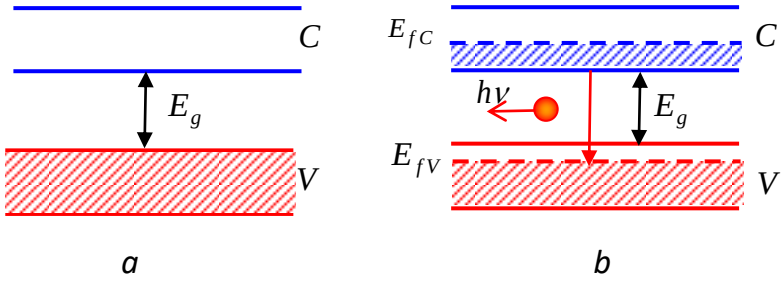


Fig 1.6.10 – Operation principle of a semiconductor laser

recombination radiation.

Excitation of the valence band electrons for the creation of the inverse population may be realized in different ways. One can use an external electron beam or another laser for excitation. However, most wide spread is the technique when the amplifying medium is created with the use of the *p* – *n*-junction across which the electric current is flowing. There are two main types of semiconductor lasers: homojunction (homostructure) lasers and double heterojunction lasers.

Homojunction lasers utilize a single *p* – *n*-junction with the applied external forward voltage, similar to the standard light-emitting diodes (LEDs). Semiconductor homojunction lasers consist of the same semiconductor layers (e.g., *GaAs*), where impurities are added to create the *p*- and *n*-type conductivity. A semiconductor of the *n*-type possesses donor impurities donating electrons to the conduction band. A semiconductor of the *p*-type, on the contrary, possesses acceptor impurities picking up electrons from the valence band to the acceptor level. In the process holes are formed in the valence band. In this way the Fermi level for the *n*-type semiconductor is lying in the conduction band and for the *p*-type semiconductor – in the valence band. With the *p* – *n*-junction formation, the energy bands are transformed so that both Fermi levels have identical energies (Fig. 1.6.11, *a*). When the voltage *U* is applied to the *p* – *n*-junction in the forward direction, Fermi levels are separated by the interval $\Delta E = eU$ (*e* – electron charge) and the band structure of the *p* – *n*-junction takes the form shown in Fig. 1.6.11, *b*. As seen, the inverse population in this case is created in the junction area.

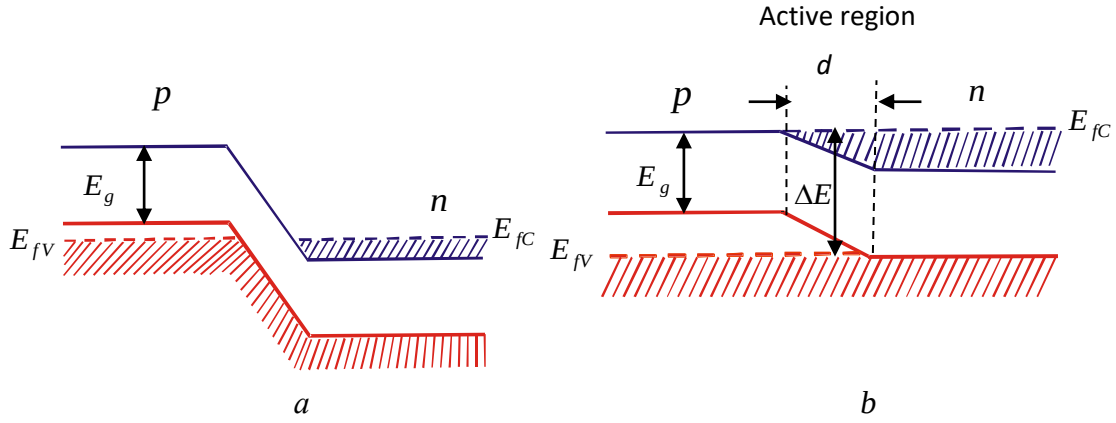


Fig 1.6.11 – Operation principle of a semiconductor laser at the a p – n-junction:
a – positions of the energy bands without the voltage applied;
b – with the voltage applied in the forward direction

In other words, the inverse population is resultant from injection of the nonequilibrium charge carriers across the *p* – *n*-junction. When the voltage is applied to the *p* – *n*-junction in the forward direction, electrons are injected from *n*- to *p*-region, and in the backward direction − injection of holes takes place (Fig. 1.6.12). The higher the current flowing across the *p* – *n*-junction, the greater number of electrons is injected into the *p*-region, contributing to the population of electrons in the excited state. The inverse process – injection of holes into the *n*-region – results in lowering of the electron number in the ground state, electron vacancies are formed.

For a specific current the inverse population condition is fulfilled: the number of electrons in the excited state (lower region of the conduction band) is

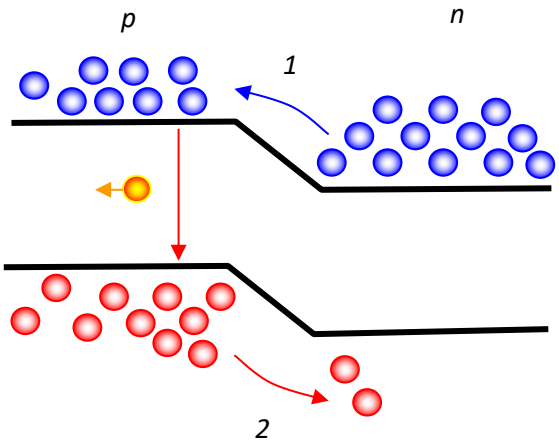


Fig. 1.6.12 – Injection of carriers at the p – n-junction:
1 – electron injection; *2* – hole injection

higher than that in the ground state (upper region of the valence band). On transition from conduction to the valence band electrons recombine with holes. The recombination process is accompanied by photon emission. A width of the light region (active region width l_d) is determined by a diffusion length (diffusion distance of the minority carriers injected into the corresponding regions of the *p* – *n*-junction until the moment of recombination). As a diffusion length of electrons is much greater than that of holes, light emission is observed mainly in the *p*-region. The electron diffusion length l_d is defined as $l_d = \sqrt{D\tau}$, where D – diffusion coefficient; τ – lifetime of an electron in the excited state. For the typical values of the above-mentioned parameters (D = 10 cm^2/s, τ = 10^{-9} s) the active region width of a semiconductor laser comes to l_d ~ 1 μm.

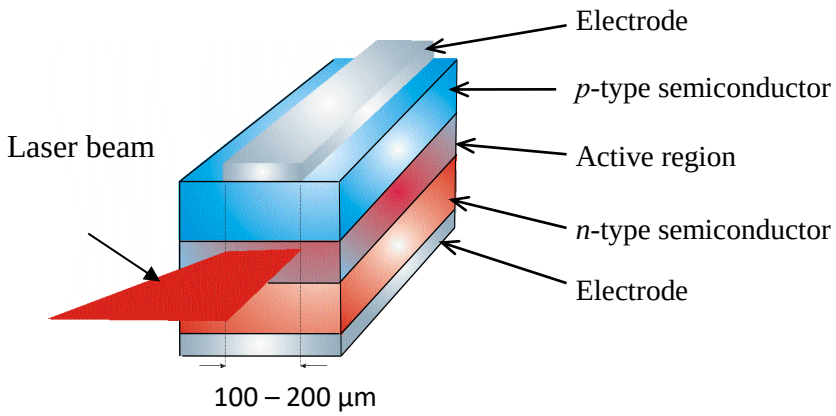


Fig 1.6.13 – Semiconductor laser utilizing the p – n-junction

Generally, resonator mirrors of a laser are polished faces of the semiconductor crystal itself with the reflection factor about 30 %. Sometimes external resonators are used. Fig. 1.6.13 demonstrates the typical design of a laser utilizing the *p* – *n*-junction.

The generation occurs when optical pumping exceeds the energy loss due to coupling of radiation out of the resonator, intracavity absorption and scattering. It should be noted that the active region of a homojunction laser is bound by the electron diffusion length that is considerably shorter than the light propagation region in the crystal. This results in a high value of the current associated with the generation initiation. For example, the threshold current density of a *GaAs* laser at room temperature comes to ~1 кA/cm^2. This direct current leads to a considerable heat release, hindering the continuous generation initiation at high temperatures. Ordinary, homojunction lasers are operating in the pulsed mode.

To lower the threshold current density, semiconductor lasers utilizing heterostructures (heterolasers) have been developed. At the present time these lasers are used most extensively. Double heterostructure lasers, as seen from Fig. 1.6.14 illustrating a $GaAs/Al_xGa_{1-x}As$-laser, represent a combination of semiconductor layers with a narrow forbidden band or band gap (*GaAs*) and semiconductor layers with a large band gap (*AlGaAs*). Laser amplification arises in the *GaAs*-layer that is called the active layer and acts as an amplifying medium. Layers of $Al_xGa_{1-x}As$, called the coating layers, limit spreading of the charge carriers from the active layer.

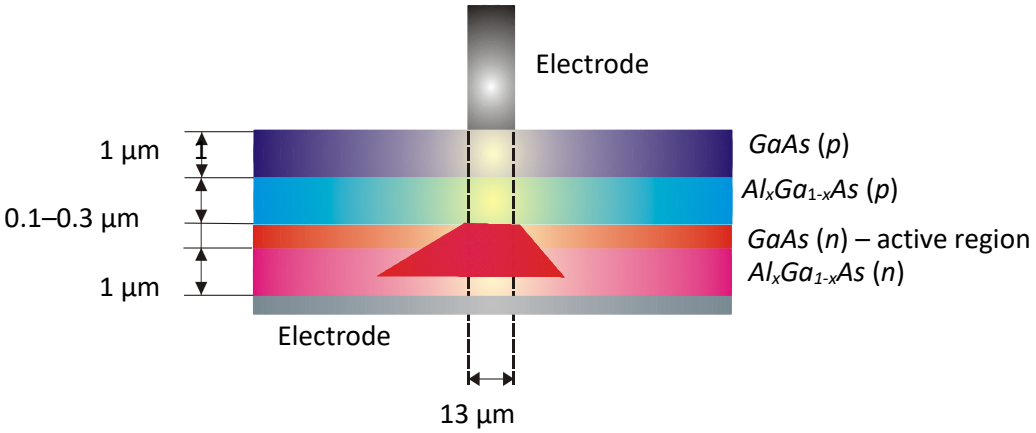


Fig 1.6.14. Double heterojunction semiconductor laser

Double heterostructure has considerable advantages over the homotropic laser structure. First, the electromagnetic field amplification is greatly increased. This is associated with the fact that the band gap *GaAs* ($E_{g1} = 1$ eV) is considerably lower than the band gap of $Al_xGa_{1-x}As$ (1.8 eV), at both junctions the

formed energy barriers effectively confine the injected electrons and holes in the active layer (Fig. 1.6.15, *a*). Second, the refractive index of *GaAs* ($n_1 = 3.6$) is higher than that of *AlGaAs* ($n_2 = 3.4$), leading to the formation of a waveguide in the region of the active medium *GaAs* (Fig. 1.6.15, *b*), where the electromagnetic radiation is localized. As the band gap E_{g2} for *AlGaAs* is much greater than the band gap E_{g1} for *GaAs*, laser radiation at the frequency $\nu = E_{g1}/h$ is actually unabsorbed in *AlGaAs*, and the wings of the transverse profile of a beam, which enter the *p*- and *n*-regions of *AlGaAs*, experience no absorption (Fig. 1.6.15, *c*).

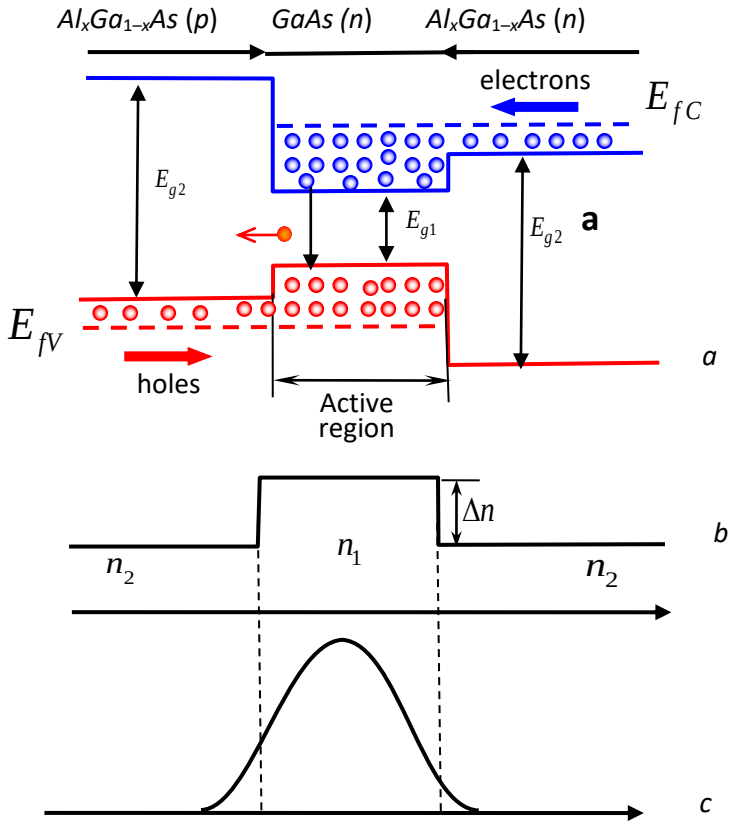


Fig. 1.6.15 – Energy band structure of the heterojunction:
a – injection of minority carriers into the active layer;
b – refractive index distribution; c – light beam profile

The watt-ampere characteristic that is one of the basic characteristics of semiconductor lasers represents the relationship between radiation power and pump current; of particular importance are the directivity diagram and the spectral composition of radiation. The typical watt-ampere characteristic of a semiconductor laser is given in Fig. 1.6.16. The transition from the LED mode to laser radiation is abrupt and is the case as soon as the pump current exceeds the threshold value. In this case the radiation power and directivity as well as polarization degree of laser radiation is drastically growing. For example,

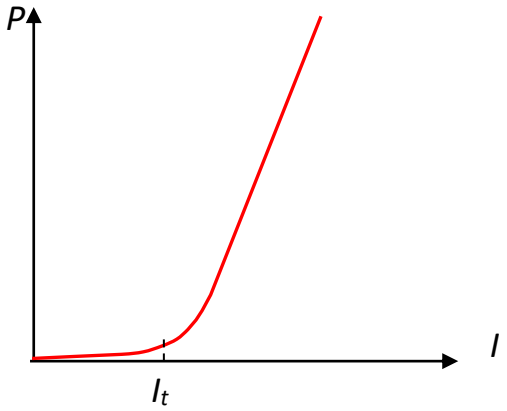


Fig. 1.6.16 – Watt-ampere characteristic of a semiconductor laser:
P – output optical power; I – pump current;
I_t – threshold value of the pump current

a degree of polarization, determined by the ratio of the polarization component intensity and the total intensity ($P = I_{\text{pol}} / I_{\text{sum}}$), with increase in the pump current is varying from dozens of percent to the values approaching unity (~80–90 %).

At the same time, critical variations are observed in emission spectrum of a laser diode depending on the pump current: from a wide continuous spectrum ($\Delta\lambda$ ~10 nm), associated with radiation of a laser diode in the LED mode (Fig. 1.6.17, curve *1*), to the lasing spectrum representing a set of longitudinal modes (Fig. 1.6.17, curve *2*). The generation is initiated at the specific modes with the frequency ν_m, which may be derived from the standing-wave formation condition $\nu_m = mc/2nL,$, where c – speed of light; L – laser resonator length; n – refractive index of the active medium; m – integer.

From here the frequency difference of the adjacent modes (spectral interval between the modes) is defined as $\Delta\nu_m = c/2nL$. For a semiconductor laser operating at the wavelength λ = 0.65 μm with the resonator base L = 250 μm, the active-medium refractive index n = 3.6, the intermode distance $\Delta\lambda_m \approx$ 0.2 nm. The radiation directivity diagram of semiconductor lasers is dependent on the transverse mode composition of radiation. In modern semiconductor lasers used in optical communications only the fundamental transverse mode TEM_{00} is excited.

The directivity diagram in this case has a single lobe (Fig. 1.6.18, *a*). As seen, the output radiation from a semiconductor laser is not symmetric. Divergence is below 20° in the plane parallel to the junction and is above 40° in

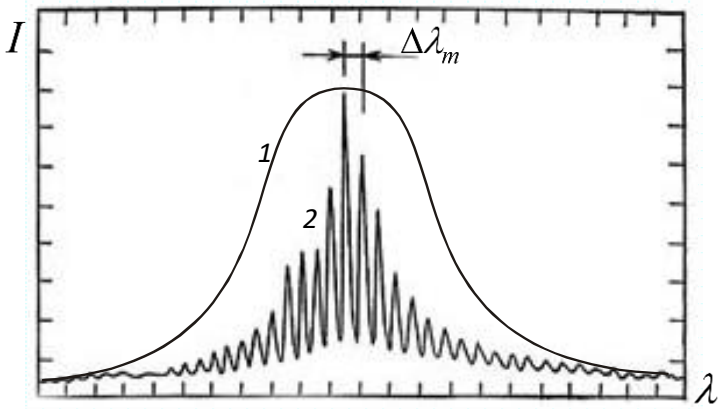


Fig. 1.6.17 – Emission spectra of a semiconductor laser operating in the LED mode (*1*) and in the lasing mode (*2*)

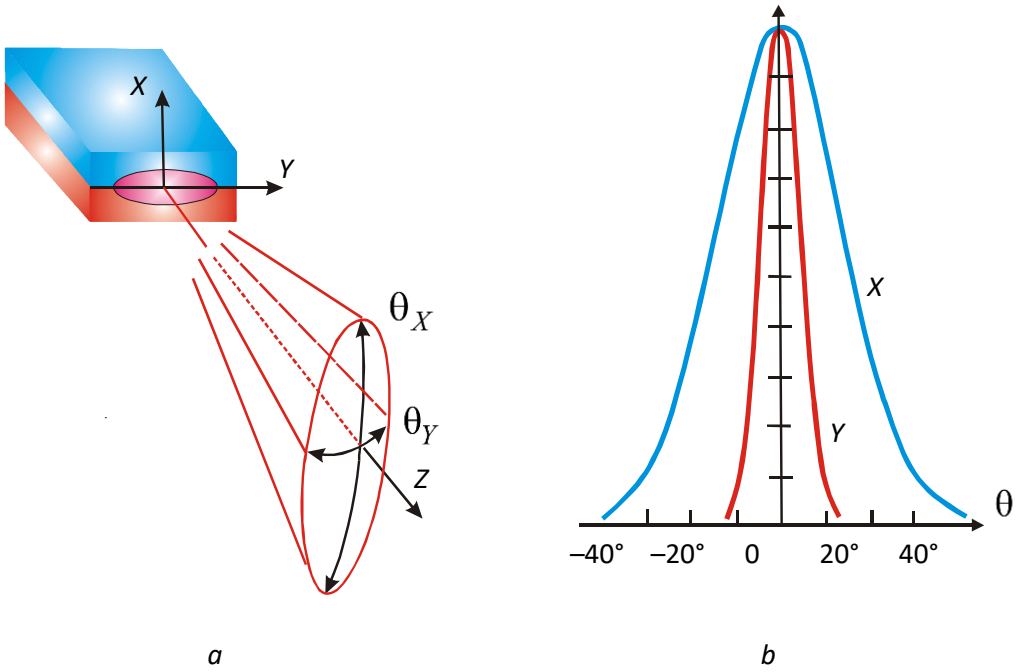


Fig. 1.6.18 – Radiation directivity diagram of a semiconductor laser in two orthogonal directions (*a*) and typical curves for the radiation power as a function of the angle in the mutually perpendicular directions X and Y (*b*)

the perpendicular plane (Fig. 1.6.18, *b*). This is caused by the differences in longitudinal and transversal sizes of the lasing region. In practice, directional radiation is formed with the use of special aspherical lenses enabling one to compensate for high asymmetric divergence at the output from a laser crystal.

It should be noted that the principal advantages of injection semiconductor lasers are their high efficiency (~50 %) and simplicity of optical signal modulation over a wide frequency range (up to several GHz) due to the electric current modulation. A great variety of advanced semiconductor materials offers

the generation in a wide spectral range from 0.3 to 30 μm (Tab.1.6.1). The radiation power comes to ~1–100 mW for continuous lasing and is up to 100 W in the pulsed mode. Owing to these properties, semiconductor lasers have found an extensive use in optoelectronics and fiber optics making it possible to develop different optical and optoelectronic devices for transmission, processing, storage, and imaging of information.

Table 1.6.1 Basic materials used in laser diodes:

Semiconductor	Radiation wavelength, μm	Semiconductor	Radiation wavelength, μm
ZnS	0,32	$In_xGa_{1-x}P$	0,60–0,91
ZnO	0,37	GaAs	0,83–0,90
$Zn_{1-x}Cd_xS$	0,32–0,49	lnP	0,90–0,91
ZnSe	0,46	$In_xGa_{1-x}As$	0,85–3,1
CdS	0,49–0,53	$InP_{1-x}As_x$	0,90–3,1
ZnTe	0,53	InAs	3,1–3,2
$CdS_{1-x}Se_x$	0,49–0,68	InSb	5,1–5,3
CdSe	0,68–0,69	PbS	3,9–4,3
CdTe	0,79	$PbS_{1-x}S_x$	3,9–8,5
GaSe	0.59	PbTe	6,4–6,5
$GaAs_{1-x}P_x$	0,62–0,9	PbSe	8,4–8,5
$Al_xGa_{1-x}As$	0,62–0,9	$Pb_xSn_{1-x}Te$	6,4–31,8

1.7. Laser technological systems

Laser material processing is one of the technologies associated with the up-to-date production level in industrialized countries. The use of laser material processing offers higher quality of the end products, the desired efficiency, and saving of material resources. Most widely spread are such technologies as laser welding, heat hardening, doping, surface cladding (fused deposition), cutting, labeling, dimensional processing, etc.

At the present time a large variety of laser equipment for material processing is available. The majority of manufacturers offer industrial laser facilities rather than individual industrial lasers. As a rule, they include external optical devices,

rotary tables, fiber-optical devices, manipulators, industrial robots for motion of work pieces during processing, and the corresponding software to realize the processes.

Among laser emitters used for material processing, the first were CO2 and solid-state lasers which have found extensive distribution. But the leading role belongs to solid-state lasers. They outperform gas lasers in manufacturability, economic efficiency, weight and dimensions, in possibility of radiation transfer by flexible optical fibers. Competitiveness of solid-state lasers has much improved since the advent of diode pumping. An important advantage is the possibility to control (comparatively easily) the energy and temporal structure of radiation by means of the controlled pumping and with the use of special gates.

In the last few years on the market of laser equipment the facilities based on high-power LED lasers have appeared. This equipment is used for heat treatment, welding, and cutting by means of fiber-optical systems.

The power of the latest fiber lasers is up to 20 кW. Their use enables one to obtain different temporal characteristics of radiation over the spectral range from 1 to 2 μm.

The potentialities of using industrial lasers for different kinds of processing depend on the spatial, energy, and temporal parameters of laser radiation.

1.7.1. Properties of laser radiation for industrial applications

The majority of laser material-processing technologies are based on thermal effect of radiation. One of the principal characteristics of continuous-wave lasers is their power; in the case of pulsed lasers – their peak power and average power that is dependent on the pulse length and on the repetition rate. When focusing radiation at the surface of the material under processing, we should take into consideration the power density and some other parameters of laser radiation including the following:

- radiation power density q;
- wavelength λ;
- pulse length τ;
- pulse repetition f;
- spatial characteristics of the radiation mode structure;
- beam divergence α.

Laser radiation power density.

Using a phenomenological theory, we can calculate the threshold power density, required for the surface heating up to the specified temperature T due to the continuous or pulsed action of laser radiation, in the following way:

$$q_{\text{имп}}^{\text{пор}} = \frac{(T-T_{\text{н}})k\sqrt{\pi}}{2A\sqrt{\alpha\tau}}\ ,\ q_{\text{непр}}^{\text{пор}} = \frac{(T-T_{\text{н}})k}{Ar_0}, \qquad (1.7.1)$$

where k – thermal conductivity coefficient; α – thermal diffusivity coefficient; $T_{\text{н}}$ – initial temperature; A – absorption factor ($A = 1 - R$ (R – reflection factor)); r_0 – cross-sectional radius of a laser beam; τ – laser pulse length.

To illustrate, knowing the threshold radiation power density of the work material, we can find the required threshold power level of laser radiation

$$P_{\text{пор}} = qS\ , \qquad (1.7.2)$$

where S – laser spot area at the lens focus ($S = 4\pi r_0^2$).

Laser radiation wavelength and angular divergence.

So that material can effectively absorb the energy of laser radiation, its wavelength should lie in the spectral region with a high absorptivity.

For example, for metals it is expedient to use lasers generating in the visible; for processing of dielectric materials such as glass – infrared (IR) lasers; for polymers – ultraviolet (UV) lasers.

Besides, one should take into account that the wavelength determines the laser action area d at the laser-beam focusing position on the processed surface. A size of the focused laser spot in the focal plane of a lens may be calculated by the following formula:

$$d = 1{,}22\lambda\frac{f}{D}\,M^2, \qquad (1.7.3)$$

where f – focal length of the lens; D – diameter of the initial (nonfocused) laser beam; M^2 – coefficient characterizing distinctions in the industrial laser radiation and the ideal Gaussian beam with minimal diffraction-limited divergence (M^2-parameter).

Proceeding from all the above, it is inferred: the shorter the wavelength and the lower the angular divergence of laser radiation, the smaller the spot to which it can be focused and hence the smaller the action area size. For visible radiation

the action area may be as small as several micrometers, for ultraviolet radiation – fractions of a micrometer.

Pulse length.

In studies of the physical processes proceeding in a material during laser radiation absorption it has been found: the shorter the pulse, the lower the thermal and deformation effect exerted on the processed material. With the use of short radiation pulses at a high power density, a small volume of metal is melted and evaporated before heat from the irradiation area has time to propagate within the processed material.

The physical principles of this process may be described within the scope of a two-dimensional evaporation model according to which laser radiation is first absorbed by free electrons (for metals). Thermalization of the electrons takes place with subsequent collisions for 10^{-15} s. Ions of the lattice which are much heavier than electrons fail to directly absorb optical radiation because they fail to keep pace with rapid oscillations of an electromagnetic field. But after collisions with the electrons gaining energy from the electromagnetic field, the lattice begins to heat up too. For temperature equilibrium, the number of collisions must be great and proportional to the ion and electron mass ratio. During this period of time, known as the electron-phonon relaxation time, a system acquires the macroscopic temperature *T*.

Despite the fact that a time of electron-phonon relaxation for different materials comes to $0.5 \div 100 \cdot 10^{-12}$ s, heating of the lattice until the onset of evaporation lasts for several nanoseconds. In this case a material remains in the melted state no more than 10 ns. Due to thermal conductivity, the material surrounding the irradiation area begins to warm up, this leading to its restructuring. The so-called heat affected zone (HAZ) is formed. In practice, this zone is always present in a submicron region down to the pulse lengths below 10 ps – even for ultrashort pulses this zone is not zero.

HAZ determines a depth of the melted layer that is approximately given by the formula

$$z \approx \sqrt{\alpha\tau}, \tag{1.7.4}$$

α – thermal diffusivity coefficient dependent on thermal and physical properties of a materiala, τ – laser pulse length.

The pulse length has a significant effect on mechanical stresses arising in the processed material close to the action zone $F \sim \sqrt{\tau}$ and on the vapor pressure that is inversely proportional to $\sqrt{\tau}$.

Pulse repetition rate.

Pulse repetition rate of laser radiation influences temperature of the work material and a character of its variations at the beginning and after the termination of every pulse. For example, after termination of the pulse, the average surface temperature is liable to lower but if the pulse repetition rate exceeds the minimum permissible value of $f_{\text{кр}} = \frac{\alpha}{30 r_0^2}$ these variations are not observed and the effect of laser radiation is dependent on the energy (power) of single pulses only.

As a rule, the pulse repetition rate is chosen higher than $f_{\text{кр}}$. Besides, an increase in the pulse repetition rate results in the improved output of a laser facility due to the increased processing rate.

Selecting a laser to be used in an industrial laser facility (ILF), one should take into consideration its operating characteristics (reliable performance; time to first failure; energy consumption; weight and dimensions) and price.

1.7.2. Optical systems for ILF

Optical systems of ILF serve for the laser radiation energy transfer to the processing place; for regulation of radiation parameters; for the formation of a light beam with a high power density; for visual radiation guidance to the action point; for the process control, and for assessment of the processing results. Optical systems used to transfer and transform high-power working laser radiation are called the power optical systems (OS). According to their functionality, these systems may be subdivided into (a) focusing; (b) scanning; (c) projecting; (d) distributive. Power OS comprise focusing, refractive, and reflecting elements. In the visible and near-IR regions totally reflecting prisms and interference mirrors with multilayer dielectric coatings are used to lower the radiation loss. The mirrors used in ILF operating at $\lambda_{\text{и}}$ = 10.6 μm have Al and Au coatings characterized by high reflection factor R and oxidation stability.

In power OS including high-power lasers, especially operating in the continuous-wave mode, the requirements to optical elements are enhanced. For radiation with λ = 0.69 and 1.06 μm, the majority of the optical glass kinds have minor absorption factors and are used up to $q = 10^3$ W/cm^2. At the wavelength λ

= 10.6 μm they are opaque; lenses are manufactured of monocrystalline KBr, NaCl, GaAs, Ge, etc.

Focusing optical elements.

Depending on the duty of a specific facility, the requirements to the characteristics of laser radiation in the processing zone are formed differently. To illustrate, the operations of perforation, cutting, welding by deep melting and the like necessitate holding of the preassigned power density in a particular volume of the work material. Most often this is attained by radiation focusing. In this case it is important to know the diameter of a laser beam in the focal plane that is given by the expression

$$d = Ftg\Theta = F\Theta, \qquad (1.7.5)$$

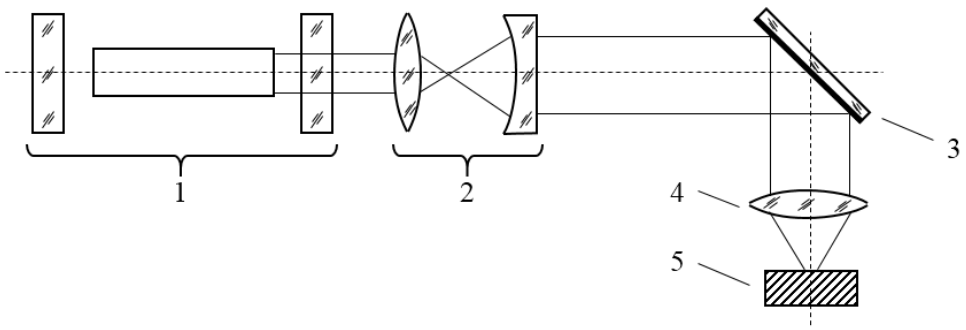
where F – focal length of a focusing lens; Θ – beam divergence of laser radiation.

As seen from the formula, to lower the value of d, we can lower F but this is not always convenient as the front lens is often damaged by the erosion products from the action zone. To have the focused radiation spot d on the order of a few micrometers, a telescopic system (see Fig. 1.7.1) is positioned before the long-focus objective at a rather great distance. In this case the diameter of a laser beam is determined as

$$d = \frac{F\Theta}{\Gamma}, \qquad (1.7.6)$$

where Γ - magnification (multiplicity) of a telescopic system.

This means: the greater magnification of a telescopic system, the smaller the focused beam diameter.

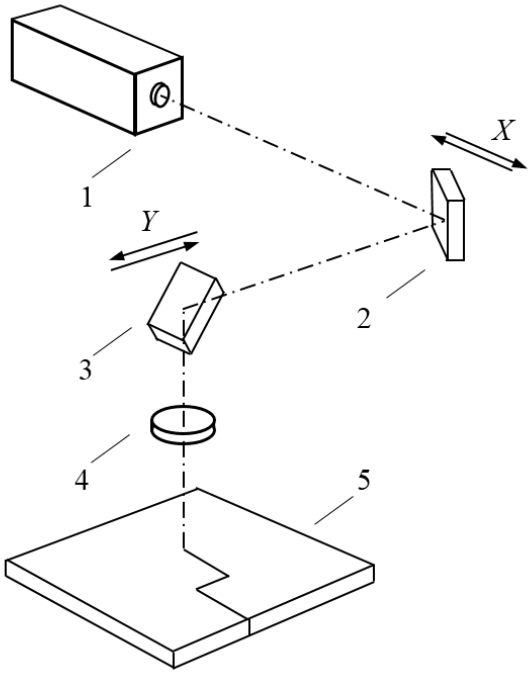


1 – laser emitter, 2 – telescopic system, 3 – turning mirror, 4 – focusing objective, 5 – surface of processed piece.

Fig. 1.7.1 – Optical scheme with the use of a telescopic system to decrease the beam diameter in the focal plane of the objective lens

Scanning optical system.

Profile processing (seam welding, cutting) or processing according to the specified positions is associated with a necessity to coordinate the beam and piece motion. This is easily realized when a piece moves with respect to the beam. However, for modern facilities with high processing rates, motion of the beam is more rational, making easier the control and improving the work precision. Fig. 1.7.2 shows schematic of a scanning system. Motion of the beam along an arbitrary path is realized by means of a system of movable mirrors, moving by the corresponding coordinates. In the *X* direction both mirrors 2 and 3, and objective lens 4 move simultaneously, whereas in the *Y* direction motion is possible only for mirror 3 and the objective lens.



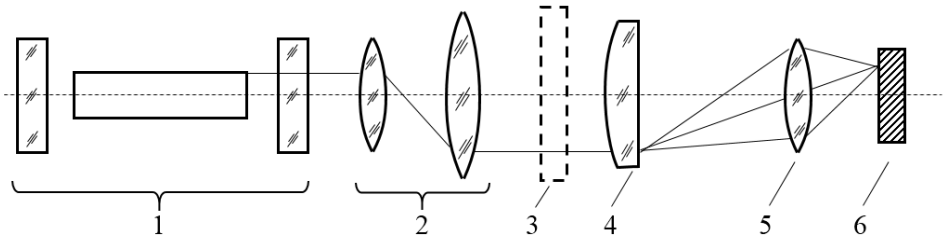
1 – laser emitter; 2, 3 – movable mirrors; 4 – focusing objective lens, 5 – processed piece

Fig. 1.7.2 – Schematic of a scanning system for profile processing with the help of two movable mirrors

Projecting optical system.

In microelectronic production there is a need for forming patterns on thin films deposited onto the dielectric substrate. Along with conventional procedures

(masked deposition of thin-films, photolithography), there are possibilities to improve the productivity. To this end, several variants of optical systems have been designed for projection of the mask image on the work zone with the down scale much greater than unity. Fig. 1.7.3 demonstrates schematic of a projecting optical system. In such a system, when a diameter of the objective lens entrance pupil 5 is smaller than a size of mask 3, collective lens 4 is used to exclude vignetting. Having no effect on the image scale, this lens directs the bundles of rays from each point of the mask to the objective. Telescopic system 2 is used to expand radiation to the size of mask 3. Processing of the whole mask area (photomask) in this system takes a single pulse.



1 – laser emitter, 2 – telescopic system, 3 – mask, 4 – collective lens, 5 – focusing objective, 6 – surface of processed piece

Fig. 1.7.3 – Schematic of a projecting optical system

Distributive and special optical systems.

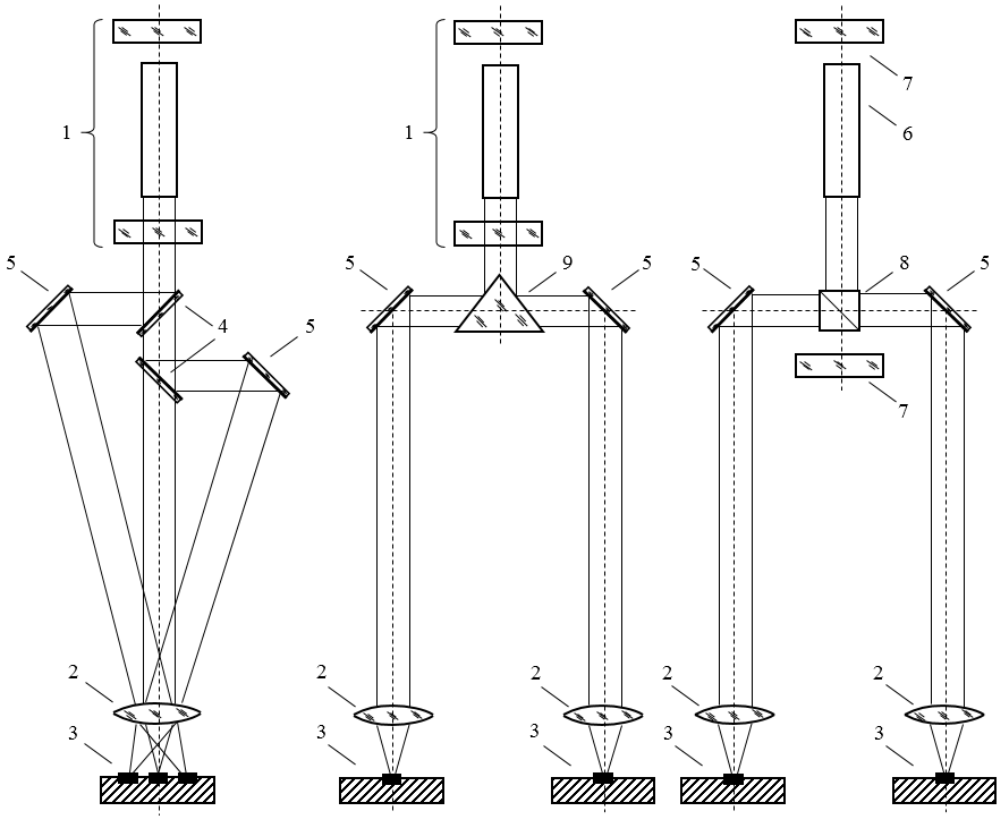
In laser technology it is often required to distribute radiation among several work positions (beam splitting). Ordinary, this is associated with the need to improve productivity of the facility. Fig. 1.7.4 illustrates several possible variants of such a distribution. In the variant shown in Fig. 1.7.4, *a* this is attained with the help of beam-splitting mirrors 4. With this scheme, the number of work branches is determined by the emitter energy and by the mirror loss, the cross-sectional form of a beam being retained. Fig. 1.7.4, *b* is associated with splitting by means of prism 9. The variant in Fig. 1.7.4, *c* is associated with two-way radiation coupling directly from a resonator by means of beam-splitting cube 8.

To form beams with the desired spatial and energy characteristics of laser radiation, one needs special optical systems.

Beams with the elliptical cross-section are formed by substitution of a cylindrical-elements objective for the spherical one.

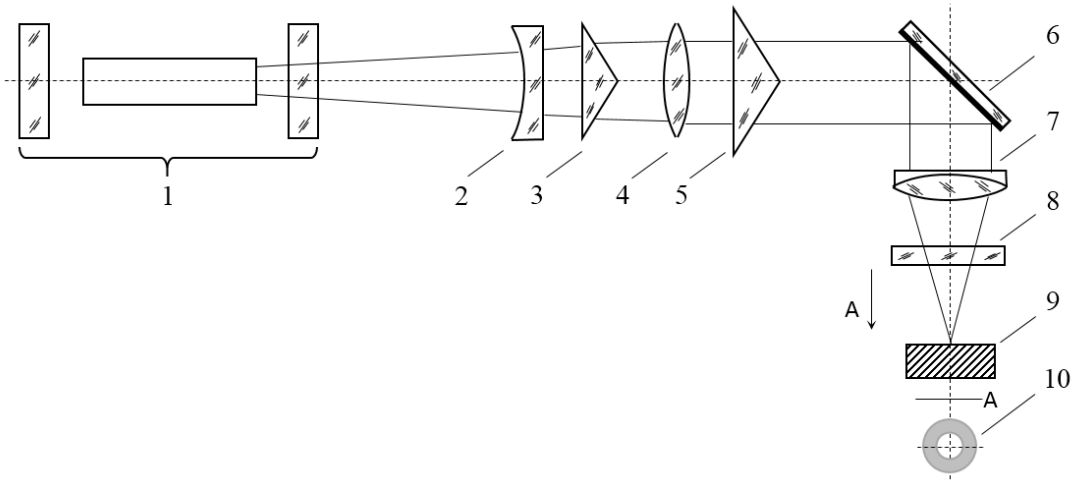
An optical system with conical optical elements is used to form the ring-section beams with a variable diameter and width of the rings. An optical scheme of such a system is shown in Fig. 1.7.5.

Laser 1 generates a cylindrical beam with a continuous circular section. This beam passes negative lens 2, first conical lens 3, and positive lens 4 to become redistributed as a circular beam due to refraction from the conical surface of lens 3. Next this beam is directed to the second conical lens 5, the base angle of which is selected so that, after refraction, the beam propagation can be parallel to the optical axis of the system.



1 – emitter, 2 – focusing lens, 3 –processed piece, 4 – beam-splitting mirror, 5 – turning mirror, 6 – active element of the emitter, 7 – full-rate resonator mirrors, 8 – beam-splitting cube, 9 – prism

Fig. 1.7.4 – Schemes for distribution of radiation among several work positions



1 – emitter, 2 – negative lens, 3, 5 – axicons, 4 – positive lens, 6 – rotary interference mirror, 7,8 – focusing objective with protective glass, 9 – processed surface, 10 – laser beam form on the processed surface

Fig. 1.7.5 – Schematic of an optical system for the formation of ring-section beams

The incidence angle α_5 on the second conical lens 5 for a parallel laser beam is given by the expression

$$\alpha_5 = \Theta_3(n-1)(1-d_2\Phi_4), \tag{1.7.7}$$

where Θ_3 – base angle of conical lens 3; d_2 – distance between conical lens 3 and positive lens 4 of a telescopic system; Φ_4 – power of lens 4; n – refractive index of the material used to manufacture conical lenses.

Diameters of the ring profiles are smoothly varied by displacement of conical lens 3 along the optical axis of the system, with a change in its position relative to one of the telescopic-system components (Fig. 1.7.5). A diameter of the ring in the processing plane can be calculated by the formula

$$D \approx 2f_{об}\Theta_3(n-1)(1-d_2\Phi_3), \tag{1.7.8}$$

where $f_{об}$ – focal length of objective 7. The value of d_2 is varying within the following limits:

$$d_{2min} \le d_2 \le d_{2max} \tag{1.7.9}$$

Values of d_{2min} and d_{2max} depend on the telescopic system design.

1.7.3. Structure of industrial laser facilities for metal and alloy processing

Schemes of industrial facilities depend on the type and field of their application. Most widely used material-processing systems form two groups:

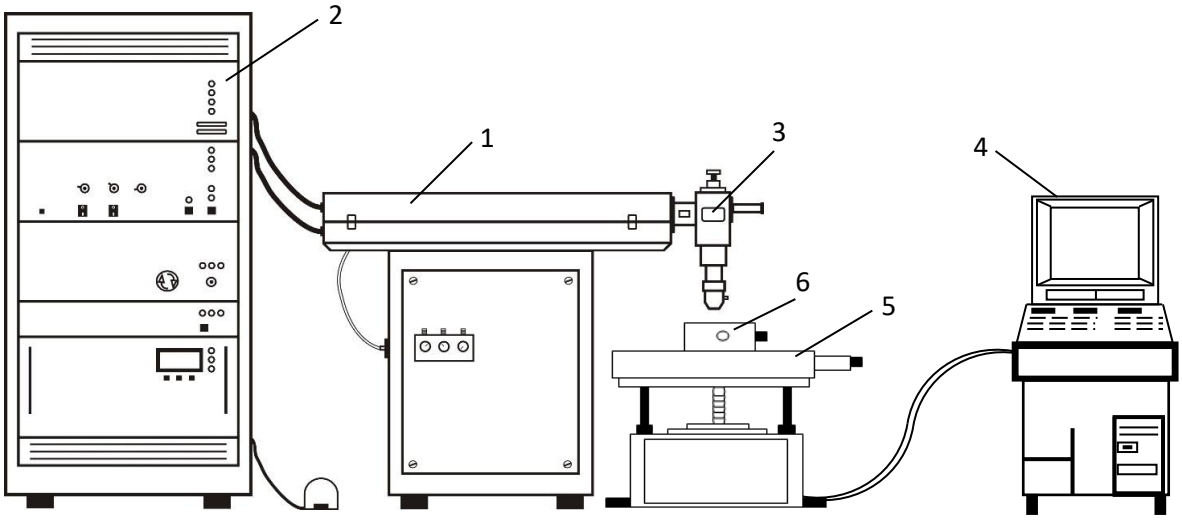
- For processing of flat sheet materials;
- For processing of three-dimensional articles.

In the first case xy systems are used, and in the second case – multiaxis systems are used to offer coordinated motion of a laser beam and of a workpiece. It is important that with all processing methods the major requirement is to perform the processing procedures at normal incidence of the beam on the work surface − for the best use of the energy of laser radiation. At the same time, it is essential to retain a distance between the focusing objective and the workpiece surface. Position of a laser beam on the workpiece surface may be specified by:

- Displacement of the workpiece with respect to the stationary beam;
- Motion of the beam with respect to the stationary workpiece;
- Relative displacement of the beam and of the workpiece.

The processing method is chosen with due regard for processing precision, workpiece dimensions, mounting and fastening technique, etc. As a rule, an industrial laser facility (ilf) is provided with a manipulator for the products or optics; auxiliary equipment; common control system for technological operations and software offering their realization.

Let us consider a structural scheme of ilf that is based on a solid-state laser (see fig. 1.7.6).



1 – laser emitter, 2 – power and control unit with a cooling system, 3 – visual control system with a focusing objective for the processing zone, 4 – PC, 5 – workpiece linear translation unit, 6 – workpiece spinner.

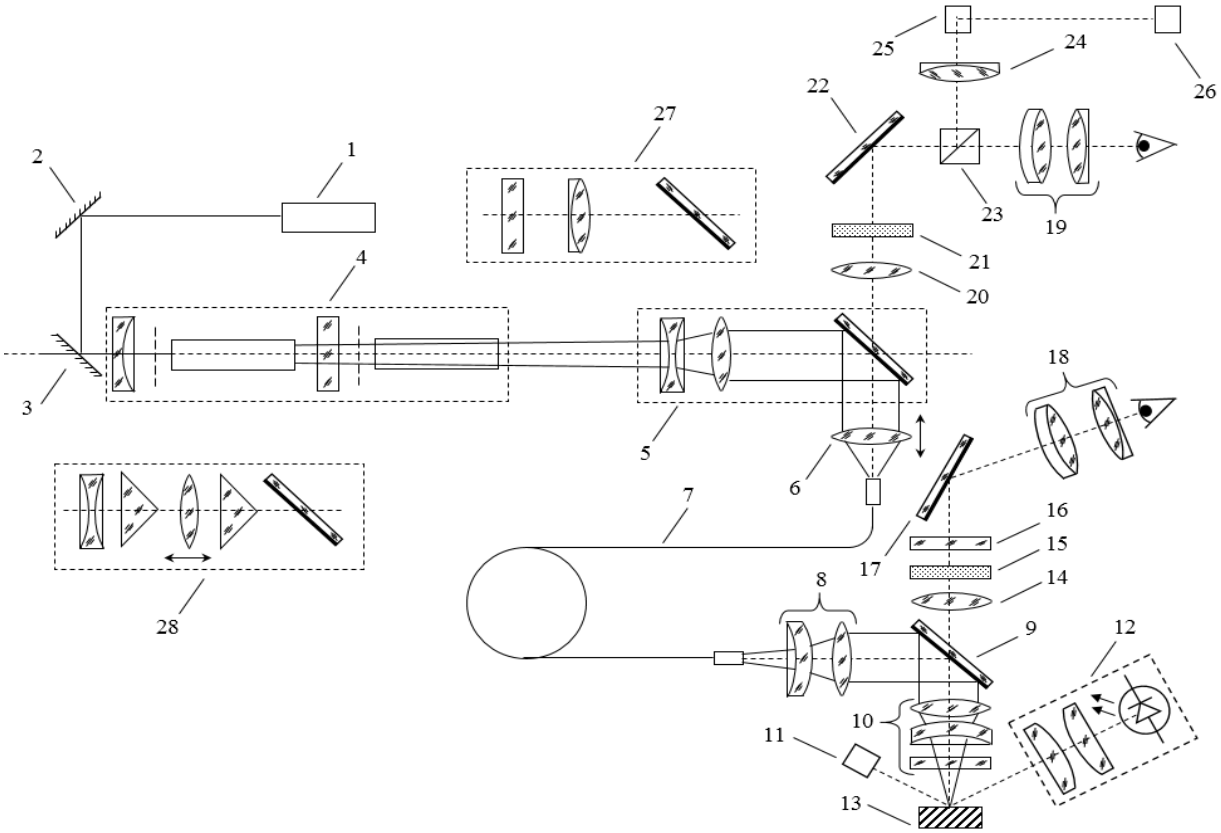
Fig. 1.7.6 – Structural scheme of ILF for metal and alloy processing

Laser emitter 1 coupled with a power and control unit forms a beam with the desired spatial, energy, and temporal characteristics. By optical system 3 this beam is focused and directed to the workpiece, the power density level being sufficient for execution of the processing procedure.

Besides, with the help of optical system 3, the workpiece position may be controlled visually and progress in operations may be monitored. Processing of bulky articles at some distance from a laser facility is realized by means of a fiber-optical adapter with a focusing objective and visual tracking channel.

The X-Y stage, with workpiece linear-translation unit 5 and spinner 6, that is intended for high-precision displacement of the processed article with respect to the focused laser beam is controlled by PC coupled with a control system for the characteristics of laser radiation, with a detector system for control of the process parameters, and with a laser power supply.

An optical scheme of ILF with a fiber-optical adapter for material processing is shown in Fig. 1.7.7.



1 – laser; 2,3 – turning mirrors; 4 – laser emitter; 5 – telescopic system with a turning interference system; 6 – objective used for radiation coupling into optical fiber; 7 – optical fiber; 8 – optical system for coupling of laser radiation from optical fiber; 9 – rotary interference mirror; 10 – focusing objective with protective glass; 11 – gas feed system; 12 – illumination system for processing zone; 13 – processing plane; 14, 20 – matching lenses; 15,21 – liquid-crystal gates; 16 – light filter; 17,22 – rotary mirrors; 18, 19 – monoculars; 23 – beam-splitting cube; 24 – matching lens for camera; 25 – camera; 26 – monitor; 27 – optical system for the formation of elliptical-section beams; 28 – optical system for the formation of ring-section beams

Fig. 1.7.7 – Optical scheme of ILF with fiber-optical system used to transfer laser radiation

This scheme includes a pulsed solid-state YAG laser, a fiber-optical adapter to transfer laser radiation to the material processing zone, a system for the formation of laser beams having circular, elliptical, or ring cross-sections with the adjusted ring diameter and width.

Pilot-laser 1 and turning mirrors 2, 3 are intended for alignment of the optical elements forming a part of the laser facility. Laser emitter 4 operates according to the oscillator-amplifier scheme including resonator mirrors and active elements of the oscillator and amplifier. The Galilean telescopic system with a turning interference mirror is used to control divergence of laser radiation and to couple radiation, with the help of focusing objective 6, into optical fiber 7.

System for radiation coupling from optical fiber 8 and turning interference mirror 9 direct laser radiation to focusing objective 10.

Protective medium is introduced into processing plane 13 by gas feed system 11. System 12 based on light-emitting diodes illuminates processing plane 13.

Visual control system for the process includes monocular 18, rotary mirror 17, light filter 16, matching lens 14. Liquid-crystal gate 15 shields operator's eyes from the radiation effect.

The optical scheme of this facility includes a system for focusing and fine alignment of laser radiation to the waveguide end that comprises: monocular 19, beam-splitting cube 23, rotary mirror 22, liquid-crystal gate 21, and matching lens 20. The waveguide end and processing plane are displayed by monitor 26 with the help of beam-splitting cube 23, matching lens 24, and camera 25. Optical system 27 is used to form laser beams with the elliptical cross-section. The ring-section beams are formed by optical-mechanical system 28.

Such an optical scheme may include Galilean and Kepler telescopic systems comprising negative and positive spherical lenses and two positive spherical lenses, respectively.

1.8. Laser processing of material

Laser processing is one of the industrial methods to process different materials, capable to sustain competition with mechanical, electro-erosion, ultrasonic, electronic or other types of material working. By their efficiency, processing quality and precision, laser facilities often outperform other technological equipment, sometimes being the only means to attain the production objectives.

Technological potentialities of lasers are great including such operations and procedures as hole drilling in metal and dielectric materials; welding of ferrous and nonferrous metals, glass, plastic, composite materials; heat treatment of surfaces and material hardening (quenching of ferrous and nonferrous metals or their alloys); dimensional cutting (through and controlled thermocracking, scribing) of semiconductor, dielectric, glass materials, etc; production of multilayer thin-film structures.

The principal advantages of laser material processing are as follows: possibility to process materials with different physical and mechanical properties in various media; high precision and efficiency of processing; enclosed area of

thermal effect; no mechanical action on material; high processing efficiency; full automation of the processing procedures.

1.8.1 Laser welding of metals and alloys

Laser welding is used to attain permanent connection of the parts by the local burn-off of metals at the adjoining surfaces. Adhesion is based on interatomic interactions. Laser flux is a heat source.

On interaction with an absorbing medium, laser radiation is partially reflected from the surface and partially penetrates the material, being absorbed by it. Variations of the flux density q, i.e. the energy quantity falling at the unit surface in a unit time, in an absorbing medium is described by the Bouguer-Lambert law as follows:

$$q(x) = q_0 A \exp\left(-\int_0^x \alpha(x) dx \right)$$

where: q_0 –density of the flux incident on a material;

A – material absorptivity;

$A = 1 - R$ (R – reflection factor);

$\alpha(x)$ – light reflection factor in a medium.

The coordinate x is measured from the surface deep into the material. The formula corresponds to the case of normal skin effect and may be applied to different materials. Specific values of the involved quantities A and α, as well as the mechanisms of light absorption and transition to heat, differ considerably for various materials. In metals light quanta are absorbed predominantly by the conduction electrons dissipating the absorbed energy for thermal oscillations of the grating during the period of time $\tau_{ei} \approx 10^{-11} \div 10^{-10}$ s.

This process proceeds in a layer with the thickness $\delta \approx 10^{-6} \div 10^{-5}$ cm.

The spatial distribution of the absorbed luminous flux in metal at the optical frequencies corresponds to the Bouguer-Lambert law. This is due to the fact that for optical frequencies a path covered by the electron in a metal during a single period of field oscillations is considerably less than the material penetration depth $\delta = 1/\alpha$. As a result, the Ohm law is satisfied and the skin-effect is normal.

In this case, $\alpha(x) = \alpha = const$

$$q(x) = q_0 A \exp(-\alpha x),$$

the values of α and A are determined by the expressions

$$\alpha \approx \frac{4}{c_0}\sqrt{\frac{\pi \cdot n_0 l_0^2}{m_0^*}},$$

$$A \approx \sqrt{\frac{m_0^*}{\pi \cdot n_0 l_0^2}}\nu',$$

where l_0 and m_0^* – electron charge and effective mass;

n_0 – concentration of free electron in a metal;

ν' – frequency of the collisions associated with a change of the pulse;

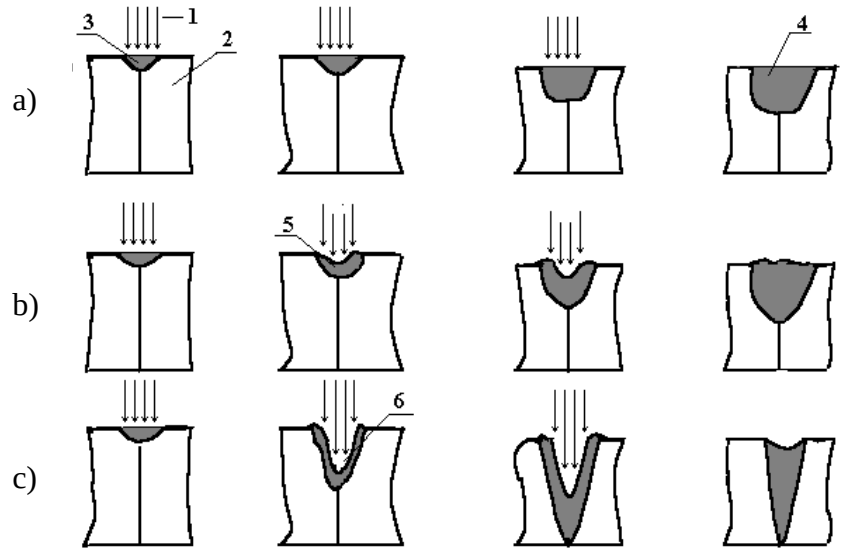
c_0 – speed of light in free space.

For the majority of metals, in the optical wavelength range the reflectivity $R = 1 - A$ (from 70% to 95%) and the absorption factor $\alpha \approx 10^5 \div 10^6$ см$^{-1}$ is high.

High power density on the surface of the welded components offered by laser sources is responsible for high heating rate that facilitates welding of metals, featuring high heat conduction (copper, silver) and high melting temperature (tungsten, tantalum, molybdenum). In the process of metal welding a spot of laser radiation on the surface of welded workpieces is the surface heating source because radiation is absorbed by a thin surface layer with the thickness about several hundredth of a micron. When the power density is $10^5 \div 10^6$ W/cm^2 and duration is about $10^{-3} \div 10^{-2}$ s, the in-depth heat transfer in the welded materials is mainly due to heat conduction.

The development of melting processes is shown in Fig. 1.8.1 for different radiation power densities. Note that the melting zone is close to the spherical form.

In the majority of the cases of laser welding one can observe strong evaporation of metal. Under the effect of vapor, the surface of a weldpool is sagging. A site of this surface directly absorbing the radiation energy is sinking (Fig. 1.8.1, b).



a) – melting due to the heat conduction, $q = 10^5 \div 10^6$ W/cm^2, b) – crater of molten metal, $q = 5 \cdot 10^5 \div 5 \cdot 10^6$ W/cm^2, c) – deep melting, $q = 10^6 \div 10^7$ W/cm^2

1 – laser radiation; 2 – welded components; 3 – molten metal; 4 – solidified pool crater; 5 – penetration volume; 6 – temporary opening due to evaporation

Fig. 1.8.1 – Development stages of the melting process at different radiation power densities

When the surface tension of molten metal still prohibits its splashing, after termination of the radiation effect, nonsolidified metal fills the formed crater. Because of sinking of the weldpool, a melting depth increases as compared to the case of heating without appreciable evaporation. The form of melting zone becomes conical (Fig. 1.8.1, b, c).

The increased melting depth on strong evaporation from the surface of a weldpool is due to mixing of the upper layers heated to the highest temperature and of the cooler lower layers of molten metal caused by inhomogeneous heating within the focused laser beam.

When the power density at the beam center increases up to $5 \cdot 10^6 \div 5 \cdot 10^7$ W/cm^2, a narrow deep crater is formed in the weldpool and a metal in this crater is partly evaporated and partly displaced out to the weldpool periphery (Fig. 1.8.1,c).On termination of the pulse, a temporary opening is filled with metal molten at the beam periphery, where the power density is insufficient for strong evaporation. The effect of deep melting is enhanced because of the spike structure of pulsed solid-state lasers as, in time of a single spike, the power density at the center of a laser beam approaches $10^7 \div 10^8$ W/cm. Variations in the energy characteristics of laser radiation make it possible to realize welding by different melting mechanisms which are dependent on the properties of welded materials

and on the joint type.

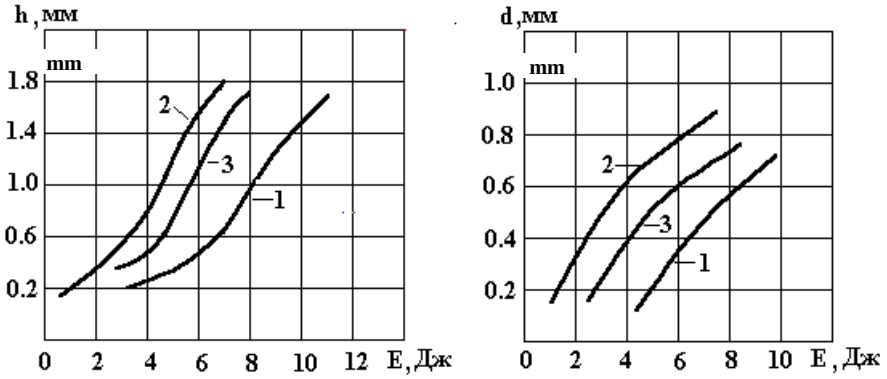
Some part of the radiation flux affecting the welded components is reflected. The reflection factor of all metals is growing with the wavelength of laser radiation. Tab. 1 lists reflection factors of several metals.

Table 1.8.1 – Reflection factors of several metals at different radiation wavelengths

Metal	Wavelength, μm		
	0.7	1.06	106
Aluminum	0.87	0.93	097
Chromium	0.56	0.58	0.93
Copper	0.82	0.91	0,98
Nickel	0.68	0.75	0.95
Silver	0.95	0.97	0.99
Steel	0.58	0.63	0.93-0.95

In the process of heating the absorptivity of many metals is markedly increased with a growth of temperature. To illustrate, silver and copper subjected to irradiation at the wavelength 1.06 μm exhibit a two times increase at heating from room to melting temperature, whereas steel reveals no significant variations in the optical characteristics in the same temperature interval.

Fig. 1.8.2 demonstrates the depth and diameter of the melting zone as a function of the radiation energy for some metals in the case of a solid state laser with a constant pulse length and beam diameter.



Pulse length 4 ms; beam diameter 0.5 mm

1 – copper; 2 – nickel; 3 – molybdenum

Fig. 1.8.2 – Depth (a) and diameter (b) of the melting zone as a function of the radiation energy (J)

It should be noted that an increase in the length of a laser pulse leads to better

removal of undissolved gases from the weldpool, lowering the possibility for porosity after hardening of the metal. Most convenient is to use for welding the trapezoid or triangular form of pulses which have a relatively steep leading edge and a low-angle trailing edge.

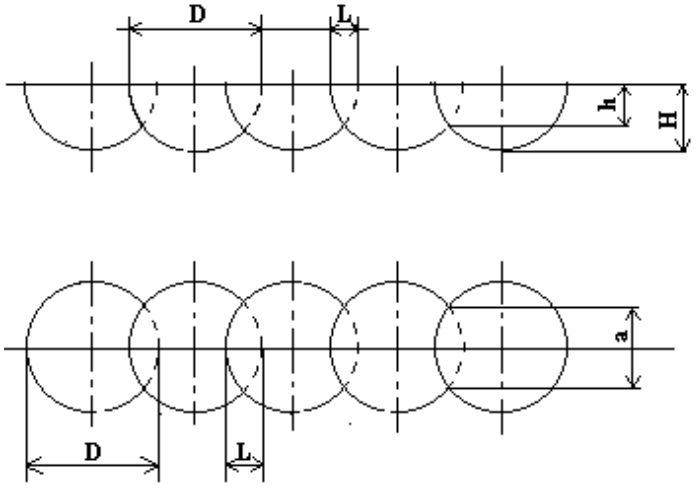
During the first part of a pulse most of metal is molten − the deep-melting regime is realized. During the period of a relatively slow decrease of the intensity, at the second part of the pulse, the liquid phase is increased due to melting of metal near the rims of the weldpool. Delay in the trailing edge contributes to filling of the pit.

A solid weld on laser welding is formed by overlapping of laser spots on the surface of welded material. To calculate the welding rate, we introduce the overlap factor as follows:

$$K = \frac{L}{D},$$

where L – overlap length of welding laser spots, D – laser spot diameter.

The geometrical parameters of weld in the process of pulsed laser welding by a circular beam are shown in Fig. 1.8.3. As weld point on laser welding is spherical or conical in form, the weld depth h is determined by the overlap factor k. Selection of the overlap factor is dictated by the minimal melting depth required, that is essential for strength and airtightness of the weld. To ensure the greatest strength, the weld depth should be close to the maximal melting depth and the overlap factor should be approaching 1.



h – weld depth, H – maximal melting depth, a – weld width

Fig. 1.8.3 – Schematic diagram of the weld formation and its characteristics

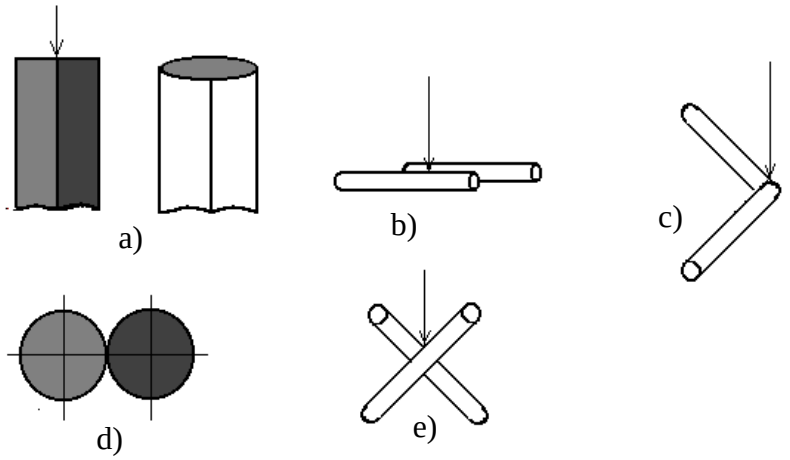
Welding rate in this case is low. It is given by the formula

$$V = f \cdot D(1 - k),$$

where f – pulse repetition rate.

The principal requirement in the process of material welding is leak (or air) tightness of weld.

Fig. 1.8.4 shows the main types of joints.



a – butt welding: b – lap welding, c – angular joint, d – longitudinal joint,
e – cross-wire connection. Directions of the incident beam are shown by arrows.
Fig. 1.8.4 – Types of welded joints

Fig. 1.8.5 presents the appearance of a weld: its depth is varying within the range 0.1 – 0.3 mm, overlap factor is ranging within 0.3-0.5.

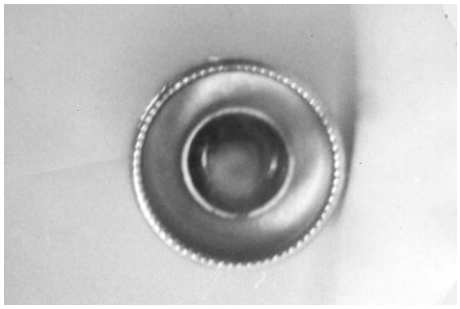
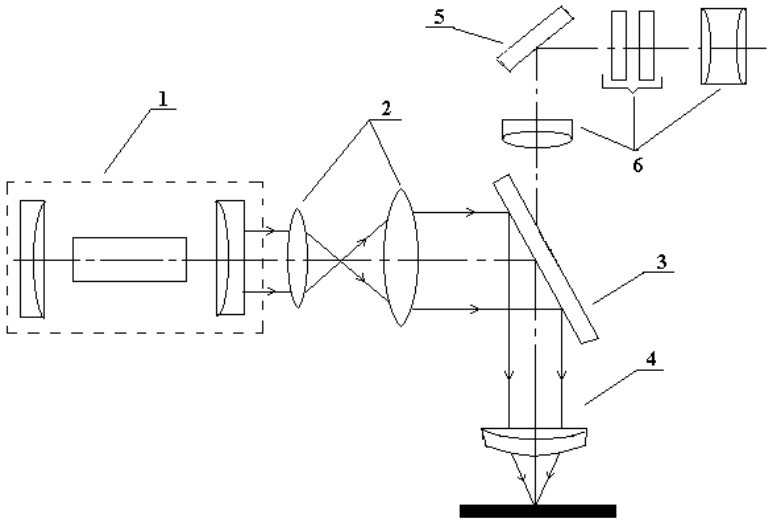


Fig. 1.8.5 – Weld appearance

The indicated joint types 4 may be realized using a setup with a neodymium glass laser, the optical scheme of which is shown in Fig. 1.8.6.



1 – neodymium glass laser; 2 – telescopic system; 3 – rotary interference mirror; 4 – focusing objective; 5 – mirror; 6 – visual observation system

Fig. 1.8.6 – Optical scheme of a setup for laser welding

1.8.2. Laser quenching of metals

Quenching process is aimed at improvement of hardness and wear resistance of materials due to variations in their structure. Quenching of metals necessitates their heating to a particular temperature, subsequent exposure, and rapid cooling.

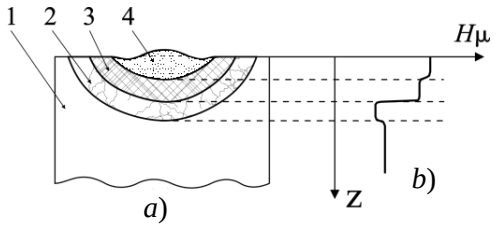
Quenching of metals and alloys by laser radiation is based on the local heating of the surface site with subsequent cooling of this site at the supercrital rate due to heat removal to the inner layers of metal. These conditions offer high (10^6 degrees/s) heating and cooling rates for the processed surface sites. Laser quenching results in better hardness of the surface layer, its thickness is dependent on the radiation intensity and exposure time and may be varied. The quenching depth is growing with the intensity. But the component as a whole remains nonquenched and plastic because only the surface layer is hardened. This is impossible with ordinary quenching procedures. Quenching may be implemented in the continuous-wave or pulsed-periodic mode. Quenching of a particular area of the surface is realized by laser-beam scanning or by motion of the processed component.

Quenching of iron-carbon alloys begins with heating to the temperature associated with stable existence of austenite. The properties of the hardened zone depend on the rate and temperature of heating, time in the heated state, cooling principle, and on the initial structure (preliminary thermal treatment and mechanical processing).

Quenching consists in diffusionless transformation on rapid cooling of the face-centered cubic (fcc) lattice of austenite to the distorted body-centered cubic (bcc) lattice of martensite. The characteristic features of martensite are high hardness and strength, on the one hand, and low plasticity and tendency to brittle facture, on the other hand. Martensite, as compared to other structural components of steel, and especially to austenite, has the greatest specific volume that is mainly responsible for high internal stresses which on hardening cause deformation of the goods or even incipient cracks.

The transformation of austenite to martensite is not completely finished. Because of this, in hardened steel some quantities of the residual austenite are present along with martensite. On cooling to temperature below A_1, austenite becomes metastable and is transformed to more stable structures. When the cooling rate is low, pearlite is formed; at a higher rate – sorbite and then troostite; finally, at a particular cooling rate (that is called the critical hardening rate) the pearlitic decomposition of austenite becomes impossible and the whole austenite is supercooled to the point $M_н$. The data concerning temperature intervals of the phase transitions in the process of continuous cooling and those concerning the formed structural components are given on the so-called thermokinetic diagrams.

The cross-sectional view (Fig. 1.8.7) of the metal surface hardened by laser processing reveals a number of the primary zones: flash-off zone (zone of quenching from the liquid state); quenching/hardening zone, tempering zone, initial structure of material.



1 – starting metal, 2 – tempering zone, 3 – quenching zone, 4 – zone of quenching from liquid state

Fig. 1.8.7 – Cross-section of the laser processed zone (a) and in-depth microhardness distributions in the processing zone (b)

Some of these zones may be omitted, for example, tempering zone on quenching of the annealed metal. Each of these zones may include several layers, differing in their microstructure, elemental composition, phase ratios, etc.

The solid-state steel hardening zones are inhomogeneously distributed over the cross section. In depth one can find, apart from martensite, elements of the initial structure: ferrite (for hypoeutectic steel) and cementite (for eutectic steel); martensite and residual austenite is formed closer to the surface after cooling of homogenized austenite. Recrystallization is accompanied by the grain size refinement and austenite homogenization, especially when it takes place for sufficient time without great overheating, i.e. when kept at a temperature above $M_н$. Dissolution of excess cementite on overheating of hypereutectic steels results in grater fractions of residual austenite and in lowering of microhardness as compared to the zone of optimal heating that, along with martensite, contains undissolved carbides

As laser radiation affects a small-volume zone, the changes occurring in depth of the laser impingement point (LIP) are best described by the microhardness. A depth of LIP depends on the following factors: material under processing, preliminary thermal treatment, radiation type, regime of laser processing. Microhardness after laser quenching is dependent on the content of carbon in steel: the higher the content the higher the microhardness.

With the use of pulsed lasers, of great importance are such parameters as overlap factor and processing interval. A choice of the parameter is influenced by sizes of the hardened and nonhardened zones, roughness of the hardened surface, thickness of the homogeneously hardened layer, and process efficiency.

Considering which of the above-mentioned parameters is limiting, one can set the maximum possible overlap factor, for which the limiting parameter will be within the optimum range. The overlap factor is chosen for the limiting parameter.

It is recommended to choose some optimal value of the overlap factor for which all the limiting parameters remain within the permissible limits. Most important for the hardened steel formation process are such energy parameters of radiation as pulse energy, focusing spot diameter and pulse length which determine the radiation power density. These parameters are used for operation of pulsed industrial laser facilities.

Fig. 1.8.8 shows a schematic diagram for processing of a component by pulsed laser radiation. Processing may be effected when spots are single or several spots forma a path with the overlap factor 50%.

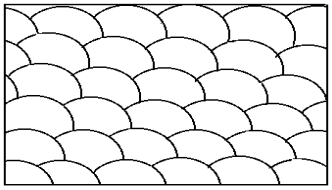


Fig. 1.8.8 − Schematic diagram of the surface processing by radiation of a pulsed laser

Fig. 1.8.9 shows the depth *z* of the hardened steel layer as a function of the radiation energy *E* on double irradiation of a sample in the air and in argon. Dashed lines show sections of the curves associated with the energy of radiation at which the surface geometry changes due to melting. A minimal energy associated with melting is called the critical energy.

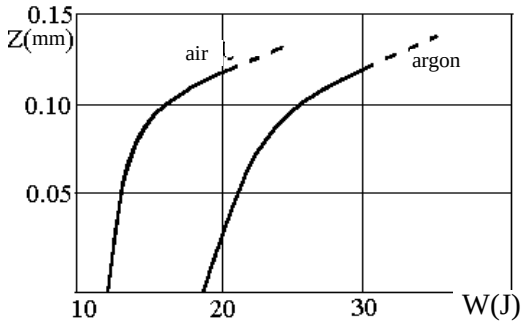


Fig. 1.8.9 – Depth of the hardened layer of steel as a function of the laser radiation energy

The critical energy on processing in the air comes to 20 J at the corresponding power density 0.8×10^4 W/cm^2; in argon − 31 J at the power density 1.2×10^4 W/cm^2.

The in-depth microhardness distribution z of a steel sample hardened in argon is demonstrated in Fig. 1.8.10.

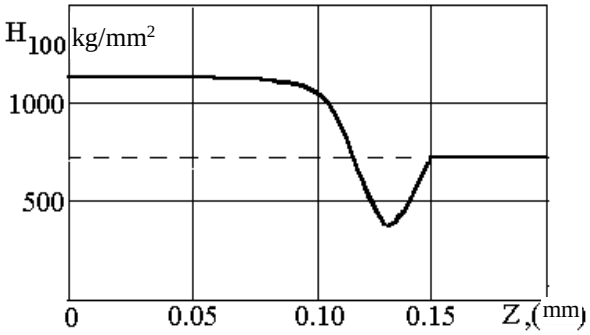


Fig. 1.8.10 – In-depth microhardness distribution for a sample of steel (mm)

The initial microhardness is indicated by a dotted line. The power density is 1.2×10^4 W/cm^2.

Microhardness near the surface and at a depth up to 0.112 mm comes to 1080 kg/mm^2. Between the hardened layer and the bulk material there is a zone of high-temperature tempering about 0.02-mm in width. As seen, microhardness in the tempering zone is lower than that of the starting material.

On hardening of the lengthy surface sites, single spots of hardening are separated by narrow bands with minor microhardness. These bands appear at the points, where laser spots are overlapping, due to the fact that the site hardened by the previous pulse is subjected to the effect of a peripheral part of the beam on irradiation by the following pulse. The power density in this part is gradually lowered towards the light spot rims. Here metal is heated less than in the central part of the beam. The resultant temperance of the surface previously hardened by laser radiation leads to lowering of the material microhardness. The microhardness distribution along the beam motion path on the surface of a steel sample is given in Fig. 1.8.11.

The hardened layer consists of the sites with the microhardness H_{100} = 1000 kg/mm^2 which are separated by bands with a lower microhardness.

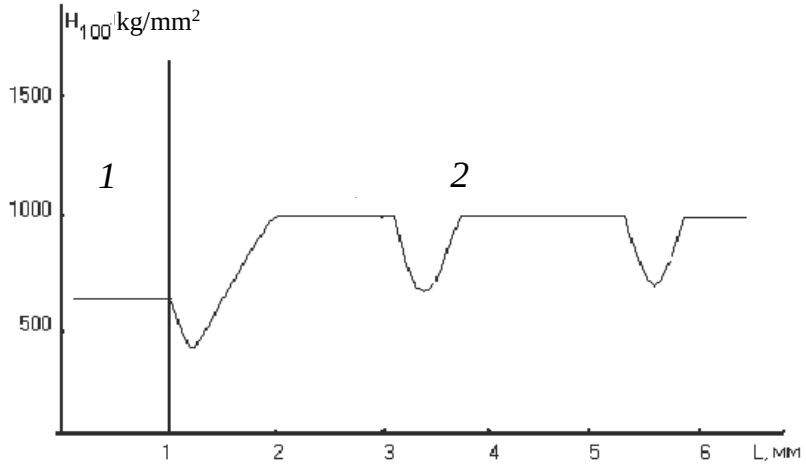


Fig. 1.8.11 – Microhardness distribution along the beam motion path for a steel sample: 1-zone with the initial structure, 2-hardened zone

The experimental results obtained in studies of steel thermostrengthening are listed in Tab. 1.8.2.

The result of laser quenching is the improved wear resistance of material surface that may be 3…4 times higher than with the ordinary thermal hardening.

Table 1.8.2 – Reflection factors for some samples at different wavelengths of radiation

Steel	Processing medium	Critical energy, J	Depth of hardened layer, mm	Microhardness, H100, kg/mm^2	
				Initial	After processing
U8	Air	26	0,135	650	1000
U10	Air	25	0,130	650	1000
ХВГ	Air	26	0,130	650	1000
XI2M	Air	24	0,110	590	1080
U8	Argon	31	0,120	650	1000
U10	Argon	30	0,115	650	1000
ХВГ	Argon	30	0,115	650	1000
XI2M	Argon	28	0,100	590	1080

1.8.3. Gas-laser cutting of metals

Much promise is offered by gas-laser cutting of metals that is based on the processes of heating, melting, evaporation; on chemical reactions of burning for removal of the melt from the cutting zone. By this cutting technique, a laser beam is directed to the processing zone together with a gas current to facilitate removal of the destruction products and initiating the chemical reaction at the point, where a material is subjected to the radiation effect. Usually this is realized with the use of oxygen, compressed air, inert or neutral gases. Depending on the properties of a material under processing and on the gas used, two mechanisms of laser cutting are distinguished: chemical and physical.

A chemical mechanism is mostly observed in the process of laser cutting of metals, forming flowable oxides, in a stream of oxygen and is characterized by a significant contribution of the combustion reaction energy to the total thermal balance. With a chemical mechanism of cutting two regimes are possible: (*i*) controlled cutting when heat from the combustion reaction only supplements the effect of laser radiation; (*ii*) uncontrolled gas-cutting when metal is burning due to heat of the combustion reaction over the whole stream diameter, whereas the energy of laser radiation only initiates this reaction. A physical mechanism consists in melting of a metal by laser radiation and removal of the melt from the cutting zone by means of a gas stream. Such a mechanism is usually observed during processing of metals having a low thermal effect of the combustion

reaction or forming refractory (high-melting) oxides and with the use of inert gases.

On gas-laser cutting of metals in a stream of oxygen the latter facilitates the formation of an oxide film on the metal surface that is responsible for its lower reflectivity.

The heat released as a result of exothermal reaction of combustion in oxygen is spent, together with laser radiation, on destruction of metal in the cutting zone. The stream carries away the destruction products and hence oxygen inflows directly to the burning front, prohibiting excessive heating of the workpiece material. In the case of inflammable materials a stream of gas predominantly clears the cutting zone and protects the surface of an optical system from ingress of the products ejected in the cutting zone. It should be noted that the reflection factor for metallic materials is rather high. Processing of metals in an oxidizing medium contributes to a drastic increase of the absorption factor.

Of great importance for gas-laser cutting are the focusing conditions of laser radiation; power of laser radiation, cutting rate, pressure of a heaved gas. For cutting of carbon steels, titanium alloys, and nonferrous metals the best results (greater cutting depth, smaller width) are achieved by focusing of laser radiation to spots with smaller diameters to increase the power density in the processing zone. To this end, we can use laser facilities with a minimal diameter of the laser beam outgoing from the resonator, operating in a single-mode regime, with short-focus lenses.

The main characteristics of gas-laser cutting are width and depth of the cut, quality of the formed rims, width of the heat-affected zone (HAZ). Besides, the process is influenced by the technological conditions.

In the majority of cases local heating on gas-laser cutting takes more than $10^{-8}c$ and thermal processes may be described using the classical theory of heat conduction. Depending on the parameters of a heat source (radiation power or energy; focusing spot diameter; duration of the effect), different models for heating are used. The cutting parameters and their interrelations may be estimated with the help of a simple model for the total thermal balance. Assuming that the whole energy of radiation is spent on melting and evaporation of material, we can write

$$AP = vbh\rho\left(c\Delta T + L_m + mL_{ev}\right),$$

where P – radiation power incident on the surface;

A – surface absorptivity;

v– motion rate of a laser beam;
b – cut width,
h – cutting depth,
ρ – material density,
c – heat capacity,
L_{ev} – latent melting heat,
L_{ev} – latent heat of evaporation,
m – fraction of evaporated material.

Based on this formula, in the approximation of the rapidly moving heat source ($vr/a >> 1$, where $d = 2r$ – spot focusing diameter, a – thermal diffusivity coefficient), one can easily calculate the function $h(P)$. In the assumption that the whole energy of laser radiation is spent to heating of the removed material up to the evaporation temperature and on transfer of the latent heat of evaporation, a maximal depth of the layer of evaporated material is determined as

$$h = \frac{2AP}{\pi r_f \rho v\left(cT_{ev} + L_{ev}\right)},$$

where the radius r_f is given by the concentration factor of the surface heat source. The linear dependence $h(P)$ is valid for small thicknesses of the material cut which are equal to several r_f. For high values of the material thickness we have $h \sim \sqrt{P}$.

A width of HAZ is found from a simple relation for heating temperature that is attributed to the action of a normal-bandpass source the intensity of which in the radial direction is distributed by the Gauss law and uniformly distributed over the plate thickness

$$T(y) = \frac{AP}{vhc\rho r_f \sqrt{\pi}} \exp\left(\frac{y^2}{r_f^2}\right).$$

It should be noted that reflection of radiation by the walls and a wave-guide character of beam propagation within the cut channel is of no significant importance for materials.

Based on theoretical and experimental studies, an optimum regime of cutting carbon steels with a great thickness is selected when the pressure of oxygen is

increased. Because of this, the cutting rate is growing. Considering that laser facilities with the increased power are more expensive, their usage is not always economically feasible for such steels. High-quality cuts of thick stainless steels necessitate the use of long-focus lenses ($F \sim 200 мм$). When using shorter focal-length lenses ($F \sim 100 мм$), one should realize cutting at lower powers of laser radiation ($P = 0{,}5 кВт$). The cut characteristics are greatly influenced by such parameters of the involved processes as power and power density of laser radiation. It is convenient, for estimation of the process efficiency and quality, to use the complex parameter representing a ratio of the radiation power and the cutting rate P/V_p or P/h (where h – metal cutting depth). If the heat conduction losses are considerably lower than the material heating and melting losses, we can write the following expression for the energy balance:

$$h \cdot V_p \cdot b\left(C \cdot \rho \cdot T_{пл} + L_{пл}\right) = \eta \cdot P,$$

where P – total power of laser radiation and of the exothermic oxidation reaction; $\eta = \eta_э \cdot \eta_m$ – process efficiency ($\eta_э$ – efficiency; η_m – thermal efficiency); $L_{пл}$ – specific energy of metal melting.

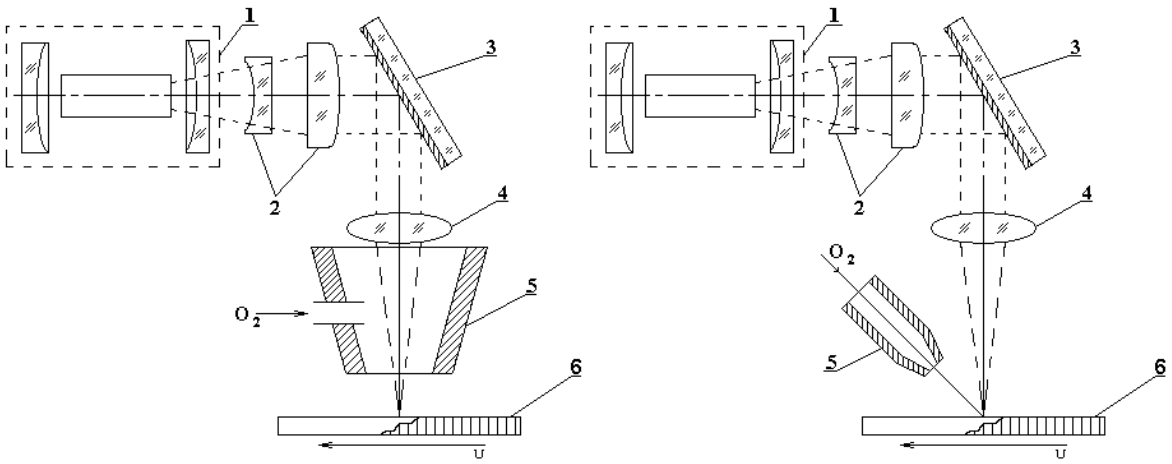
As for laser cutting a width of the cut equals the laser beam diameter, we can find that $V_p \sim P$ for $h = const$.

The dependence between the cutting rate and the plate width is less pronounced because the permissible rate may be limited by the quality of cutting. But in the general case we have $V_p \sim h^{-1}$.

The parameter P/h characterizing the energy expenditures per unit cut depth increases with the rate.

The power density E_f at the zone, where a material is affected by laser radiation, is also a complex quantity dependent not only on the power but also on the radiation focusing conditions, mode composition of the beam, its divergence, diameter of the beam at the output of a resonator.

Two schemes of gas feeding to the cut zone are used in the process of gas-laser cutting: coaxially with a laser beam through the nozzle of a gas-laser cutter (Fig. 1.8.12, a); from one side at an angle with respect to the optical axis through a special, capillary or injector (Fig. 1.8.12,b).



1 – laser, 2 – telescopic system, 3 – mirror, 4 – objective, 5 – nozzle, 6 – workpiece, 7 – jet
Fig. 1.8.12 – Schematic diagram for gas-laser cutting of materials

1.8.4. Laser hole drilling

Melting and evaporation of materials are the main processes involved in hole drilling with the use of lasers. A depth of the hole is growing due to evaporation, its diameter – due to melting of the walls and displacement of fluid by excessive vapor pressure. When laser radiation interacts with the material surface, some part of radiation is absorbed and some − scattered by the optical destruction products.

The absorption of radiation in the plume is influenced most significantly by the vapor phase of destruction products that represents the low-temperature weakly ionized plasma with the transparency determined by its temperature and concentration. As it has been demonstrated experimentally, radiation absorption in the plasma is most essential at the radiation flux densities $g \geq 10^{8} \div 10^{9}$ W/cm^3. This quantity sets a lower bound for the range of the working radiation-flux densities used in the process of laser hole drilling in materials.

A lower bound is determined by the flux density, for which the pit formation process begins at the time of pulse termination, i.e. It approximately corresponds to the beginning of material (metal) destruction $g = 10^{5} \div 10^{6}$w/cm^3.

The technological regimes of hole drilling by a laser beam is selected depending on the properties of a material under processing: absorption factor and reflectivity for the given wavelength which determine the process of energy absorption; heat conductivity and diffusivity which determine a heat flow in a material; density, specific heat, latent heat, and temperature of the phase transition which determine the process energy capacity for the material going to a new phase

state. One should take into account the effect of the energy and temporal characteristics of laser radiation. Many of industrial laser facilities have a single invariable pulse length. In this case the desired size of drilled holes is attained by selection of the adequate pulse energy. The energy dependence of the drilled-hole depth and diameter is the primary characteristic determining potentialities of laser processing. The energy is varied by the use of pulse lamps, light filters, by beam diaphragming. It should be noted that, when the energy is varied by pumping or beam diaphragming, diameters of the drilled holes are smaller than those obtained with the use of light filters.

A decrease of the diameter in the case of beam diaphragming is attributed to lower divergence and in the case of the energy variations − to pumping level of the active element, both the beam divergence and the radiation duration is lowered. Beam diaphragming at constant pumping of the active element results in the following feature: when the hole diameter is related to the diaphragm diameter, a depth of the hole very weakly depends on variations in the diaphragm over wide limits.

A hole depth is invariable on beam diaphragming due to the fact that a diaphragm has no effect on the flux density distribution over the beam cross-section, changing only the total divergence and the beam diameter. As the in-depth pit growth rate is determined by the flux power density, a depth of the hole should not be varied significantly in the case of beam diaphragming.

As demonstrated by experimental studies, hole sizes depend on the pulse length of laser radiation too. Variation of the pulse length from 0.25 to 0.85 ms at the constant pump energy results in the increased hole depth (by a factor of 1.5) with simultaneous decrease of the diameter by 30%. The dependence of the hole depth and diameter on the pulse length is associated with two factors: (1) the radiation intensity distribution is lowering from maximum (at the center of a beam) to the periphery; (2) lower shielding effect of the destruction products of a material with the lowered density of a luminous flux when the pulse length is increased. This is attributed to a greater fraction of the liquid phase remaining on the walls of a hole because it is liable to remain within the pit due to the lower vapor pressure and hence is not actually involved in the process of radiation shielding.

An important factor in the formation of molten puddle across the walls is the low-angle trailing edge of a radiation pulse. As demonstrated by the experiments with a cut-off of the trailing edge, in this case the surface microrelief of a hole becomes more smooth (practically without puddle of the molten metal), though

the hole depth varies insignificantly (15% when the 300-μs trailing edge is cut-off).

According to a phenomenological model, a rigorous analysis of the pit growth with wall melting necessitates solution of a hydrodynamic problem for motion of vapor and viscous fluid along the walls considering all the factors of their heating. Let us consider a simplified model: focusing of radiation, close to the front surface of a material, results in the formation of a pit with a nearly cylindrical profile (Fig. 1.8.13).

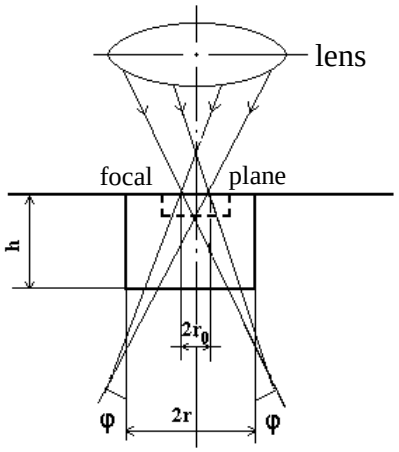


Fig. 1.8.13 Time variations in the pit depth *h* and radius *r*

Changes of the pit dimensions in time is so that in the first approximation the actual radius *r(t)* and depth *h(t)* are related with each other by the direction of a light cone for rim rays of the beam with the apex angle 2φ, and we have

$$r(t) = r\text{o} + h(t)\, tg\varphi$$

Deriving a phenomenological model, we neglect the temperature dependence of a complete heat of evaporation for a material and the surface shielding by the destruction products. According to a phenomenological theory, when laser radiation is focused at the material surface, by the time of the pulse action termination, the drilled hole has the depth

$$h = \sqrt[3]{\left(\frac{r_0}{tg\varphi}\right)^3 + \frac{3E}{\pi \cdot tg^2\varphi \cdot L_0}} - \frac{r_0}{tg\varphi}$$

and the diameter

$$d = 2\sqrt{r_0{}^3 + \frac{3E \cdot tg\varphi}{\pi \cdot L_0}} ,$$

where $E = P\tau$ – peak radiation energy ; P – pulse power; r_0 – initial radius of a pit that is equal to that of a light spot; φ – apex half-angle of a light spot; L_0 – latent specific heat of material evaporation.

In this way the decisive factor for the depth-to-diameter ratio of a hole is the quantity $tg\varphi$ determining the light cone angle after the focal plane of an optical system. The lower $tg\varphi$ the higher the depth-to-diameter ratio. To form a deep narrow hole, one should set the optimum condition $tg\varphi = 0$, and we have

$$h = \frac{E}{\pi \cdot r_0 L_0}; d = 2r_0; \frac{h}{d} = \frac{E}{2\pi \cdot r_0 L_0}$$

These relations describe the hole growth process within a cylindrical light tube into which the light cone is degenerated ($tg\varphi = 0$).

When $tg\varphi \neq 0$, defocusing of a light beam results in decrease of the energy density in its lower part.

As a consequence, there are limiting dimensions of a hole formed in a material under the effect of the unlimited number of light pulses with the specific energy E

$$d_{удел.} \approx 2\sqrt{\frac{E'}{\pi \cdot Q}} , \quad h_{удел} \approx \frac{\sqrt{\frac{E'}{\pi \cdot Q} - r_0}}{tg\varphi} ,$$

where Q – threshold energy density corresponding to initiation of the quasi-stationary evaporation process by the end of the pulse.

Of great importance for the hole profile are the focusing conditions of laser radiation: focal distance of a focusing system and shifting of the focal plane with respect to the surface of a workpiece. Fig. 1.8.14 shows profiles of the holes drilled in a material at the same energy of radiation but for different focal positions with respect to the surface of a workpiece.

As seen in Fig. 1.8.14, in a convergent beam the conical profile of a hole is characteristic; actually, its walls do not absorb the luminous flux propagating along these walls and a mechanism of destruction is material evaporation due to heat conduction. When irradiation is realized in a divergent laser beam with a sufficient energy density, intensive melting of the walls takes place and, though in this case the total amount of the removed material is somewhat greater due to

the liquid phase, a depth of the hole is decreased because of the beam defocusing. Let us consider the conditions for the formation of holes with a maximum depth by a single pulse at the specified energy and pulse length of laser radiation.

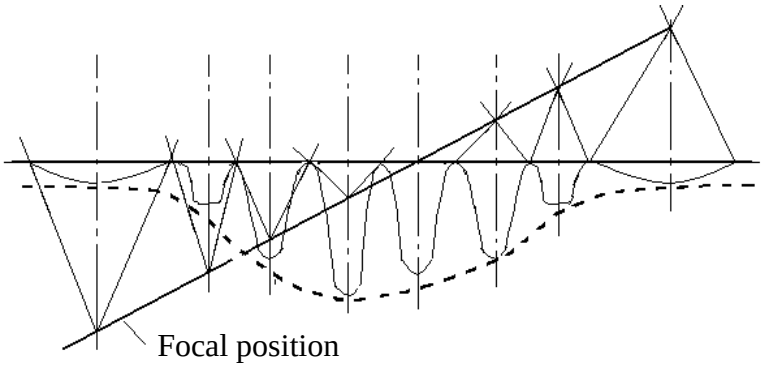


Fig. 1.8.14 Variations in the hole profile for different focal positions with respect to the material surface

As demonstrated by experiments, the formation of holes with maximum depths necessitates focusing of a light beam to the distance *h* from the surface of a material.

We can calculate l_0 in the assumption that, up to the focus of an optical system, the hole replicates the form and dimensions of a light cone, evaporation being its mechanism of formation:

$$l_0 = \frac{r_0}{\beta}\left[\sqrt[3]{1+\frac{3E\beta \cdot tg\varphi}{\pi \cdot L_0 \cdot r_0{}^3(\beta + tg\varphi)}} - 1\right],$$

$$h_{max} = l_0 \frac{\beta + tg\varphi}{tg\varphi},$$

where $\beta = \frac{D + 2\alpha \cdot l}{2f}$, D – light beam diameter at the output mirror of a laser; 2α - radiation divergence angle; l – distance form laser to the front focus of a lens with the focal distance *f*.

In practice, having the given *h* and *d* for a hole, it is essential to calculate the energy E, pulse length τ, spot size of focused laser radiation. It is important to take into consideration that, according to the equation for a light cone, $d = do + 2h\,tg\varphi$.

To have more accurate values of the diameter, we can use a simple relation between the volume of material stripped from a cylindrical hole and the radiation energy spent

$$\frac{\pi \cdot d^2}{4} h L_p = E \,,$$

where L_p – specific energy for destruction of a unit volume of material that is estimated experimentally.

This formula, together with a formula for the hole depth by the time of the pulse action termination, may be used to calculate the light beam parameters h and d when the optical and thermophysical characteristics of material are known

$$h = \sqrt[3]{\left(\frac{r_0}{tg\varphi}\right)^3 + \frac{3E}{\pi \cdot tg^2\varphi \cdot L_0}} - \frac{r_0}{tg\varphi}$$

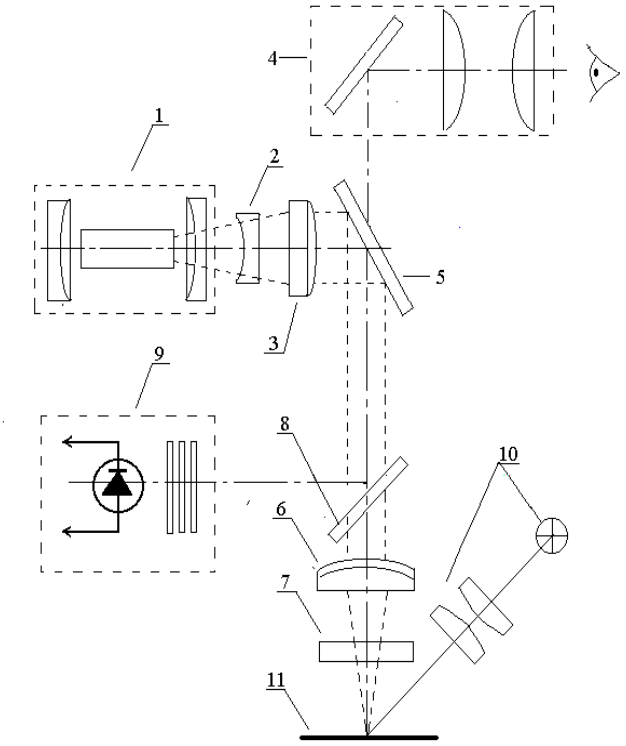
Using these formulae, we can find the energy and focal distance of a lens. The other parameters (pulse length τ, radiation divergence angle 2α, and distance between a laser and the front focus of a lens) may be selected in accordance with the laser type used. Because of this, the quantities τ, α, and l are considered known. Then the values of E and f are determined as follows:

$$E = \pi L \mathrm{p} r 2 h,$$

$$f = \left\{ \frac{3\left(r^2\varepsilon - \alpha \cdot h \cdot S\right) + \left[9\left(r^2\varepsilon - \alpha \cdot h \cdot S\right)^2 - 12\alpha^2 \cdot S^2 h^2\right]^{1/2}}{6\alpha^2} \right\}^{1/2},$$

where $r = \frac{d}{2}; S = \frac{D - 2\alpha \cdot l}{2}; \varepsilon = \frac{L_p}{L_0}$.

Experimental studies of laser hole drilling have been performed using the industrial laser facility shown in Fig. 1.8.15.



1 – laser source; 2,3 – telescopic system; 4 – visual channel; 5 – rotary mirror; 6 – optical system for focusing of laser radiation; 7 – protective plate; 8,9 – control unit for the laser beam energy; 10 – illumination; 11 – sample for hole drilling

Fig. 1.8.15 Optical scheme of an industrial laser facility for hole drilling in materials

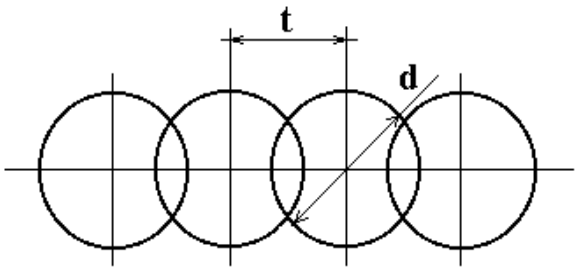
1.8.5. Laser scribing of dielectric materials

Similar to other types of laser processing, laser scribing of materials is based on thermal effect of radiation when heating takes place up to the temperature of evaporation. A pulse of the focused laser radiation forms a pit on the processed surface.

Due to relative displacement of radiation or of the plate, pits are partially overlapping to form a continuous scribe line on the plate surface. Lowering of the local strength with the formation of a scribe line is influenced by three factors: decreasing of the cross-sectional area of plates; formation of the stress concentrator; variations in strength of the material layer close to the scribe line due to the effect of laser radiation. The formation of a scribe line is schematically shown in Fig. 1.8.16, where d is the diameter of a single pit and *t* is the center distance of adjacent pits. The cut width is equal to the radiation spot diameter in the focal plane, whereas the rate of scribing is determined as

$$V = t \cdot f,$$

where f – pulse repetition rate.



d – diameter of a single pit; *t* – center distance for adjacent pits
Fig. 1.8.16 Schematic diagram for the formation of a scribe line

The overlap factor (characterizing overlapping of individual pits) is found as $K = \frac{d-t}{d}$, and the rate of scribing is derived as

$$V = f \cdot d\,(1 - K).$$

For $K > 0.5$, the points positioned at the center line of individual pit are under the effect of several pulses. The averaged number of pulses affecting a single point is given by

$$n = \frac{1}{1-K}.$$

In the case of single-mode generation the radiation intensity at the spot is distributed by the Gauss law. But, as the focal spot center is shifted with respect to the fixed point, every subsequent pulse makes a different contribution to the increased depth of marks. A maximal contribution is made by the pulse whose center of the spot is coincident with this point. Fig. 1.8.17 presents a characteristics graph for scribing depth as a function of pulse numbers and of the overlap factor: $h = F(n) = F(K)$. These experimental curves have been obtained at the average power of laser radiation in the continuous mode, i.e. at the pump power P = 2.6 kW. As seen in Fig. 1.8.18, for every value of the mark depth h the function $V = F(f)$ is at maximum. This points to the fact that for scribing to the specified depth one can always determine the frequency and the overlap factor

associated with a maximal rate of scribing.

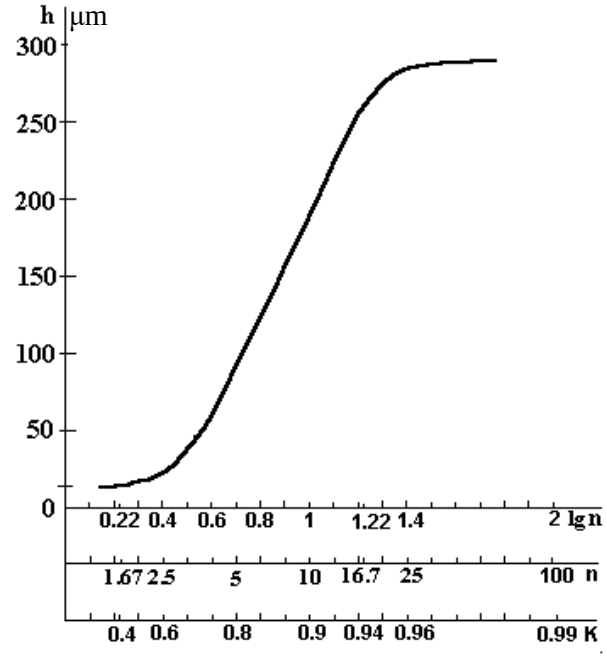


Fig. 1.8.17 – Depth of scribing as a function of the pulse number and of the overlap factor

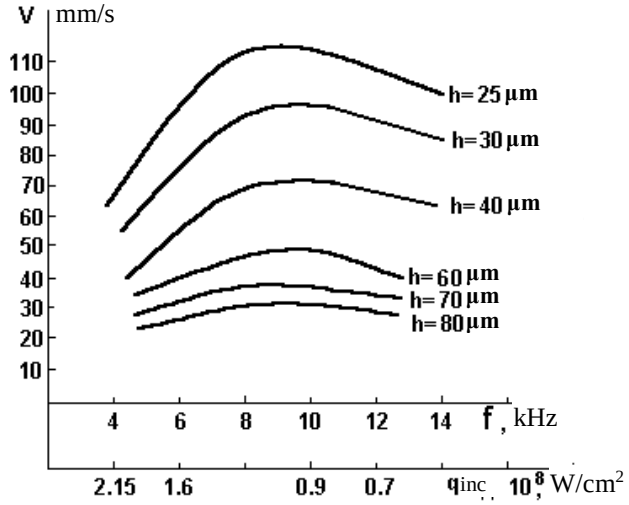


Fig.1.8.18 − Scribing rate as a function of the pulse repetition frequency and of the laser-radiation power density

The above-mentioned functions enable one to find the conditions and optimal mode for scribing. The mode is dictated by a choice of the plate displacement and of the pulse repetition rate. The pulse repetition rate and frequency may be found by the following procedures.

1. From the curve for the optimal power density given in Fig. 1.8.19 we can find a value of q_{inc} for scribing to the specified depth.

2. The pulse power is derived from the relation $P_{pulse} = q_{inc\ opt} \cdot \frac{\pi d^2}{4}\ W$.

3. From the curve in Fig. 1.8.20 one can obtain the pulse repetition frequency.

4. An optimum overlap factor is found from the curve shown in Fig. 1.8.21.

5. The scribing rate is given by the expression $V = f \cdot d\ (1 - K)$.

Scribing of silicon, gallium arsenide, and other materials for subsequent fragmentation into individual elements along the cut line leads to greater output of the accepted parts.

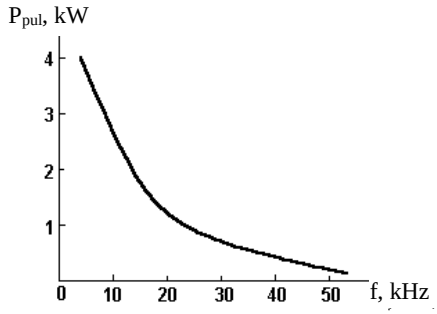


Fig. 1.8.19 − Optimal power density of laser radiation for scribing to the specifeid depth

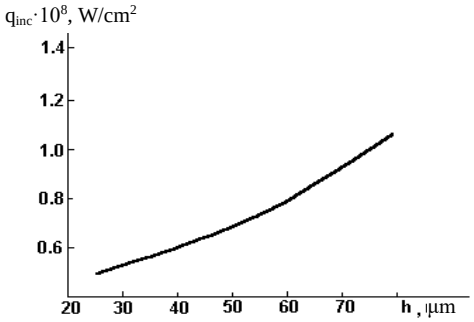


Fig.1.8.20 − Pulse power of laser radiation as a fucntion of the repetition frequency

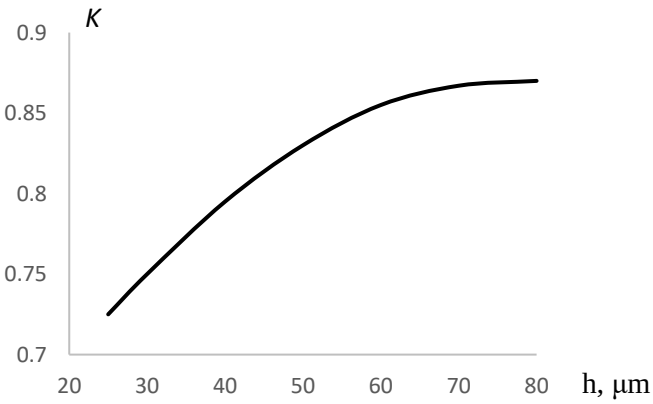


Fig.1.8.21 – Scribing depth as a fucntion of the overlap factor

As a rule, scribing of semiconductor plates is realized with the use of continuous-wave Q-switched YAG lasers (radiation wavelength λ = 1.06 μm, pulse length 0.15 – 0.3 μs, pulse repetition frequency 1-50 kHz, average power of pulse series 0.5-16 W at the peak power 1-20 kW). A depth of the cut made by a focused laser beam comes to 40 - 125 μm, the width is 20-40 μm at the plate thickness 150-300 μm. Ordinary, the scribing rate is 100-150 mm/s. Full detachment of the plate is also possible. The heat-affected zone in the semiconductor layer adjacent to the cut is below 50μm.

Due to high power density ($10^8 \div 10^9$ W/cm^2), silicon and gallium arsenide are melting, in the process of scribing a temperature at the point affected by a focused laser beam comes to 2000 ^{0}C. Owing to very small lengths of laser pulses and to motion of the cut plate with respect to the focused laser beam, heat released into the region adjacent to the notch on a semiconductor material has no significant effect on this material.

In the process of scribing some fractions of the removed semiconductor may be settled near the cut at the plate surface. This contamination is removed, for example, in an ultrasonic bath. Besides, special coatings are available, which are applied to the plate surface before scribing and are removed after the procedure.

Scribing of ceramic plates, with the formed structures or without them, are divided into modules along the cut line. Scribing of electrocorundum and berrylium ceramic plates is implemented with the use of CO_2 or YAG lasers operating in the pulsed mode with a high pulse repetition frequency. The average power of a radiation beam in the case of CO_2 laser is 20-50 W, peak power comes to 40-150 W at the pulse length 0.1 - 5 ms and pulse repetition rate from 100 Hz to 1 kHz. Radiation with the wavelength λ = 10.6 μm (CO_2 laser) is better absorbed by ceramics than with the wavelength λ = 1.06 μm (YAG lasers) making CO_2 lasers more advantageous in this case.

Scribing with the use of a CO_2 laser consists in forming a series of blind holes 75-200 μm in diameter, 100-200 μm in depth at the distance 75-200 μm from each other.

With the use of YAG lasers for scribing of ceramic plates, the cut is uninterrupted; even full detachment of the plate parts is possible due to high frequency and power of laser pulses.

Also, laser scribing is used for plates of sapphire, glass and other dielectric materials.

1.8.5. Laser marking and engraving

The technology of laser marking and engraving is associated with the formation of textual and graphical images on the surface of the processed workpiece under the effect of high-intensity laser radiation. In the case of marking a thickness of the stripped material is up to 100 μm, in the case of engraving it is about 0.5 mm , and in the case of deep engraving – 3.5 mm. Modern laser technologies make it possible to mark various materials: metal, wood, organic glass, plastic, acryl, rubber, leather, etc.

Generally, marking is realized with the use of continuous-wave CO_2 lasers, YAG:Nd lasers operating in the Q-switching mode with a high pulse repetition rate (0.5…100 kHz), which generate single-mode (TEM_{00}) radiation with small divergence (4…15 mrad) and with the output power ranging 10…16 W, and fiber lasers.

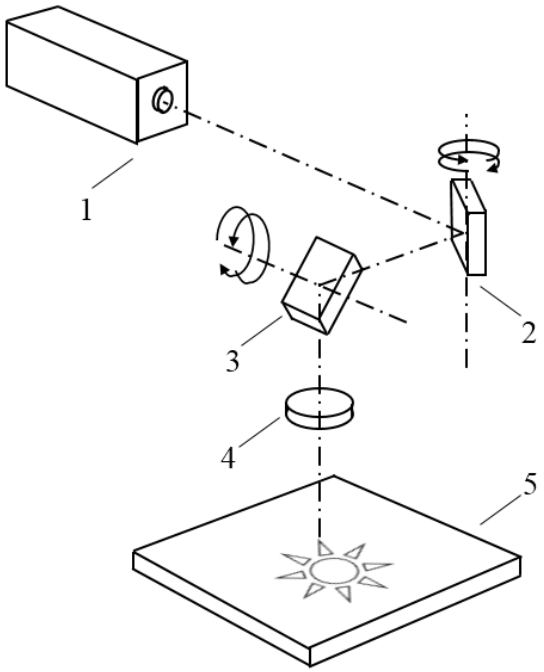
Engraving is realized with the use of continuous-wave CO_2 lasers at the power from 10 to 50 W and pulsed YAG:Nd lasers operating in the free running mode (peak energy up to1 J, pulse repetition rate up to 100 Hz).

Marking and engraving is implemented by scanning of the focused laser beam over the processed surface. Motion of a laser beam is provided with the help of fast double-axis galvanoscanners or the projection method is used.

Fig. 1.8.22 shows a laser setup for marking and engraving with the use of a galvanoscanner. The type of a laser emitter is determined by the material under processing and process parameters. Scanning of processed surface 5 is effected with the use of computer-controlled mirrors 2,3 of a galvanoscanner. The aperture of long-distance objective 4 is dependent on the size of a processing field.

When the projection method is used, a laser beam, of the form defined by the special mask, is projected on the processed surface in the required scale to form an image reproducing the mask form. This approach is common in the production of microelectronic components and devices.

One of the kinds of laser marking is laser processing of surfaces for subsequent formation of thin color films. Such films are used as inhibiting coatings of the end products, which protect metal surface from corrosion and scratches. Besides, laser processing is used for the formation of color images on metal surfaces (specialty goods).



1 – laser emitter; 2, 3 – galvanoscanner mirror; 4 – focusing objective, 5 – workpiece
Fig. 1.8.22 – Laser setup for marking and engraving of workpieces

Color marking of metals is based on the process of forming oxide or nitride films on the surface of processed material under the effect of laser radiation. Oxide films are formed during processing in the air. Nitride films are formed on metals in a special chamber with heaving of nitrogen. In both cases the film color is determined by its chemical composition and thickness. In practice the formation of oxide films is more common as this requires no special technological equipment for nitriding. Oxide films are formed due to homogeneous heating of the metal surface that stimulates the process of its oxidation.

Thus produced oxide films may be of different color shades. The formation and variation of color is caused by light interference due to summation of the waves reflected from the surface layer of an oxide film and from the surface of the metal itself. As a thickness of an oxide is growing, the conditions are formed for quenching of beams with some or other wavelengths − that offers the possibility to obtain films with variations of color from violet to red. A degree of the surface roughness is also important for the film color.

Color films are usually formed with the use of excimer lasers operating in the UV spectral region or fiber lasers with the wavelength 1.06…1.07 μm. This is associated with the fact that the majority of metals have high reflection factors in the IR region

References

1. A.Yariv. Quantum Electronics, John Wiley and Sons, Inc., 1975.
2. O. Svelto. Principles of Lasers. Prenum, Press-New York and London, 1989.
3. W.T. Silfvast. Lasers Fundamentals, Cambridge University Press, 2003.
4. Edited By C.E. Webb, J.D.C. Jones. Handbook of Laser Technology and Applications (Three- Volume Set). CRC Press, 2003.
5. S. Hooker. Laser Science and Quantum Information Processing, Laser Physics, 2006.
6. S. Hooker. Radiation and Matter. – Basic Laser Physics, 2006.
7. A.L.Tolstik, I.N.Agishev, E.A.Melnikova. Laser Physics. Practical Laboratory Works. Minsk: Belarusian State University, 2006.
8. P.W.Milonni, J.H.Eberly. Laser Physics. Wiley, 2010.
9. S. Hooker, C. Webb. Laser Physics. University of Oxford, 2016.
10. Parfenov V. A. Laser micro-processing of materials: Tutorial / V.A. Parfenov. - Saint-Petersburg Electrotechnical University "LETI", 2011. – 59 p.

Chapter 2. Nonlinear optics

2.1. Nonlinear medium and nonlinearity mechanisms

Introduction

Owing to the advent of the high-power sources of coherent optical radiation (lasers), we have the possibility to observe and use the nonlinear optical effects, a character of which depends on the radiation intensity. In nonlinear optics, the medium polarization is nonlinearly dependent on the light-wave field strength, the superposition principle is violated, and light beams can affect each other due to variations in the medium characteristics (refractive index and/or absorption factor).

The first nonlinear effect discovered in 1923 by S.I. Vavilov is a change in the absorption factor (medium bleaching) under the effect of high-power radiation. With the use of a spark, it has been found that the absorption factor of glass activated by uranium salts is lowered by 1.5% (at the measurement accuracy $\pm 0.3\%$). This effect is quite understandable when using the concept of the molecular transition to the excited states, for which the absorption cross-section is differing from the initial one. The decisive factor is a rather long lifetime of the molecules in the excited state ($\tau \approx 03$ms).

No other nonlinear-optical phenomena have been detected until the use of lasers. With the advent of lasers (1960), numerous works in nonlinear optics were underway. Several nonlinear effects have been discovered during the period of five years: second-harmonic generation, sum- and difference-frequency generation, self-focusing, autocollimation and defocusing of light beams, stimulated Raman scattering of light, stimulated Mandelshtam-Brillouin scattering.

The unique phenomenon of the wavefront conjugation was discovered in 1970. This phenomenon permits returning in time to the light beam as it was before passing through a scattering medium. The wavefront conjugation phenomenon makes it possible to compensate for distortions of light propagating in phase-inhomogeneous media (e.g., in high-power laser amplifiers) and to realize self-guidance of laser radiation to the target. The phenomenon of optical bistability discovered in 1974 enables one to have two different states of an optical system for the same input intensity. At the same time, an optical analog of the transistor capable to amplify optical signals without the use of amplifying media was demonstrated; optical hysteresis was realized. Discovery of the squeezed states of light fields, offering the possibility to lower noise in the process of photorecording below the shot noise level, was an outstanding achievement.

The performed experiments necessitated the development of new methods for theoretical description of the studies. The earlier calculation methods for the optical phenomena have been based on separate solutions of the two problems: finding of the material parameters (e.g., spatial distribution of the absorption factor or refractive index) and calculations of the radiation propagation within the material with the specified properties. Such approach is applicable only in the case when the radiation effect on the material is insignificant. In the general case it is required to take into account the effect exerted by both the medium on radiation and by radiation on the medium. This is offered by the calculation methods which are based on the probabilistic balance equations for populated levels and on the mathematical apparatus of the quantum-mechanical density matrix. In so doing the commonly used assumption that photons are not interacting with each other is violated.

The medium polarization vector P related to the electric field strength of the light wave describes the interaction of a light wave with some material as follows:

$$P = \chi E, \qquad (2.1.1)$$

where χ – optical susceptibility of the medium. As high-intensity laser radiation is transmitted through a medium, the susceptibility χ is not assumed to be a constant parameter. It is a function of the light-wave field strength. This function may be series expanded in terms of the electric field strength in the assumption of infinitesimal field strength of the light wave E as compared to the intraatomic field E_{atom}:

$$\chi(E) = \chi_1 + \chi_2 E/E_{atom} + \chi_3(E/E_{atom})^2 + \chi_4(E/E_{atom})^3 + \chi_5(E/E_{atom})^4 + \ldots . \qquad (2.1.2)$$

For the electron on the outer atomic shell, the intraatomic field is calculated based on the relations $E_{atom} = e/4\pi\varepsilon_0 r^2$ (SI) and $E_{atom} = e/r^2$ (CGSE system). When a radius of the atomic electronic shell is 0.1 nm, the electric field strength comes to $\sim 10^{11}$ V/m ($\sim 5*10^6$ CGSE electrostatic units), that is associated with the intensity $\sim 10^{16}$ W/cm^2. Such an intensity is higher than that of the classical light sources by many orders of magnitude and may be attained only by focusing of high-power laser pulses.

Substituting the variables, we can introduce the nonlinear optical susceptibilities of different orders: $\chi^{(1)} = \chi_1$, $\chi^{(2)} = \chi_2/E_{atom}$, $\chi^{(3)} = \chi_3/E_{atom}^2$, $\chi^{(4)} = \chi_4/E_{atom}^3$, $\chi^{(5)} = \chi_5/E_{atom}^4$, etc.

In this case the optical susceptibility of a medium may be given as

$$\chi(E) = \chi^{(1)} + \chi^{(2)}E + \chi^{(3)}E^2 + \chi^{(4)}E^3 + \chi^{(5)}E^4 + \ldots . \qquad (2.1.3)$$

Then the polarization P (2.1.1) is represented in the following form:

$$P = \chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \chi^{(4)}E^4 + \chi^{(5)}E^5 + \ldots , \quad (2.1.4)$$

where $\chi^{(1)}$, $\chi^{(2)}$, $\chi^{(3)}$, $\chi^{(4)}$. $\chi^{(5)}$ – medium parameters characterizing its polarizability. The greatest nonlinear term is polarization quadratic in the field $P^{(2)} = \chi^{(2)}E^2$. However, note that, for isotropic media and centrosymmetric crystals, expression (2.1.4) has no even terms, and the greatest nonlinearity is cubic, $P^{(3)} = \chi^{(3)}E^3$.

In the case of anisotropic media the components of the optical susceptibility $\chi^{(N)}$ are different-order tensors. For such media the nonlinear polarization of a medium is given by the expression

$$P_i = \sum_k \chi_{ik}^{(1)} E_k + \sum_k \sum_l \chi_{ikl}^{(2)} E_k E_l + \sum_k \sum_l \sum_m \chi_{iklm}^{(3)} E_k E_l E_m + \ldots \quad (2.1.5)$$

To find the nonlinear optical susceptibility, we use the well-known relations connecting the medium dielectric constant ε to the optical susceptibility χ

$$\varepsilon = 1 + 4\pi\chi . \quad (2.1.6)$$

Representing the optical susceptibility as a sum of the linear and nonlinear parts

$$\chi = \chi_l + \chi_{nl} , \quad (2.1.7)$$

the dielectric constant of a medium may be derived as

$$\varepsilon = 1 + 4\pi\chi_l + 4\pi\chi_{nl} = \varepsilon_l + 4\pi\chi_{nl} . \quad (2.1.8)$$

On the other hand, we have

$$\varepsilon = n^2 = (n_0 + \Delta n)^2 = n_0^2 + 2\Delta n \cdot n_0 + (\Delta n)^2 . \quad (2.1.9)$$

Considering that $\Delta n \ll n_0$ (as a rule, $\Delta n \sim 10^{-3\text{–}5}$), an expression for the nonlinear susceptibility may be written as follows:

$$\chi_{nl} = n_0 \Delta n / 2\pi . \quad (2.1.10)$$

In this way, finding nonlinear variations in the refractive index, one can find the nonlinear optical susceptibility. Let us consider the specific mechanisms associated with variations of the refractive index under the effect of optical radiation.

2.1.1. Thermal nonlinearity

Thermal nonlinearity is due to the most obvious nonlinearity mechanism – variation of the refractive index as a result of heating of the optical-radiation absorbing medium. To calculate thermal variations in the refractive index, we use

the notion of the thermo-optic coefficient $\partial n/\partial T$. Then variations in the refractive index may be of the form

$$\Delta n = \frac{\partial n}{\partial T}\Delta T\,. \tag{2.1.11}$$

Variations in the medium temperature ΔT under the effect of optical radiation are found from the formula

$$W_{abs} = C_\rho \Delta T\,, \tag{2.1.12}$$

where W_{abs} - energy absorbed in a unit volume, C_ρ - heat capacity of a unit volume.

When a light beam is propagating in the absorbing medium at the intensity I, the energy absorbed in a unit volume may be represented by

$$W_{abs} = kIt\,, \tag{2.1.13}$$

where k - absorption factor, t - duration of the effect.

Considering (2.1.11) – (2.1.13), thermal variations in the refractive index may be described as follows:

$$\Delta n = \frac{kIt}{C_\rho}\frac{\partial n}{\partial T}\,. \tag{2.1.14}$$

It should be noted that a value of the thermo-optic coefficient $\partial n/\partial T$ is dependent on the material used and on the pulse length. At short length of the pulse, when the medium expansion may be neglected, the thermo-optic coefficient is found for the constant volume $(\partial n/\partial T)_{V=const}$, its values hardly exceeding $\sim 10^{-5} K^{-1}$. An increase in temperature in the region of illumination results in the local growth of the pressure responsible for the subsequent medium expansion. A value of the thermo-optic coefficient is considerably greater than $(\partial n/\partial T)_{P=const} \sim -10^{-4}$. The characteristic expansion time is determined by the acoustic-wave propagation time $t_{acoustic} \sim L/v_{acoustic}$, where L - illumination region, $v_{acoustic}$ - velocity of the acoustic wave. At the initial stage the thermo-optic coefficient is positive (the temperature and pressure growth leads to the increased refractive index). Then the medium expansion results in lowering of the refractive index, the thermo-optic coefficient is negative. Note that a value of the thermo-optic coefficient is greatly dependent on the material used. To illustrate, for ethyl alcohol having a high thermal nonlinearity we have $(\partial n/\partial T)_{P=const} = -4\cdot 10^{-4}$, whereas for water – $(\partial n/\partial T)_{P=const} \sim -2\cdot 10^{-5}$.

Assuming that heating of a medium leads to the temperature growth by 0.1° (as for the alcoholic solution of a dye at the intensity $I \sim 1$ MW/cm^2, pulse length

$_t$~ 10 ns, absorption factor k = 20 cm^{-1}, specific heat c_T =2400 J/kg K, and density ρ = 800 kg/m^3), variation in the refractive index comes to $\Delta n = -4 \cdot 10^{-5}$ only.

Apart from nonlinear variations in the refractive index, of great importance for practical applications is fastness of the medium response that is determined by the time interval sufficient for relaxation of the medium. For media with thermal nonlinearity this time is associated with heat removal from the excitation region. The characteristic time of thermal relaxation is found from the heat diffusivity equation and is greatly dependent on the boundary conditions. We can use a simple expression

$$\tau = d^2 / a, \tag{2.1.15}$$

where a - heat diffusivity, d - size of the excitation region. For many media such as water, ethyl alcohol, glass, polymeric materials the heat diffusivity comes to ~ 10^{-3} cm^2/s. When the excitation region is changing from 1 μm to 1 mm, the time of thermal relaxation τ is varying over a wide range from 10 μs to 10 s.

Using the expressions for nonlinear susceptibility (2.1.10) and thermal variations of the refractive index (2.1.14), we can write an equation for the nonlinear optical susceptibility

$$\chi_{nl} = \frac{n_0 k I t}{2\pi C_\rho} \frac{\partial n}{\partial T}. \tag{2.1.16}$$

Considering that the medium nonlinear polarization is given by $P_{nl} = \chi_{nl} E$, and the light-field intensity – by $I = c n_0 |E|^2 / 4\pi$ (CGS electrostatic units), it is inferred that thermal nonlinearity is cubic, and we have

$$\chi^{(3)} = \frac{c n_0^{\,2} k t}{8\pi C_\rho} \frac{dn}{dT}. \tag{2.1.17}$$

Nonlinear properties are often described with the use of the nonlinearity parameter n_2 that is related to nonlinear variations in the refractive index

$$\Delta n = n_2 E^2, \tag{2.1.18}$$

from whence it follows that for thermal nonlinearity we can write

$$n_2 = \frac{c n_0}{4\pi} \frac{k t}{C_\rho} \frac{\partial n}{\partial T}. \tag{2.1.19}$$

For the above-mentioned parameters (pulse length $_t$~ 10 ns, absorption factor k = 20 cm^{-1}, specific heat c_T =2400 J/kg K, density ρ = 800 kg/m^3, refractive index n_0 = 1.36, thermo-optic coefficient $\partial n / \partial T = -4 \cdot 10^{-4}$) the

nonlinearity parameter is $n_2 \sim$ - 10^{-8} CGS electrostatic units and the cubic susceptibility is $\chi^{(3)} = n_0 n_2 / 2\pi \sim$ - 2 10^{-9} CGS electrostatic units.

2.1.2. Resonance nonlinearity

Resonance absorption of light by atomic and molecular media is determined by the structure of energy levels. It is the case when the radiation frequency is coincident with the energy gap between the levels. When particles are going to the excited levels, the absorption factor changes because, as a rule, atoms and molecules have other probability of absorption than from the ground level. S.I. Vavilov was the first to observe variations in the absorption factor at high intensities in 1923. With the advent of lasers, the optical bleaching effect has been observed at high intensities for the majority of absorbing media. Besides, it has been noted that the absorption factor is changing together with the refractive index of a resonant medium.

To describe resonant media, we use a two-level model including the two singlet levels S_0 and S_1 (Fig. 2.1.1). Transitions between the singlet states S_0 - S_1 are associated with absorption in the visible and near ultraviolet spectral regions. After transition of atoms (molecules) to the excited state, we can observe spontaneous and stimulated radiative transitions S_1 - S_0 as well as radiationless relaxation of the state S_1 by means of the transition to the state S_0.

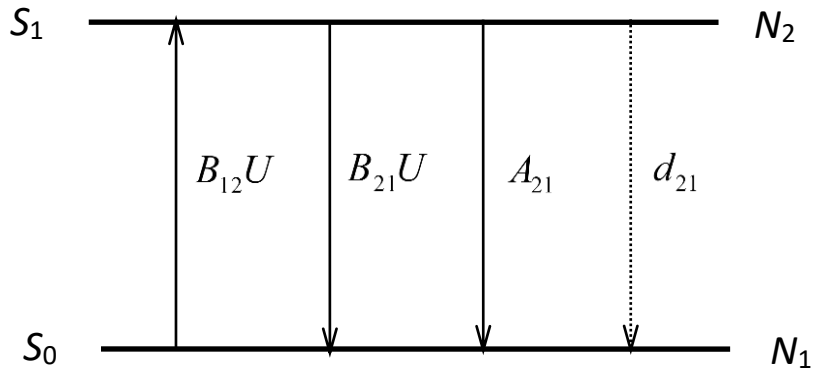


Fig. 2.1.1. Two-level scheme for the energy levels of a resonant medium with regard to the stimulated $B_{12}U$, $B_{21}U$, spontaneous radiative A_{21}, and radiationless d_{21} transitions between the singlet states S_0, S_1

To describe the light-induced processes in resonant media, we use kinetic equations for the populations of the energy levels N_1 and N_2

$$\frac{\partial N_2}{\partial t} = N_1 B_{12} U - N_2 (B_{21} U + A_{21} + d_{21}), \qquad (2.1.20)$$

where $B_{12}U$, $B_{21}U$ - probability of stimulated transitions under the effect of laser radiation, A_{21} - probability of spontaneous radiative transitions, and d_{21} - probability of radiationless transitions (the probability of a transition is understood as a number of transitions in a unit time), $U = I/v$ - volume energy density in a medium, I - radiation intensity, v - speed of light in a medium. Here the factors A_{21}, B_{12}, B_{21} are the Einstein coefficients for the spontaneous A_{21} and stimulated transitions with absorption B_{12} and with emission B_{21}.

On the other hand, the total population on transitions between the levels is retained

$$N_1 + N_2 = N, \qquad (2.1.21)$$

where N – number of the absorbing particles in a unit volume.

Using expressions (2.1.20) and (2.1.21), we get

$$\frac{\partial N_2}{\partial t} + N_2((B_{12} + B_{21})U + A_{21} + d_{21}) = NB_{12}U . \qquad (2.1.22)$$

A solution of the differential equation (2.1.22) may be obtained in the analytical form in the assumption of the rectangular-form pulse of laser radiation

$$N_2 = N \frac{B_{12}}{B_{12} + B_{21}} \frac{\alpha I}{1 + \alpha I} (1 - \exp(-(1 + \alpha I) t_{pulse} / \tau), \qquad (2.1.23)$$

where $\alpha = (B_{12} + B_{21})/v(A_{21} + d_{21})$ - nonlinearity factor, $\tau = 1/(A_{21} + d_{21})$ - intrinsic lifetime of the particles in the excited state (relaxation time of a medium in the absence of stimulated transitions), t_{pulse} - laser pulse length.

As follows from (2.1.23), population of the excited level is growing with the pulse length. For small $t_{pulse} << \tau/(1 + \alpha I)$, we have

$$N_2 = N \frac{B_{12}}{B_{12} + B_{21}} \alpha I \, t_{pulse} / \tau . \qquad (2.1.24)$$

For $t_{pulse} > \tau/(1 + \alpha I)$, the population N_2 approaches the stationary value

$$N_2 = N \frac{B_{12}}{B_{12} + B_{21}} \frac{\alpha I}{1 + \alpha I}, \qquad (2.1.25)$$

From (2.1.25) it follows that a population of the excited level is a saturable function of intensity, the population at low intensities being proportional to the intensity.

After finding the excited level population, we can find the absorption factor. To this end, we use the known formula for the absorption factor

$$k(\omega) = \frac{\hbar\omega}{v}(N_1 B_{12}(\omega) - N_2 B_{12}(\omega)). \quad (2.1.26)$$

Then, considering equation (2.1.23), we get

$$k(\omega) = N\frac{\hbar\omega}{v}B_{12}\left(1 - \frac{\alpha I}{1+\alpha I}(1-\exp(-(1+\alpha I)t_{pulse}/\tau)\right). \quad (2.1.27)$$

For the intensity $I = 0$, the initial absorption factor is found as

$$k(\omega) = N\frac{\hbar\omega}{v}B_{12} = k_0(\omega). \quad (2.1.28)$$

For a short pulse length when $t_{pulse} << \tau/(1+\alpha I)$, from (2.1.27) we can derive

$$k(\omega) = k_0 \alpha I t_{pulse}/\tau, \quad (2.1.29)$$

and in the stationary state we have

$$k(\omega) = k_0/(1+\alpha I). \quad (2.1.30)$$

As follows from equation (2.1.30), the absorption factor decreases monotonically as the radiation intensity is growing. At $I = \alpha^{-1}$ the absorption factor is halved (Fig. 2.1.2). This intensity is referred to as the saturation intensity

$$I_{sat} = v(A_{21} + d_{21})/(B_{12} + B_{21}). \quad (2.1.31)$$

For the typical parameters of molecular resonant media, the intrinsic lifetime is $\tau = 1/(A_{21} + d_{21}) = 10^{-9}$s, Einstein coefficients for the stimulated transitions $B_{12} \approx B_{21} \sim 10^{13}$cm^3/J c, and the saturation intensity is on the order of 1 MW/cm^2. For media with long-lived levels, the saturation intensity is lowered considerably. For example, at the lifetime $\tau \sim 1$ ms that is possible for population of a triplet level the saturation intensity is $I_{sat} \sim 1$ W/cm^2.

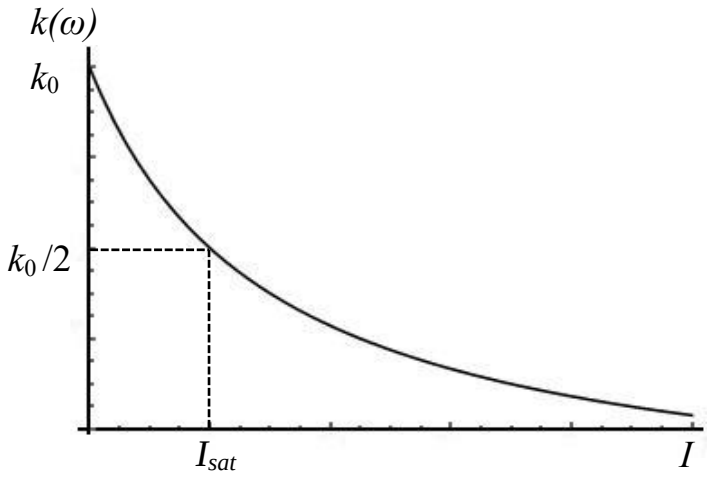


Fig. 2.1.2. Absorption factor versus intensity

On atomic or molecular transitions to the exited states, concurrent with variation of the absorption factor, the refractive index is also changed. This may be attributed to the fact that, when the electron goes to a higher stationary orbit, the molecular electronic shell is changed together with the intermolecular coupling forces.

The refractive index $n(\omega)$ is correlated to the extinction coefficient $\kappa(\omega)$, that is proportional to the absorption factor.

$$n(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\kappa(\omega')}{\omega' - \omega} \partial\omega' \qquad (2.1.32)$$

This relation is known as the dispersion relation that is a particular case of the Kramers-Kronig relations connecting the real and imaginary parts of any complex function, analytical in the upper half-plane. In the case under study a complex function is represented by the complex refractive index

$$\hat{n}(\omega) = n(\omega) + i\kappa(\omega). \qquad (2.1.33)$$

The physical meaning of such a complex function is as follows: if, in analogy, we introduce the complex wave too

$$\hat{k} = \omega\hat{n}/c, \qquad (2.1.34)$$

then, taking the electric-field strength in the form $E = A\exp(i(\hat{k}z - \omega t))$, we automatically include the medium absorption

$$E = A\exp(i(\hat{k}z - \omega t)) = A\exp(i(kz - \omega t)) \cdot \exp(-\omega\kappa z/c), \qquad (2.1.35)$$

where $k = \omega n/c$ - wave vector, and the extinction coefficient is related to the absorption factor by $\kappa(\omega) = ck(\omega)/2\omega$.

Using the Kramers-Kronig relation in the expression for the absorption factor (2.1.26), we get

$$n(\omega)=\frac{\hbar c}{2v}\left(N_1\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{B_{12}(\omega')}{\omega'-\omega}\partial\omega'-N_2\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{B_{21}(\omega')}{\omega'-\omega}\partial\omega'\right). \qquad (2.1.36)$$

The integral value is dependent on spectral form of the profile and may be found analytically for the Lorentzian form of absorption and emission profiles as follows:

$$B_{ij}=B_{ij}^{\max}/(1+\eta^2), \qquad (2.1.37)$$

where $B_{ij}^{\max}$ - Einstein coefficients at the profile maxima, $\eta=(\omega-\omega_0)/\Delta$ - spectral radiation-frequency detuning ω from the profile maximum ω_0, Δ - profile half-width.

For convenience, we introduce the function

$$\vartheta_{ij}(\omega)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{B_{ij}(\omega')}{\omega'-\omega}\partial\omega'=-\eta\frac{B_{ij}^{\max}}{1+\eta^2} \qquad (2.1.38)$$

It is easily seen that the function $B_{ij}(\omega)$ approaches a maximum at $\eta=-1$ and a minimum – at $\eta=+1$ (Fig.2.1.3).

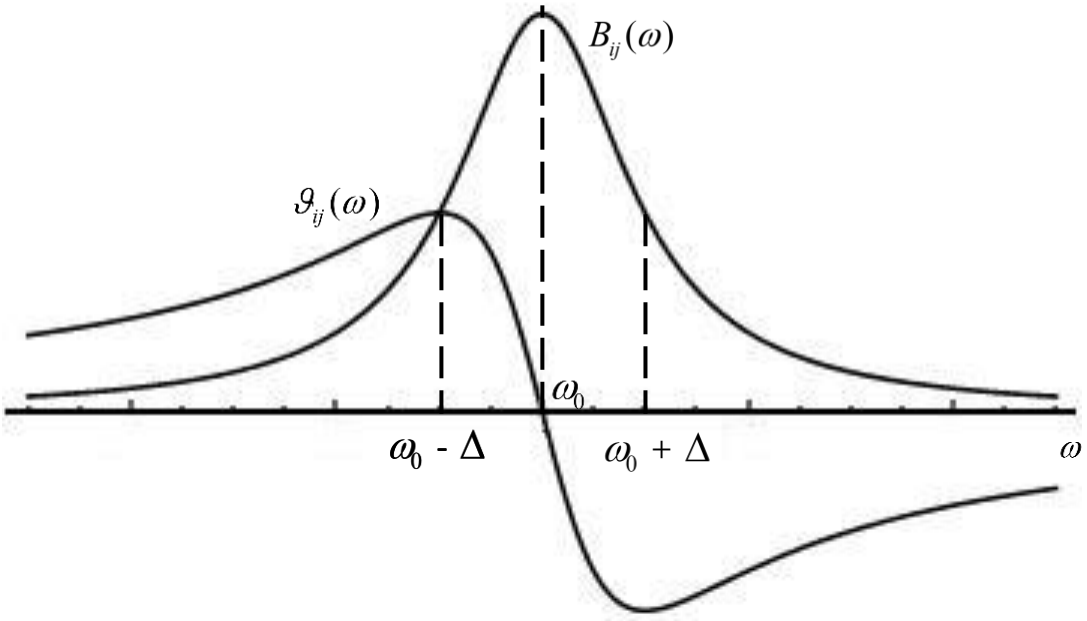


Fig. 2.1.3. Spectral dependences of the Einstein coefficients $B_{ij}(\omega)$ and of the function $B_{ij}(\omega)$ for the Lorentzian profile

Then the refractive index (2.1.36) is given by

$$n(\omega)=\frac{\hbar c}{2v}(N_1\vartheta_{12}(\omega)-N_2\vartheta_{21}(\omega)). \qquad (2.1.39)$$

Using relations (2.1.21) and (2.1.23) for the level populations we have

$$n(\omega) = N\frac{\hbar c}{2v}\vartheta_{12}(\omega) - N\frac{\hbar c}{2v}B_{12}(\omega)\frac{\vartheta_{12}(\omega)+\vartheta_{21}(\omega)}{B_{12}(\omega)+B_{21}(\omega)}\frac{\alpha I}{1+\alpha I}(1-\exp(-(1+\alpha I)t_{pulse}/\tau)) \tag{2.1.40}$$

For simplicity, we introduce the initial values of the refractive index and of the extinction coefficient in a resonant medium

$$n_0(\omega) = N\frac{\hbar c}{2v}\vartheta_{12}(\omega) \tag{2.1.41}$$

and

$$\kappa_0(\omega) = N\frac{\hbar c}{2v}B_{12}(\omega), \tag{2.1.42}$$

and involve the parameter

$$a = (\vartheta_{12} + \vartheta_{21})/v(A_{21} + d_{21}) \tag{2.1.43}$$

connected by the dispersion relation to the nonlinearity factor $\alpha = (B_{12} + B_{21})/v(A_{21} + d_{21})$.

In this case an expression for the refractive index (2.1.40) may be written in the following form:

$$n(\omega) = n_0 - \kappa_0\frac{aI}{1+\alpha I}(1-\exp(-(1+\alpha I)t_{pulse}/\tau)). \tag{2.1.44}$$

For a short pulse length, when $t_{pulse} << \tau/(1+\alpha I)$, from (2.1.44) we get

$$n(\omega) = n_0 - \kappa_0 aIt_{pulse}/\tau, \tag{2.1.45}$$

and in the stationary state

$$n(\omega) = n_0 - \kappa_0\frac{aI}{1+\alpha I}. \tag{2.1.46}$$

Note that for a short pulse length the effect is dependent on the laser pulse energy rather than on the intensity. Because of this, for observation of the nonlinear effect in these conditions, of importance is the energy of a pulse but not its intensity. In the stationary state the situation is radically changed – the nonlinear effect is determined by the intensity.

Based on equation (2.1.46), the light-induced variations in the refractive index are given by the expression

$$\Delta n = -\kappa_0\frac{aI}{1+\alpha I}. \tag{2.1.47}$$

As follows from (2.1.47), the light-induced variation in the refractive index is monotonically growing with radiation intensity. At the saturation intensity $I_{sat} = \alpha^{-1}$ this variation approaches a half of its maximal value.

Considering the form of the parameter a (2.1.43) and the form of the function $\vartheta_{ij}(\omega)$ (2.1.38), it is inferred that, for the coincident absorption and emission profiles, the light-induced variation of the refractive index is $\Delta n > 0$ in the short-wavelength spectral region ($\omega > \omega_0$) and $\Delta n < 0$ in the long-wavelength region ($\omega < \omega_0$). At a maximum of the absorption (emission) profile we have $\Delta n = 0$.

When the emission profile has a Stokes shift, that is characteristic for complex organic compounds (dyes), all the foregoing expressions are valid, excluding only the changing spectral dependence of the light-induced variations in the refractive index: $\Delta n > 0$ in the region of the absorption band and $\Delta n < 0$ in the region of the luminescence band.

To find a quantitative estimate of the light-induced variations in the refractive index, we select the medium parameters which are close to those used for thermal nonlinearity: intensity $I = 1$ MW/cm^2, absorption factor $k = 20$ cm^{-1}, saturation intensity $I_{sat} = 1$ MW/cm^2, frequency detuning $\eta = 1$, radiation wavelength 0.5 μm. In this case the extinction coefficient is $\kappa = \lambda k / 4\pi = 8*10^{-5}$; nonlinear variation in the refractive index comes to $\Delta n = 4 \cdot 10^{-5}$ and is coincident with thermal variations of the refractive index.

2.1.3. Electrooptic nonlinearity

The electrooptic effect is associated with variations in the optical properties of materials due to the action of an electric field. We distinguish the linear electrooptic effect and the quadratic electrooptic effect (Kerr effect). The linear electrooptic effect was first studied by Pockels in 1893 and was called the Pockels effect. Origination of the Pockels effect is caused by redistribution of bound electric charges and by deformation of the ionic lattice in crystals due to the action of the applied electric field. The quadratic electrooptic effect was detected by Kerr in 1875 during studies of isotropic media (liquids, glass) and was named after him. Initially consideration had been given to variations in the refractive index of different media in a direct-current field but similar results have been obtained subsequently under the action of the electromagnetic field of a light wave.

In the case of the linear electrooptic effect a change in the refractive index is proportional to the electric field strength ($\Delta n \square E$). Consequently, the nonlinear polarization of a medium associated with the Pockels effect is quadratic

$$P_{nl} = \chi^{(2)} E^2 . \qquad (2.1.48)$$

Such nonlinearity may be revealed only in anisotropic media (electrooptic crystals) having no center of inversion. Among these media, we can name crystals of lithium niobate ($LiNbO_3$), potassium dihydrophosphate (KH_2PO_4), barium titanate ($BaTiO_3$), bismuth silicate ($Bi_{12}SiO_{20}$), cadmium telluride (CdTe), gallium arsenide (GaAs), etc. The typical values of quadratic susceptibility come to $\chi^{(2)}$ ~ 10^{-8} CGS electrostatic units.

As distinct from the linear electrooptic effect, the quadratic electrooptic effect occurs in media of any symmetry isotropic media including. Variation in the refractive index at the quadratic electrooptic effect is proportional to the squared electric-field strength ($\Delta n \square E^2$). Consequently, the nonlinear polarization of a medium associated with the Kerr effect is cubic

$$P_{nl} = \chi^{(3)} E^3 . \qquad (2.1.49)$$

In solids the effect is associated with the induced polarization due to redistribution of bound electric charges, in liquids and gasses it is mainly due to the molecular orientation in the external field when , from the optical viewpoint, the material behaves like a uniaxial crystal.

Let us consider a medium, in the liquid or gaseous phase, composed of anisometric molecules with the polarization dependent on the observation direction. In the absence of an electromagnetic field, the molecules are chaotically oriented and hence the refractive index is isotropic. The light-wave field interacting with the molecules lines up the dipole moments and variations of the refractive index in this case are described by

$$\Delta n \cong \frac{1}{15 n_0} \left(\frac{n_0^2 + 2}{3} \right)^4 \frac{4\pi}{3} N \frac{(\alpha_{ZZ} - \alpha_{XX})^2}{kT} E^2 , \qquad (2.1.50)$$

where N - molecular concentration, T – temperature, α_{ZZ} and α_{XX} - polarizabilities of the molecules along the corresponding axes. The quantity $\alpha_{ZZ} - \alpha_{XX}$ is proportional to the molecular dipole moment and so the product $(\alpha_{ZZ} - \alpha_{XX})E$ gives the orienting force. On the other hand, thermal motion interferes with a strict orientation over the field as demonstrated by the factor kT in the denominator. Note that formula (2.1.50) is true for the electric field strengths not associated with the saturation effect. In this case variations of the refractive index are described by formula (2.1.18), similar to thermal nonlinearity ($\Delta n = n_2 E^2$). For high values of the field strength E, the molecules are actually lined up over the field and the function $\Delta n(E)$ is saturated. The nonlinearity

parameter n_2 is rather minor. To illustrate, for carbon bisulfide (CS_2) that is one of the best Kerr media $n_2 = 1{,}7 \cdot 10^{-11}$ CGS electrostatic units. In this case variation in the refractive index for the intensity $I = 1$ MW/cm^2 is only $\Delta n \cong 5 \cdot 10^{-8}$, i.e. by three orders of magnitude lower than for resonance or thermal nonlinearities. The situation is changed when we use picosecond pulses at $I = 1$ GW/cm^2. An increase in the intensity with shortening of the pulse length has minor effect on thermal or resonance variations of the refractive index because of their fast response (level of nano- or microseconds). Fast response of a Kerr medium is determined by relaxation of the orientation of polar molecules. As demonstrated by Debye, relaxation time is related to viscosity, and we use the so-called time of Debye relaxation $\tau_D = 4\pi\eta d^3 / kT$, where η - viscosity, d - size of a molecule, T – temperature. Calculating the Debye relaxation time for carbon bisulfide ($\eta = 3{,}8 \cdot 10^{-4}$ Pa · s, $kT = 1{,}38 \cdot 10^{-23} \cdot 300 \cong 4 \cdot 10^{-21}$ J, $d \sim 10^{-10}$ m), we get $\tau_D \sim 10^{-12}$ s, that allows for effective operations with picosecond laser pulses.

Apart from the orientational Kerr effect, there exists the electronic Kerr effect (electronic hyperpolarizability). It is associated with deformation of the electronic molecular shell under the action of an electromagnetic field and is reached virtually immediately ($\tau \sim 10^{-14}$- 10^{-15}s), this time being comparable to the light-wave oscillation period. The electronic Kerr effect is of paramount importance for molecules having no intrinsic dipole moment. But its value is lower by one-two orders of magnitude than for the orientational Kerr effect ($n_2 \sim 10^{-13}$ CGS electrostatic units). The electronic Kerr effect is revealed in liquids and gasses composed of the molecules having their intrinsic dipole moment, though in this case a relative contribution from the electronic polarizability into the total Δn is, as a rule, below ~10%.

2.1.4. Electrostriction nonlinearity

It is known that a high-intensity electromagnetic wave exerts pressure on a medium that results in changes of the medium density and hence in the refractive index Δn with the value given by

$$\Delta n = \frac{n_0 \beta}{2\pi} \left(\rho \frac{\partial n}{\partial \rho} \right)^2 E^2, \qquad (2.1.51)$$

where ρ- density, β - coefficient of volumetric compression. For the typical material parameters $n_2 \sim 10^{-11} - 10^{-12}$ CGS electrostatic units, i.e. it is close to the orientational Kerr effect. Still note that the electrostriction effect is inertial. A

speed of response is determined by the acoustic-wave propagation (similar to thermal nonlinearity) $t_{acoustic} \sim L/v_{acoustic}$, where L - illumination region, $v_{acoustic}$ - speed of an acoustic wave. Because of this, the electrostriction effect is observed on the interaction between laser pulses and transparent media (absorption factor $k < 0.01$ cm^{-1}), in the case of absorbing media being generally suppressed by thermal nonlinearity.

2.1.5. Photorefractive nonlinearity

The photorefractive effect is caused by the formation, due to nonuniform illumination, of a space charge, whose field changes the refractive index because of the electrooptic effect. The photorefractive effect was discovered in 1965 in the process of laser beams propagation in electrooptic crystals used for nonlinear optics, when it has been found that the propagating light beams are distorted. Initially, the effect was regarded to be adverse until its use for recording of dynamic gratings.

Let us consider the main stages of the formation of photorefractive nonlinearity. First, due to the effect of light, the electrons in electrooptic crystals are leaving the valence band for the conduction band. Then, in the process of their motion due to diffusion (or drift under the action of an electric field) within the crystal, the electrons are trapped with a certain probability by the traps positioned in the forbidden energy band. As the transition of the electrons from the valence band to the conduction band takes place only in the illuminated spatial areas and trapping is realized over the whole volume of a crystal, the electric charge in this crystal is distributed inhomogeneously. A gradient of the electric charge creates an intrinsic electric field leading, owing to the Pockels or the Kerr effect, to variations in the refractive index of the electrooptic crystal.

In photorefractive crystals two most important mechanisms of the formation of a space-charge field are possible: diffusion and drift. With a diffusion mechanism, the light-excited electrons are moving from the illuminated areas, where their concentration is higher, to nonilluminated areas with lower concentrations of the carriers and are trapped. The charge and space-charge field distribution, when the radiation intensity distribution is sinusoidal, is shown for this mechanism in Fig. 2.1.4. Here L_d – diffusion length that is a specific average distance covered by the electrons from the excitation point to the trapping point.

The second mechanism for the space-charge field formation is drift. As distinct from the diffusion mechanism, in this case motion of the photoexcited electrons is realized in the external electric field E_0 applied to a photorefractive

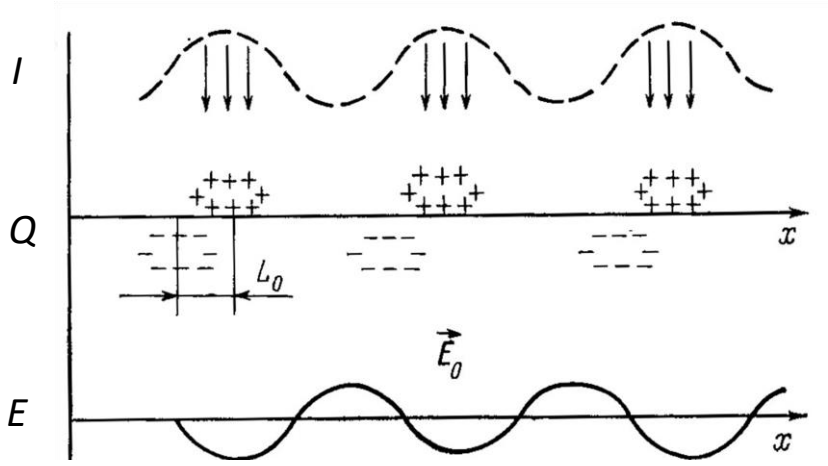


Fig. 2.1.4. Distribution of the electric charge q and of the electric field strength E when a space-charge field in a photorefractive crystal is formed by a diffusion mechanism

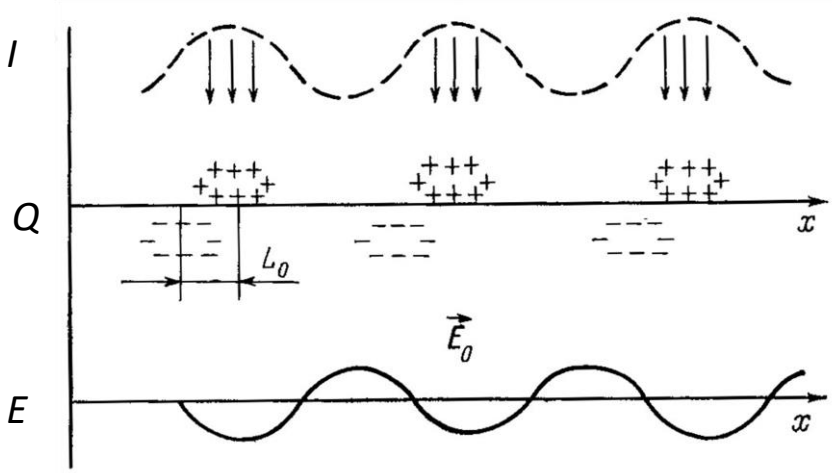


Fig. 2.1.5. Distribution of the electric charge q and of the electric-field strength E when the formation of a space-charge field in a photorefractive crystal is realized by a drift mechanism

crystal for amplification of the photorefractive effect. As this takes place, the electrons are moving in the same direction, on the average covering some characteristic distance L_d until the time of trapping. This distance L_d is referred to as the drift length of transport (Fig. 2.1.5). The electrons are drifting under the action of an external field till the moment when the space-charge drift in the illuminated areas compensates the applied field, that may be on the order of several dozens of kilovolt per centimeter.

The process of the photo-excited charge transport involves both electrons and holes. Since, as a rule, mobility of the holes is significantly lower than that of

the electrons and the lifetime is shorter, their contribution into photorefractive nonlinearity is generally small.

In conclusion, note that an increasing interest to photorefractive crystals is due to the possibility to use low-power light-fields (μW, nW) but the response in this case takes from milliseconds and more to several minutes.

2.2. Light beam self-focusing and autocollimation

The phenomenon of self-focusing was predicted by the Soviet physicist G. Askaryan in 1960 practically immediately after the advent of lasers. The essence of this phenomenon is that in the case of a light beam propagation in a nonlinear medium with the refractive index $n = n_0 + n_2 E^2$, when $n_2 > 0$, the refractive index in the center of the beam is greater than at the edges. And a plane-parallel layer of such a medium has the properties of a focusing lens because an optical length of the medium in the beam center is greater than at the edges. Another model is associated with a light beam having the rectangular profile. At the boundaries of such a beam, when it propagates within a nonlinear medium, the interface is formed between the regions with different refractive indices, the refractive index in the central region being higher than that at the edges. In this case the total internal reflection effect (i.e., the light beam is not extended beyond its boundaries) is expected.

Despite the fact that the self-focusing effect was predicted in 1960, it was observed for the first time in 1966. The researchers have recorded capture of the beam by self-focused filaments with the diameter 5 – 10 μm and waveguide propagation along the filaments. To understand, what was the problem with detection of self-focusing, we analyze theoretically the light beam propagation in a nonlinear medium.

We describe the nonlinear effects by the classical wave equation

$$\frac{\partial^2 E}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = \frac{4\pi}{c^2}\frac{\partial^2 P}{\partial t^2}\ , \qquad (2.2.1)$$

where c – speed of light in the vacuum, and light fields are represented in a complex form as

$$E(\omega) = \frac{1}{2}\left(A\exp(i(kz - \omega t)) + A^* \exp(-i(kz - \omega t))\right). \qquad (2.2.2)$$

Next we use the approximation of slowly varying amplitudes: variation of the light field amplitude is minor for time on the order of the light oscillation

period ($dA/dt << \omega A$) or for distances on the order of the light wave length ($dA/dz << kA$). Then for the stationary mode interaction a reduced wave equation of a light field is of the following form:

$$\Delta_{\perp} A + 2ik\frac{\partial A}{\partial z} = -\frac{2k^2 \Delta n}{n_0} A, \qquad (2.2.3)$$

where $\Delta_{\perp} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ – transverse Laplacian, $A = ae^{i\varphi}$ – complex light-wave amplitude, $\Delta n = n_2 E^2$ – nonlinear variations of the refractive index.

For simplicity of our theoretical analysis, expression (2.2.3) is transformed and we have the following equations for the amplitude and phase of a light wave:

$$\frac{\partial \phi}{\partial z} + \frac{1}{2k}\left(\frac{\partial \phi}{\partial x}\right)^2 + \frac{1}{2k}\left(\frac{\partial \phi}{\partial y}\right)^2 - \frac{k}{2}\left(\frac{\partial^2 a/\partial x^2 + \partial^2 a/\partial y^2}{k^2 a} + \frac{2\Delta n}{n_0}\right) = 0, \qquad (2.2.4)$$

$$k\frac{\partial a^2}{\partial z} + \frac{\partial}{\partial x}(a^2 \frac{\partial \varphi}{\partial x}) + \frac{\partial}{\partial y}(a^2 \frac{\partial \varphi}{\partial y}) = 0. \qquad (2.2.5)$$

Equation (2.2.4) determines the light beam path and equation (2.2.5) determines variations in the spatial intensity distributions of a light field. Expression $\frac{\partial^2 a/\partial x^2 + \partial^2 a/\partial y^2}{k^2 a}$ in equation (2.2.4) describes the diffraction effect (diffraction-limited divergence). This expression is below zero and it is compensated due to nonlinear variations in the refractive index $\frac{2\Delta n}{n_0} > 0$. When we have

$$\frac{\partial^2 a/\partial x^2 + \partial^2 a/\partial y^2}{k^2 a} + \frac{2\Delta n}{n_0} = 0 \qquad (2.2.6)$$

and in the particular point z the wave front is plane ($\partial\phi/\partial x = 0$, $\partial\phi/\partial y = 0$), this wave front remains plane in any other point z because $\partial\phi/\partial z = 0$. Equation (2.2.6) is thought to be the light-beam autocollimation condition. If expression (2.2.6) is below zero, the beam is divergent and its intensity and hence nonlinear variations in the refractive index in the process of propagation are lowered. When expression (2.2.6) is above zero, the beam is focused, and the narrower the beam the higher the electric field strength and hence the nonlinear variations in the refractive index, even greater contributing to the light beam compression. So, equation (2.2.6) sets a threshold condition for realization of the self-focusing effect.

For our analysis we further select the light beam Gaussian profile $a = a_0 \exp(-x^2/2\rho^2)$, where ρ – beam radius. Taking the second derivatives with respect to the coordinates and considering that $\Delta n = n_2 a^2$, we obtain the condition for the threshold amplitude in the center of a light beam as follows:

$$a_{0\,th}^2 = \frac{n_0 \lambda^2}{4\pi^2 \rho^2 n_2}. \qquad (2.2.7)$$

Then the threshold intensity in the beam center is given by

$$I_{0_{th}} = \frac{c n_0}{8\pi} a_{0_{th}}^2 = \frac{c\lambda_0^2}{32\pi^3 \rho^2 n_2}, \qquad (2.2.8)$$

where $\lambda_0 = n_0 \lambda$ - wave length in the vacuum.

The threshold light beam intensity is easily found by multiplication of the intensity in the beam center into its area

$$P_{th} = I_{0_{th}} \pi \rho^2 = \frac{c\lambda_0^2}{32\pi^2 n_2}. \qquad (2.2.9)$$

Substituting the parameters of a nonlinear medium, e.g. carbon bisulfide ($n_2 = 1{,}7 \cdot 10^{-11}$ CGS electrostatic units), for the wave length 0.5 μm, we get the threshold power $P_{th} \sim 10^{10}$ Erg/s or $P_{th} \sim 1$ kW. This and much higher values of the power of laser radiation were obtained in the beginning of the 60ies of the XX century but the researchers failed in realization of the self-focusing effect.

To elucidate the reason, why this was the case, let us consider the so-called self-focusing length – distance at which the light beam compression (focusing) occurs. To this end, it is convenient to consider the condition for realization of the total internal reflection at the interface of two media with the refractive indices n_0 and $n = n_0 + \Delta n$.

When a light beam is incident on the interface at the angle α (Fig.2.2.1), the total internal reflection condition is of the following form:

$$\sin\alpha_{tir} = n_0/(n_0 + \Delta n). \qquad (2.2.10)$$

Then for the angle Θ characterizing divergence of a light beam, with regard to $\cos\Theta = \sin\alpha$ and considering a series expansion of $\cos\Theta \cong 1 - \Theta^2/2$, we have

$$\Theta_{tir}^2 \cong 2\Delta n/n_0. \qquad (2.2.11)$$

If the light beam divergence is below the angle Θ_{tir} value, the propagation of such a light beam is divergence free.

On the other hand, the angle Θ_{tir} indicates an angle by which the light beam is deflected at the interface of two media, see Fig. 2.2.1. It is expected that the

forward propagating light beam in a nonlinear medium should be deflected by the same angle. As seen in Fig. 2.2.2, in this case the distance, where compression of a light beam occurs, is approximately given by

$$l_{SF} = \rho \Big/ \sqrt{2\Delta n / n_0} \,. \qquad (2.2.12)$$

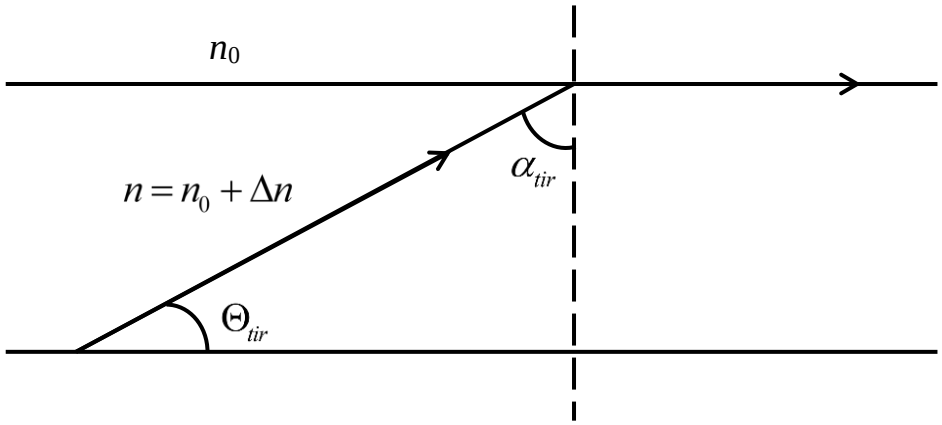


Fig. 2.2.1. Total internal reflection at the interface of two media with the refractive indices n_0 and $n = n_0 + \Delta n$

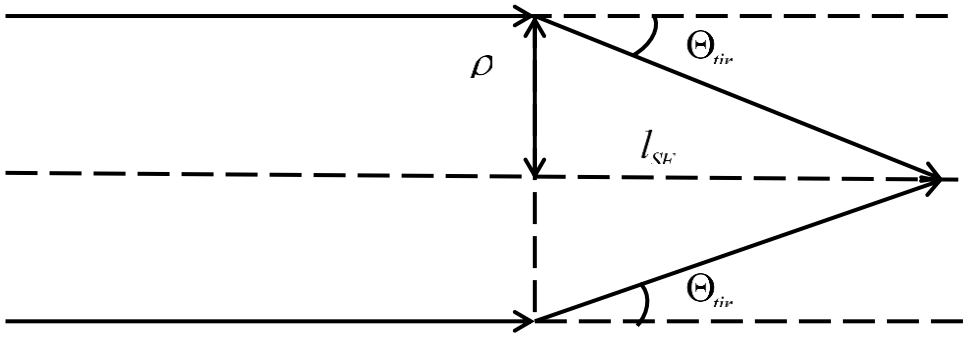


Fig. 2.2.2. Total internal reflection at the interface of two media with the refractive indices n_0 and $n = n_0 + \Delta n$

Considering that ρ and with regard to formulae (2.2.7) – (2.2.9), we can write the following condition for the self-focusing length:

$$l_{SF} = \frac{\rho^2 n_0}{4} \sqrt{\frac{c}{n_2 P}}, \qquad (2.2.13)$$

where P - light beam power.

The above formula was derived disregarding the diffraction-limited divergence and it is valid for the light beam powers significantly exceeding the

threshold power P_{th} (2.2.9). Approximately, the requirement for the light beam power exceeding the threshold one may be taken into account if we introduce the power difference into the denominator

$$l_{SF} = \frac{\rho^2 n_0}{4}\sqrt{\frac{c}{n_2}}\frac{1}{\sqrt{P}-\sqrt{P_{th}}}. \qquad (2.2.14)$$

To obtain a qualitative estimate, we use the earlier-mentioned parameters of a nonlinear medium: $n_2 = 1.7\cdot 10^{-11}$ CGS electrostatic units, refractive index n_0 = 1.63, light beam radius ρ=0.5 cm, and radiation power P = 1 MW/cm^2. In this case the self-focusing length is more than 10 m, that clarifies the situation with a failure in detection of this effect. Proceeding from (2.2.14) and from the quadratic dependence of the self-focusing length on the light beam radius, it is inferred that the self-focusing length may be reduced considerably by decreasing the light beam radius. To illustrate, on going to the use of a beam with the radius 0.5 mm the self-focusing length comes to 10 см. Another approach to shortening of the self-focusing length is an increase of the nonlinearity parameter and of the light beam power. However, in this case the dependence is less distinct (square root). For a medium with resonance nonlinearity ($n_2 \sim$ -10^{-8} CGS electrostatic units), with a beam having the radius 0.5 mm we can get the self-focusing length $l_{SF} \approx$ 0.5 cm.

In real experimental situations the threshold power and self-focusing length may deviate from the above-mentioned values due to the formation of waveguide modes in a nonlinear light-guide channel and also due to the saturation effect of nonlinear variations in the refractive index or fine-scale focusing effects, etc. Nevertheless, the considered relationships of the light beam self-focusing allow for estimation of the conditions associated with the self-focusing effect development. It should be noted that in the case of self-focusing a diameter of the light beam is as small as several microns leading to the intensities in excess of GW/cm^2 even for a comparatively low power of laser radiation (~ kW). As this takes place, the self-focusing effect can result in destruction of the medium itself as it has been observed on propagation of laser pulses in solids, e.g. in optical elements of lasers. On the other hand, realization of the autocollimation effect makes it possible to form waveguide channels. Such structures are termed as spatial solitons; during the period of 20 years these structures have received much attention of the researchers. Fig. 2.2.3 demonstrates the condition for their formation. As seen, a spatial soliton (*c*) is formed when the diffraction-limited

divergence responsible for the light beam expansion (*a*) is compensated by the self-focusing effect (*b*).

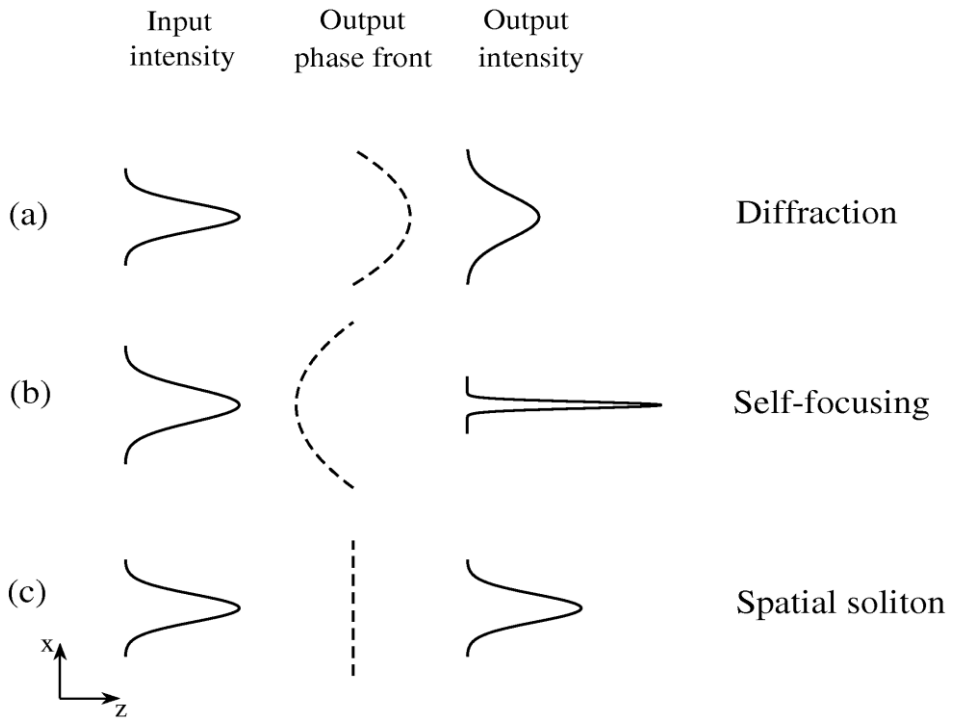


Fig. 2.2.3. Formation of a spatial soliton in the conditions when the diffraction-limited divergence is compensated for by the self-focusing effect

As this takes place, the light beam retains its plane wave front on propagation in a nonlinear medium. Note that there is no need to use high-power light beams for the formation of spatial solitons. Owing to the use of photorefractive crystals as a nonlinear medium, one can realize the autocollimation mode of continuous-wave laser radiation even at the powers at a nano- and microwatt-level.

In conclusion, note that in nonlinear media with the refractive index $n = n_0 + n_2 E^2$, where $n_2 < 0$, we observe the inverse effect – light beam defocusing. This situation is typical for, e.g., media with thermal nonlinearity. Since in the beam center, where the intensity is maximal, the refractive index is lower as compared to the edges, within the medium a negative lens is formed to increase the light beam divergence. As distinct from the self-focusing effect, the defocusing effect is threshold-free. We can use equations (2.2.3) – (2.2.5) for its description with due regard for the fact that nonlinear variations of the refractive index are given as $\Delta n < 0$. In this case expression (2.2.6) is also below zero.

To estimate the focal length of a lens, we use an approach similar to that used for the derivation of formula (2.2.12) and can write the following expression:

$$f_{DF} = -\rho \Big/ \sqrt{2|\Delta n| / n_0} \ . \qquad (2.2.16)$$

Assuming that defocusing is caused by thermal nonlinearity and heating of a medium results in an increase of the temperature by 0.1° (realized for an alcoholic solution of the dye at the energy density of laser radiation 0.01 J/cm^2 and at the absorption factor $k = 20$ cm^{-1}), the refractive index is varying as $\Delta n = -4 \cdot 10^{-5}$. Consequently, the focal length of a negative lens for a beam with the radius 0.5 mm is greater than 50 cm. It is concluded that a marked effect of thermal defocusing is expected for the laser-radiation energy density above 1 J/cm^2.

Thus, when laser radiation is propagating in nonlinear media, we can expect for a light beam the development of the self-focusing, defocusing, and autocollimation effects. The self-focusing and autocollimation effects are observed at the light beam power that is equal to or higher than the threshold value in media with positive variations of the refractive index. As a rule, these are media with Kerr and resonant nonlinearity when the operation is realized in the short-wavelength spectral region with respect to the absorption profile maximum. In the case of media with thermal nonlinearity or resonance media on operation in the long wavelength region one can expect the defocusing effect of a light beam.

2.3. Second harmonic generation, phase matching condition

2.3.1. Second harmonic generation phenomenon

One of the first nonlinear effects detected on propagation of laser radiation in a material was observation of the second harmonic generation with doubling of the frequency of laser radiation. The first experiment was conducted by P.Franken and his co-workers in 1961, practically immediately after the advent of lasers. Focused radiation of a ruby laser (pulse energy 3 J, pulse length 1 ms) was directed to a thin plate of crystal quartz. Apart from the initial red laser radiation at the frequency ω (λ_1 = 0.694 μm), after the plate one could observe ultraviolet radiation at the doubled frequency (λ_2 = 0.347 μm). In the first experiments about 10^{-10} of the energy of the initial radiation was converted into the second harmonic energy. As a result of further studies, highly effective (featuring the conversion factor above 50 %) laser radiation frequency doublers and cascade frequency amplifiers for the third, fourth, and higher harmonics have been developed and found an extensive use in modern laser systems.

The second harmonic generation process is studied with the help of phenomenological models. From the quantum point of view, coalescence of two photons takes place in the presence of an atomic field (model of virtual levels). Eventually, from two photons at the frequency ω a photon at the frequency 2ω is generated. From the classical viewpoint, we can figure propagation of a light wave that affects the electron rotating about the atomic nucleus. Under the effect of an electromagnetic wave, the electron, apart from rotational motion about the nucleus, is involved in oscillatory motion with the frequency ω. An increase in the light wave intensity results in the increased amplitude of the electron oscillations which, due to the electron-electron interactions, are no longer sinusoidal (anharmonic oscillator model). In this case the periodic nonsinusoidal motion of the electron may be expanded in a Fourier series and represented as a set of sine-wave oscillations with multiple frequencies. When the electron oscillates at the frequency 2ω, the emitted electromagnetic wave is also at the frequency 2ω.

To describe qualitatively the second harmonic generation process, we perform a series expansion of the nonlinear medium polarization in terms of the electric field strength as follows:

$$P = \chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \dots , \qquad (2.3.1)$$

where $\chi^{(1)}$, $\chi^{(2)}$, $\chi^{(3)}$ – parameters of the medium characterizing its polarizability. The greatest nonlinear term here is polarization quadratic in the field ($P^{(2)} = \chi^{(2)}E^2$), just this polarization is responsible for the second harmonic generation. Note that, for isotropic media and crystals with inversion center, expression (2.3.1) has no even terms but nonlinearity is the highest – cubic $P^{(3)} = \chi^{(3)}E^3$. In these media the second harmonic generation is impossible.

Nonlinear effects we describe with the use of the classical wave equation

$$\frac{\partial^2 E}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = \frac{4\pi}{c^2}\frac{\partial^2 P}{\partial t^2} , \qquad (2.3.2)$$

where c – speed of light in the vacuum, and the light fields are represented in the complex form

$$E(\omega) = \frac{1}{2}\left(A\exp(i(kz - \omega t)) + A^* \exp(-i(kz - \omega t))\right). \qquad (2.3.3)$$

Considering the propagation of light fields at the fundamental frequency of lasing and at the second harmonic frequency in a medium with the quadratic nonlinearity, polarization of the medium takes the form

$$P = \chi^{(1)}E + \chi^{(2)}E^2 , \qquad (2.3.4)$$

where $E = \frac{1}{2}(E_1 \exp(-i\omega t) + E_2 \exp(-i2\omega t) + E_1^* \exp(i\omega t) + E_2^* \exp(i2\omega t))$ - total field of the interacting waves, $E_1 = A_1 \exp(ik_1 z)$, $E_2 = A_2 \exp(ik_2 z)$ - spatial parts of the electric field strength for a light wave.

In this case the quadratic nonlinear polarization $P^{(2)} = \chi^{(2)} E^2$ is represented as

$$P^{(2)} = \frac{1}{4}\chi^{(2)}(E_1^2 \exp(-i2\omega t) + E_2^2 \exp(-i4\omega t) + E_1^{*^2} \exp(i2\omega t) +$$

$$E_2^{*^2} \exp(i4\omega t) + 2E_1 E_2 \exp(-i3\omega t) + 2E_1 E_1^{\ *} + 2E_1 E_2^* \exp(i\omega t) + \quad (2.3.5)$$

$$+2E_2 E_1^{\ *} \exp(-i\omega t) + 2E_2 E_2^* + 2E_1^{\ *} E_2^{\ *} \exp(i3\omega t))$$

Then the medium polarization at the frequencies ω and 2ω is described by the following equations:

$$P_1(\omega) = \frac{1}{2}\chi^{(1)}\left(E_1 \exp(-i\omega t) + E_1^* \exp(i\omega t)\right) +$$

$$+\frac{1}{2}\chi^{(2)}\left(E_1 E_2^* \exp(i\omega t) + E_1^* E_2 \exp(-i\omega t)\right) \quad (2.3.6)$$

and

$$P_2(2\omega) = \frac{1}{2}\chi^{(1)}\left(E_2 \exp(-i2\omega t) + E_2^* \exp(i2\omega t)\right) +$$

$$+\frac{1}{4}\chi^{(2)}\left(E_1^2 \exp(-i2\omega t) + E_1^{*2} \exp(i2\omega t)\right) \quad (2.3.7)$$

Substituting expressions (2.3.6) and (2.3.7) into wave equation (2.3.2), we obtain a system of the second-order nonlinear differential equations solved by the numerical methods only. Because of this, in what follows we use an approximation of the slowly varying amplitudes: amplitude of a light field has minor variations during time on the order of the light oscillation period $dA/dt << \omega A$ or at the distances on the order of the light wave length $dA/dz << kA$. Then, in the stationary interaction mode, for plane waves we get a system of the reduced wave equations

$$\frac{\partial A_1}{\partial z} + \alpha_1 A_1 = i\frac{2\pi\omega}{cn}\chi^{(2)} A_1^* A_2 \exp(i\Delta k z), \quad (2.3.8)$$

$$\frac{\partial A_2}{\partial z} + \alpha_2 A_2 = i\frac{2\pi\omega}{cn}\chi^{(2)} A_1^2 \exp(-i\Delta k z), \quad (2.3.9)$$

where $\Delta k\, z = (k_2 - 2k_1)\, z$ – phase mismatch of the waves at the fundamental and double frequencies; α_1 and α_2 – amplitude absorption factors at these frequencies.

An exact solution for a system of equations (2.3.8) and (2.3.9), when $\alpha_1 = \alpha_2 = 0$, was obtained by N. Bloembergen in the form of elliptical integrals. When the phase synchronism condition is met ($\Delta k = 0$), a system of equations (2.3.8) and (2.3.9) with the boundary conditions $A_1(0) = A_0$ and $A_2(0) = 0$ takes the following form:

$$\begin{aligned} A_1 &= A_0 \operatorname{sech}\left(A_0\, \sigma L\right), \\ A_2 &= iA_0 \operatorname{th}\left(A_0\, \sigma L\right), \end{aligned} \tag{2.3.10}$$

where $\sigma = \dfrac{2\pi\omega}{cn}\chi^{(2)}$; A_0 – amplitude of a plane wave with the frequency ω at the input to a nonlinear medium; L – medium length. The obtained equations point to the fact that the fundamental frequency radiation is completely converted to the second harmonic.

It should be noted that a system of equations (2.3.8) and (2.3.9) may be solved analytically, with regard to the phase mismatch and wave absorption in a nonlinear medium, in the approximation of low factors for radiation conversion to the second harmonic. In this case from equation (2.3.8) we obtain

$$A_1 = A_0 \exp(-\alpha_1 z), \tag{2.3.11}$$

and from equation (2.3.9) it follows that

$$A_2 = \frac{i\sigma\left(A_0\right)^2}{2\alpha_1 - \alpha_2 + i\Delta k}\left[\exp(-\alpha_2 z) - \exp(-(2\alpha_1 + i\Delta k)z)\right]. \tag{2.3.12}$$

Then the wave intensity of the second harmonic at the output from a nonlinear medium is given by

$$\begin{aligned} I_2 = \frac{cn}{8\pi}\left|A_2\left(L\right)\right|^2 = \frac{32\pi^3\omega^2}{c^3n^3}\left|\chi^{(2)}\right|^2 \cdot I_0^2 L^2 \times \\ \times \frac{\exp(-2\alpha_2 L) - 2\exp(-(2\alpha_1 + \alpha_2)L)\cos(\Delta kL) + \exp(-4\alpha_1 L)}{(2\alpha_1 - \alpha_2)^2 L^2 + \left(\Delta kL\right)^2}, \end{aligned} \tag{2.3.13}$$

where $I_0 = \dfrac{cn}{8\pi}\left|A_0\right|^2$ – intensity of a wave with the frequency ω at the input to a nonlinear medium.

Expression (2.3.13) is simplified for low-absorption media ($\alpha_1 = \alpha_2 = 0$)

$$I_2 = \frac{32\pi^3\omega^2}{c^3 n^3}\left|\chi^{(2)}\right|^2 I_0^2 L^2 \frac{\sin^2(\Delta kL/2)}{(\Delta kL/2)^2} \ . \qquad (2.3.14)$$

As follows from equation (2.3.14), when the phase-matching condition is met $\Delta k = k_2 - 2k_1 = \frac{2\omega}{c_0}(n_2 - n_1) = 0$, an intensity at the output from the crystal is growing in direct proportion to L^2. Increasing the crystal length L, one can attain a significant factor of the fundamental wave conversion to the second harmonic wave. However, in the case of dispersion ($n_1 \neq n_2$, $\Delta k \neq 0$) the relationship between the second harmonic intensity and the medium length is a periodic fucntion (Fig. 2.3.1), with the maxima meeting the condition $\Delta k\ L = \pi, 3\pi, 5\pi \ldots$

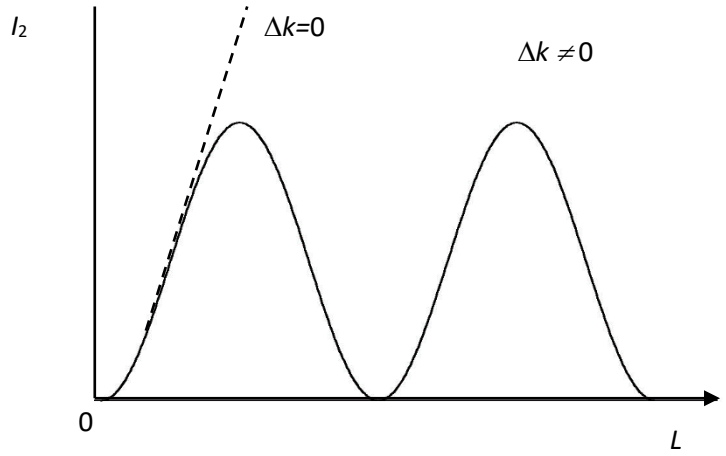


Fig. 2.3.1. Second harmonic intensity as a function of length of a nonlinear crystal

The first maximum is reached at a minimal length of the crystal $L_{\text{coherent}} = \lambda/4\,(n_2 - n_1)$, that is called the coherent interaction length L_{coherent}, where λ – wave length of laser radiation. This length in optically transparent crystals comes to 10–20 wave lengths and is considerably less than an ordinary length of crystals (several millimeters or centimeters). Shortness of the coherent interaction length in turn dictates a low efficiency of radiation conversion to the second harmonic, as it has been observed in the first experiments. For example, rotation of a quartz plate with the thickness $L = 750$ μm caused periodic fluctuations of the second harmonic intensity I_2 for the conversion efficiency about $10^{-8} - 10^{-10}$.

From the viewpoint of physics, these fluctuations are due to different phase velocities of photons at the frequency ω and 2ω. If a medium has a normal dispersion, the refractive index at double frequency is higher than that at the

fundamental one ($n_2 > n_1$). Consequently, the phase velocity of photons at the double frequency is lower ($v_\phi^{2\omega} < v_\phi^{\omega}$). In this way, as the wave is propagating at the frequency ω, other phase-shifted waves are formed within the crystal at the frequency 2ω. When the phase shift is greater than π, we observe interference attenuation; when the phase shift between the waves equals 2π, interference quenching takes place and we have $I_2 = 0$. In this case, without absorption, all energy again goes to the wave with the frequency ω.

2.3.2. Phase-matching condition

The interaction-efficiency improvement problem was solved by the American physicists J. Gordmane and R. Terhune in 1962. They suggested to use for the second harmonic generation birefringent uniaxial crystals, where the phase-matching condition may be achieved ($\Delta k = 0$). Ordinary and extraordinary waves are propagating in such crystals with different velocities. The refractive index surface cross-sections of the ordinary n_o and of the extraordinary n_e waves in a uniaxial negative ($n_o > n_e$) crystal are shown in Fig. 2.3.2. Thin lines denote the frequency ω (refractive index n_1), thick lines – the double frequency 2ω (refractive index n_2); the optical axis is denoted as Z.

Current values of the extraordinary refractive indices (for the arbitrary angle Θ between the wave vector and the crystal optical axis Z) are denoted by n with the superscript «*e*» (n^e). In Fig. 2.3.2 the curves n_{1o} and n_2^e are intersecting each other. Their intersection points are associated with the directions, for which the phase-matching condition is fulfilled between an ordinary wave with the frequency ω and its harmonic (extraordinary wave) with the frequency 2ω. These directions are termed the phase-matching directions, and the angle Θ_S between the directions and the crystal optical axis – phase-matching (or synchronism) angle. The phase-matching directions are lying on the surface of the cone of revolution about the crystal axis at the cone angle Θ_S. As seen in Fig. 2.3.2, phase matching is the case in the direction determined by the angle Θ_S^{ooe} ($n_{10} = n_2^e$), when two ordinary waves at the frequency ω are summed up to form an extraordinary wave at the frequency 2ω («*ooe*» - interaction).

Knowing the principal values of the refractive indices n_{1o}, n_{2o}, and n_{2e}, we can calculate the phase-matching angle Θ_S^{ooe}. Because section of the refractive

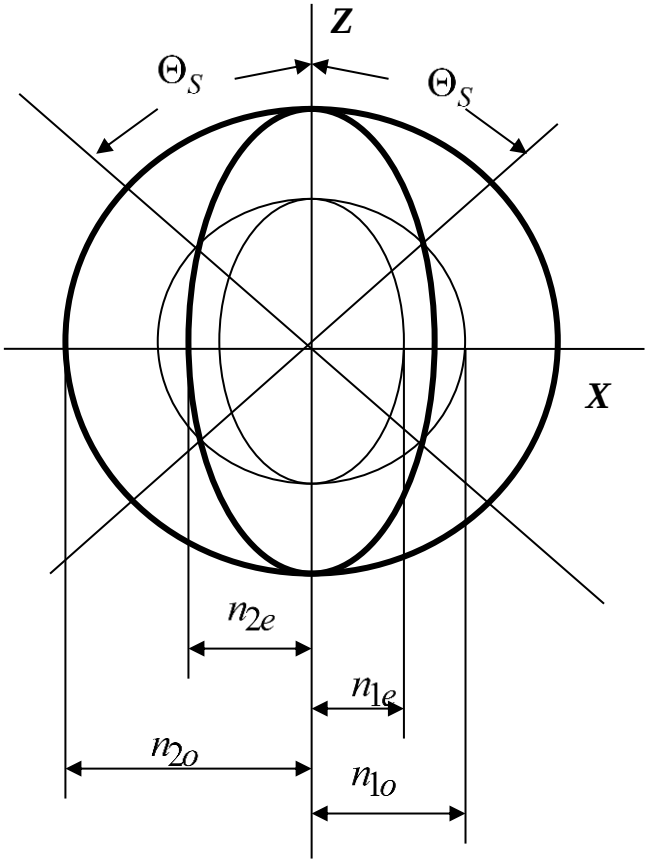


Fig. 2.3.2. The refractive index surface cross-sections
for a uniaxial negative crystal

index surface by the figure plane is a circle or an ellipse, we have the following equalities:

$$n_{1o}(\Theta) \equiv n_{1o}; \qquad n_2^e(\Theta) = \frac{n_{2e}}{\sqrt{1-\varepsilon_2^2 \cos^2\Theta}}, \tag{2.3.15}$$

where $\varepsilon_2 = \sqrt{1-\left(n_{2e}/n_{20}\right)^2}$ – eccentricity of ellipse. Substituting (2.3.15) into the phase-matching condition $n_{1o} = n_2^e$, we obtain

$$\cos^2\Theta_c^{ooe} = \frac{1}{\varepsilon_2^2}\left[1-\left(\frac{n_{2e}}{n_{1o}}\right)^2\right]. \tag{2.3.16}$$

Along with «*ooe*» - interaction in specific conditions, the possibility exists for the phase-matched generation of the second harmonic on the interaction of an ordinary and an extraordinary wave of fundamental radiation with an extraordinary wave of the second harmonic («*oee*» – interaction). When the waves

are propagating in the same direction, the phase-matching condition for this type of interactions is of the form $k_{1o} + k_1^e = k_2^e$, from where we have

$$\frac{n_{10} + n_1^e}{2} = n_2^e \ , \qquad (2.3.17)$$

and the phase-matching angle, to within a significant accuracy, is equal to

$$\cos^2 \Theta_c^{oee} \cong 2\,\frac{\left(n_{10} + n_{1e}\right)/2 - n_{2e}}{n_{2e}\varepsilon_2^2 - n_{1e}\varepsilon_1^2/2}, \qquad (2.3.18)$$

where $\varepsilon_i = \sqrt{1-\left(n_{ie}/n_{i0}\right)^2}$, $i = 1, 2$.

2.3.3. Phase-matching angular width

Realizing the second-harmonic generation effect, one should take into consideration that laser radiation possesses finite divergence and hence the phase-matching condition could not be fulfilled simultaneously for the whole beam of fundamental radiation.

The relationship between the second-harmonic intensity and the crystal orientation angle Θ may be derived from expressions (2.3.13) or (2.3.14) by finding the phase mismatch $\Delta k = \dfrac{2\omega}{c_0}\Delta n \quad (\Delta n = n_2 - n_1)$. For «*ooe*» -interaction, we differentiate expression (2.3.15) with respect to the angle Θ close to the phase-matching direction $\left(|\beta| = |\Theta - \Theta_c| \langle\langle 1\right)$ to get

$$\Delta n^{ooe} = n_2^e - n_{10} \cong \left.\frac{\partial n_2^e}{\partial \Theta}\right|_{\Theta_c} \beta = -\frac{n_{2e}\varepsilon_2^2\left(\sin 2\Theta_c\right)}{2(1-\varepsilon_2^2\cos^2\Theta_c)^{3/2}}\beta = \gamma^{ooe}\beta \ . (2.3.19).$$

For «*oee*» - interaction, the phase mismatch is determined as $\Delta n = n_2^e - \left(n_{1o} + n_1^e\right)/2$ and connected to the angle β by the following relation:

$$\Delta n^{oee} \cong \left.\frac{\partial n_2^e}{\partial \Theta}\right|_{\Theta c}\beta - \frac{1}{2}\left.\frac{\partial n_1^e}{\partial \Theta}\right|_{\Theta c}\beta \cong$$

$$\cong \frac{1}{2}\left[\frac{n_{1e}\varepsilon_1^2}{2\left(1-\varepsilon_1^2\cos^2\Theta_c\right)^{3/2}} - \frac{n_{2e}\varepsilon_2^2}{\left(1-\varepsilon_2^2\cos^2\Theta_c\right)^{3/2}}\right]\left(\sin 2\Theta_c\right)\beta = \gamma^{oee}\beta. \qquad (2.3.20)$$

Fig. 2.3.3 shows the curve for the second-harmonic generation as a function of the angle β close to the phase-matching direction that is calculated according

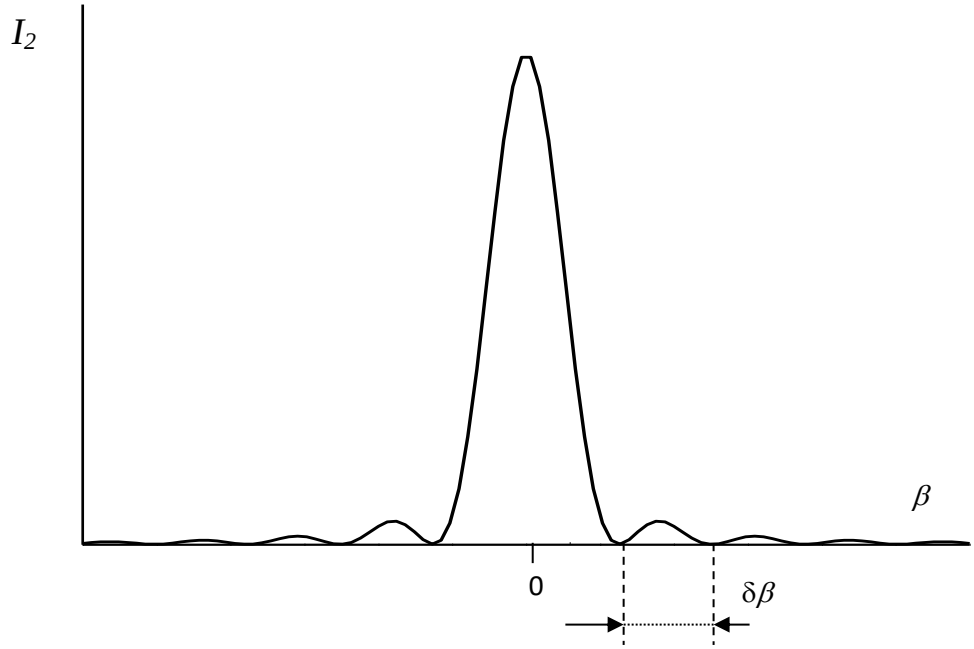


Fig. 2.3.3. Second harmonic intensity as a function of crystal orientation angle

to (2.3.14). Distances between the minima are found from the condition of the vanishing second-harmonic generation intensity I_2 (2.3.14) that is met when the condition $\frac{\Delta kL}{2} = \frac{2\pi L}{\lambda}\Delta n = \pi; 2\pi; 3\pi \ldots$ is fulfilled. From this it follows that, with regard to expressions (2.3.19) and (2.3.20), a width of each band (angular distance between the adjacent minima) equals

$$\delta\beta = \frac{\lambda}{2\gamma L}\ , \qquad (2.3.21)$$

where $\gamma = \left|\gamma^{ooe}\right|$ or $\gamma = \left|\gamma^{oee}\right|$ is selected depedning on the interaction type.

The quantity $\delta\beta$ is termed the angular phase-matching width. As seen from Fig. 2.3.3, the angular phase-matching width is a half of the central maximum determining the phase-matching direction. Therefore, for the efficient conversion of radiation to the second harmonic, divergence of a light beam should be lower than the angular phase-matching width.

When the beams with the divergence exceeding the angular phase-matching width are involved, one should consider the angular intensity distribution of fundamental radiation. In the Gaussian distribution approximation the angular intensity density of a light beam incident on a crystal is given by the expression

$$\rho_1(\beta) = \partial I_1/\partial\beta = I_1 \exp\left[-(\beta-\beta_0)^2/\psi^2\right]/\sqrt{\pi}\,\psi, \qquad (2.3.22)$$

where β_0 – axial orientation of the beam with respect to the phase-matching direction; ψ – light beam divergence. Using expression (2.3.22), by manipulations with expression (2.3.13) derived for a plane light wave, considering the phase mismatch and wave absorption, we obtain the following equation of the form suitable for a diverging beam:

$$I_2 = \frac{32\pi^2\omega^2}{c^3 n^3 \psi^2}\left|\chi^{(2)}\right|^2 L^2 I_1^2 \int_{-\infty}^{\infty}\int \frac{\exp(-2\alpha_2 L) - 2\exp(-(2\alpha_1+\alpha_2)L)\cos(\Delta kL) + \exp(-4\alpha_1 L)}{(2\alpha_1-\alpha_2)^2 L^2 + (\Delta kL)^2} \times$$

$$\times \exp\left(-\left((\beta_1-\beta_0)^2 + (\beta_2-\beta_0)^2\right)/\psi^2\right)\partial\beta_1\,\partial\beta_2, \qquad (2.3.23)$$

where the phase mismatch is given by $\Delta kL = (k_2 - 2k_1)L = \omega\gamma L(\beta_1+\beta_2)/c$. In this case equation (2.3.23) takes into account that the formation of a photon at the ferquency 2ω involves two photons at the frequency ω, their propagation directions being determined by the angles β_1 and β_2.

A numerical solution of expression (2.3.23) demonstrates a significant lowering of the radiation conversion efficiency to the second harmonic when the divergence of laser radiation is in excess of the angular phase-matching width. Proceeding from (2.3.21), we can find a limitation on the nonlinear crystal length, for which the conversion efficiency is close to a maximal value in the approximation of a quasi-linear light beam

$$L_{\max} = \frac{\lambda}{2\gamma\psi}, \qquad (2.3.24)$$

where ψ - divergence of laser radiation. The quantity $\gamma = \left|\gamma^{ooe}\right|$ or $\gamma = \left|\gamma^{oee}\right|$ is dependent on the interaction type according to expressions (2.3.19), (2.3.20).

2.4. Parametric amplification and generation

2.4.1. Frequency summation or subtraction in media with quadratic nonlinearity

In the previous section we have considered the case of the second harmonic generation effect when two photons at the frequency ω are summed up to generate

photons at the frequency 2ω. In this section we consider the wave interaction at different frequencies ω_1 and ω_2 in a medium with quadratic nonlinearity

$$P^{(2)} = \chi^{(2)} E^2, \tag{2.4.1}$$

where $E = \frac{1}{2}\left(E_1 \exp(-i\omega_1 t) + E_2 \exp(-i\omega_2 t) + E_1^* \exp(i\omega_1 t) + E_2^* \exp(i\omega_2 t) \right)$ - total field of the interacting waves, $E_1 = A_1 \exp(ik_1 z)$, $E_2 = A_2 \exp(ik_2 z)$ - spatial parts of the electric field strength for a light wave.

The quadratic nonlinear polarization in this case is represented by

$$P^{(2)} = \frac{1}{4}\chi^{(2)}(E_1^2 \exp(-i2\omega_1 t) + E_2^2 \exp(-i2\omega_2 t) + E_1^{*2} \exp(i2\omega_1 t) + E_2^{*2} \exp(i2\omega_2 t) +$$

$$+2E_1E_2 \exp(-i(\omega_1+\omega_2)t) + 2E_1E_1^{\ *} + 2E_1E_2^* \exp(-i(\omega_1-\omega_2)t) +$$

$$+2E_2E_1^{\ *} \exp(i(\omega_1-\omega_2)t) + 2E_2E_2^* + 2E_1^{\ *}E_2^{\ *} \exp(i(\omega_1+\omega_2)t). \tag{2.4.2}$$

As seen from the above expression, the interaction of two waves at the frequencies ω_1 and ω_2 results in the generation of new waves at other frequencies. Apart from the second harmonic generation ($2\omega_1$ and $2\omega_2$), there also appear the waves at the frequencies $\omega_1 + \omega_2$ and $\omega_1 - \omega_2$. These effects are termed the sum- or difference-frequency generation of the waves. To find, which of the four indicated frequencies are generated at the output from the medium, we write the reduced wave equations similar to those used when considering the second harmonic generation:

$$\frac{\partial A(2\omega_1)}{\partial z} \exp(ik(2\omega_1)z) = i\frac{2\pi\omega_1}{cn}\chi^{(2)} A_1^2 \exp(i2k(\omega_1)z), \tag{2.4.3}$$

$$\frac{\partial A(2\omega_2)}{\partial z} \exp(ik(2\omega_2)z) = i\frac{2\pi\omega_2}{cn}\chi^{(2)} A_2^2 \exp(i2k(\omega_2)z), \tag{2.4.4}$$

$$\frac{\partial A(\omega_1+\omega_2)}{\partial z} \exp(ik(\omega_1+\omega_2)z) = \\ = i\frac{2\pi(\omega_1+\omega_2)}{cn}\chi^{(2)} A_1 A_2 \exp(i(k(\omega_1)+k(\omega_2))z) \tag{2.4.5}$$

$$\frac{\partial A(\omega_1-\omega_2)}{\partial z} \exp(ik(\omega_1-\omega_2)z) = \\ = i\frac{2\pi(\omega_1-\omega_2)}{cn}\chi^{(2)} A_1 A_2^* \exp(i(k(\omega_1)-k(\omega_2))z) \tag{2.4.6}$$

To realize the effective transformation, the exponents in both parts of these equations should be coincident. In this case the wave amplitude is linearly

dependent on the interaction length, and we observe the quadratic relationship between the wave intensity and the crystal length – phase-matching condition considered in detail in the preceding section for the second harmonic generation case. But in the case under study the phase-matching condition is given in the following form:

$$k(2\omega_1) = 2k(\omega_1), \tag{2.4.7}$$

$$k(2\omega_2) = 2k(\omega_2), \tag{2.4.8}$$

$$k(\omega_1 + \omega_2) = k(\omega_1) + k(\omega_2), \tag{2.4.9}$$

$$k(\omega_1 - \omega_2) = k(\omega_1) - k(\omega_2). \tag{2.4.10}$$

The conditions of (2.4.7), (2.4.8) are associated with the second harmonic generation. For the conditions of (2.4.9), (2.4.10), the sum- or difference-frequency generation is realized, respectively. The phase-matching conditions for the sum- or difference-frequency generation are achieved in birefringent crystals, similar to the second harmonic generation. With the use of negative birefringent crystals ($n_o > n_e$), the medium dispersion is compensated for by selection of the crystal orientation so that a wave of the highest frequency be extraordinary. To illustrate, the wave at the sum frequency $\omega_1 + \omega_2$ in expression (2.4.9) and the wave at the frequency ω_1 in expression (2.4.10) should be extraordinary. When the remaining waves are ordinary, we have «*ooe*» - interaction.

An interesting effect is associated with the quadratic polarization components

$$P^{(2)} \sim E_1 E_1^* + E_2 E_2^*. \tag{2.4.11}$$

As seen from the expression, the induced polarization is independent of time. This points to the fact that a stationary electric field will be induced. This effect is called the optic rectification or inverse electro-optic (inverse Pockels) effect. The electric field strength is proportional to the quadratic polarization and hence to the light beam intensity. To measure the light induced voltage, the electrodes are deposited on the opposite crystal faces. In the approximation of a blanket exposure of the crystal, ΔU is proportional to the light intensity I and to the distance L between the electrodes

$$\Delta U = aIL, \tag{2.4.12}$$

where a - proportionality factor dependent on the parameters of an electro-optical crystal. For a crystal of potassium dihydrogen phosphate (KDP) $a \approx 10 \frac{\text{mV/cm}}{\text{MW/cm}^2}$, for lithium niobate ($LiNbO_3$) $a \approx 100 \frac{\text{mV/cm}}{\text{MW/cm}^2}$. In this way at the frequencies on

the order of MW/cm^2 we expect that the voltage is about 10 – 100 mV. Despite a small value of the induced voltage, the optic rectification effect, due to ultrafast response, is widely used to measure ultrashort laser pulses. For example, in the case of picosecond pulses at the typical intensities on the order of GW/cm^2 an amplitude of the initiated electric signal may come to several hundreds of volts.

2.4.2. Parametric amplification

The difference frequency generation is used for amplification of weak optical signals. If a weak signal at the frequency ω_1 and a high-power wave, so-called pump, at the frequency ω_P, is incident on a medium with the quadratic nonlinearity, then, when the phase-matching condition is met, the difference frequency is generated $\omega_2 = \omega_P - \omega_1$. On the other hand, the generated wave with the frequency ω_2, interacting with a pump wave, generates a wave at the frequency $\omega_1 = \omega_P - \omega_2$. Considering that the pump intensity is much higher than intensities of the waves at the frequencies ω_1 and ω_2, these processes may be treated as a decay of a pump photon into two photons with lower frequencies ($\omega_P = \omega_1 + \omega_2$). Another situation is also possible: only a pump wave is injected into a medium, whereas photons with the frequencies ω_1 and ω_2 are generated from noise.

To describe the parametric amplification, we use expression (2.4.1) for the quadratic nonlinearity $P^{(2)} = \chi^{(2)} E^2$, now with the following total field of the interacting waves:

$$E = \frac{1}{2}(E_1 \exp(-i\omega_1 t) + E_2 \exp(-i\omega_2 t) + E_P \exp(-i\omega_P t) + \\ + E_1^* \exp(i\omega_1 t) + E_2^* \exp(i\omega_2 t) + E_P^* \exp(i\omega_P t)) \quad . \qquad (2.4.13)$$

Substituting expression (2.4.13) into equation (2.4.1), we derive polarization at the frequencies ω_1, ω_2, and ω_P as follows:

$$P(\omega_1) = \frac{1}{2}\chi^{(2)}\left(E_P E_2^* \exp(i\omega_1 t) + E_P^* E_2 \exp(-i\omega_1 t)\right), \qquad (2.4.14)$$

$$P(\omega_2) = \frac{1}{2}\chi^{(2)}\left(E_P E_1^* \exp(i\omega_2 t) + E_P^* E_1 \exp(-i\omega_2 t)\right), \qquad (2.4.15)$$

$$P(\omega_P) = \frac{1}{2}\chi^{(2)}\left(E_1 E_2 \exp(i\omega_P t) + E_1^* E_2^* \exp(-i\omega_P t)\right). \qquad (2.4.16)$$

Then, in analogy with expressions (2.4.5), (2.4.6), we can write reduced wave equations of the form

$$\frac{\partial A_1}{\partial z}\exp(ik_1 z) = i\frac{2\pi\omega_1}{cn}\chi^{(2)}A_P A_2^*\exp(i(k_P - k_2)z), \qquad (2.4.17)$$

$$\frac{\partial A_2}{\partial z}\exp(ik_2 z) = i\frac{2\pi\omega_2}{cn}\chi^{(2)}A_P A_1^*\exp(i(k_P - k_1)z), \qquad (2.4.18)$$

$$\frac{\partial A_P}{\partial z}\exp(ik_P z) = i\frac{2\pi\omega_P}{cn}\chi^{(2)}A_1 A_2\exp(i(k_1 + k_2)z). \qquad (2.4.19)$$

Note that, with the reduced equations, the Manley-Rowe relation for the number of photons is valid

$$\frac{1}{\omega_1}\frac{\partial(A_1 A_1^*)}{\partial z} = \frac{1}{\omega_2}\frac{\partial(A_2 A_2^*)}{\partial z} = -\frac{1}{\omega_P}\frac{\partial(A_P A_P^*)}{\partial z}, \qquad (2.4.20)$$

this relation may be transformed as

$$\frac{1}{\omega_1}\frac{\partial I_1}{\partial z} = \frac{1}{\omega_2}\frac{\partial I_2}{\partial z} = -\frac{1}{\omega_P}\frac{\partial I_P}{\partial z} \qquad (2.4.21)$$

or

$$\frac{\partial N_1}{\partial z} = \frac{\partial N_2}{\partial z} = -\frac{\partial N_P}{\partial z}. \qquad (2.4.22)$$

where N_1, N_2, N_P - number of photons at the frequencies ω_1, ω_2, ω_P, respectively.

Relation (2.4.22) indicates that an increase in the number of photons at the frequencies ω_1 and ω_2 equals a decrease in the number of photons at the pump frequency ω_P, i.e. the pump photon decays into two photons with lower frequencies ($\omega_P = \omega_1 + \omega_2$).

For further consideration we use the high-power pumping approximation, when lowering of the pump intensity in the process of interaction may be neglected ($A_P = const$). Then the parametric amplification process is described by equations (2.4.17), (2.4.18) in the following form:

$$\begin{aligned}\frac{\partial A_1}{\partial z} &= i\sigma_1 A_2^*\exp(i\Delta k z),\\ \frac{\partial A_2^*}{\partial z} &= -i\sigma_2 A_1\exp(-i\Delta k z),\end{aligned} \qquad (2.4.23)$$

where $\sigma_{1,2} = 2\pi\omega_{1,2}\chi^{(2)}A_P \big/ cn$, $\Delta k = k_P - k_1 - k_2$ - phase mismatch of the waves.

A system of the two first-order differential equations is transformed to the second-order differential equation

$$\frac{\partial^2 A_1}{\partial z^2} - i\Delta k \frac{\partial A_1}{\partial z} - \sigma_1\sigma_2 A_1 = 0, \qquad (2.4.24)$$

its solution may be found as

$$A_1 = C_+ \exp(\gamma_1 z) + C_- \exp(\gamma_2 z), \qquad (2.4.25)$$

where $\gamma_{1,2}$ - solutions of the characteristic equation $\gamma^2 - i\Delta k\gamma - \sigma_1\sigma_2 = 0$

$$\gamma_{1,2} = \frac{1}{2}(i\Delta k \pm \sqrt{4\sigma_1\sigma_2 - (\Delta k)^2}). \qquad (2.4.26)$$

If the pump intensity is low

$$4\sigma_1\sigma_2 = 16\pi^2\omega_1\omega_2\chi^{(2)^2}A_P^2 / c^2n^2 < (\Delta k)^2, \qquad (2.4.27)$$

the factors $\gamma_{1,2}$ are pure imaginary and an amplitude of the wave at the frequency ω_1 fluctuates along the axis *z*. When $4\sigma_1\sigma_2 > (\Delta k)^2$, the wave amplification takes place with the amplification factor $\alpha_{ampl} = \sqrt{\sigma_1\sigma_2 - (\Delta k/2)^2}$.

Based on equation (2.4.27), we can write the threshold amplification condition as follows:

$$A_{P_{th}} = \frac{cn|\Delta k|}{2\pi\sqrt{\omega_1\omega_2}\chi^{(2)}}. \qquad (2.4.28)$$

A maximal amplification factor is attained at $\Delta k = 0$

$$k_{ampl} = 2\alpha_{yc} = \frac{4\pi\sqrt{\omega_1\omega_2}}{c^2 n}\chi^{(2)}A_P = \frac{8\pi^2\chi^{(2)}A_P}{n\sqrt{\lambda_1\lambda_2}}. \qquad (2.4.29)$$

Taking the typical parameters for the medium and radiation $\chi^{(2)} = 2.7\cdot10^{-8}$ CGS electrostatic units ($LiNbO_3$), $n = 2.23$, $\lambda_1 = \lambda_2 = 1$ μm, $A_P = 60$ CGS electrostatic units (associated with the intensity $I_P \cong 1$ MW/cm^2), we can obtain $k_{ampl} \cong 0.6$ cm^{-1}.

Considering such a character of equations (2.4.17), (2.4.18) and the fact that numbers of the generated photons at the frequencies ω_1 and ω_2 are identical, a similar reasoning is proper for a wave at the frequency ω_2 – its amplification factor is also determined by expression (2.4.29).

The calculated amplification factor $k_{ampl} \cong 0.6$ cm^{-1} is low to realize in practice the amplification of weak waves but it is comparable with amplification factors of many laser media. All these aspects encouraged the use of the parametric amplification effect for laser generation with frequency tuning.

2.4.3. Parametric generation

To realize the parametric generation, a cavity is used, like in an ordinary laser, to create a positive feedback. We differentiate between single- and two-cavity parametric oscillators. The difference between them is that a cavity feedback is realized for one, e.g., ω_1, or for two frequencies, ω_1 and ω_2. Schematically such a laser is shown in Fig. 2.4.1. A nonlinear crystal is positioned within the cavity comprising two mirrors; the mirrors are transparent at the pump frequency and they reflect waves at the frequency ω_1 or at the two frequencies ω_1 and ω_2. The generation is effected at the two frequencies, for which the phase-matching condition is fulfilled $\Delta\vec{k} = \vec{k}_P - \vec{k}_1 - \vec{k}_2 = 0$ and the amplification factor is at maximum (2.4.29). The phase-matching condition is given in the vector form as in the parametric oscillator schematic under study the waves are propagating not only in the concurrent but also in the counter-directions.

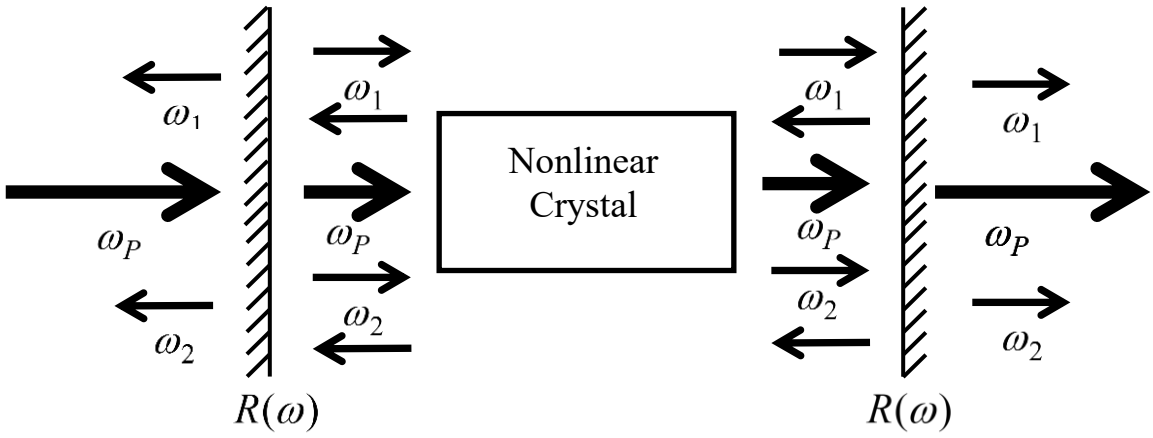


Fig. 2.4.1. Schematic of a parametric light oscillator

First, we consider a two-cavity parametric oscillator. It is assumed that for reflection factors of the mirrors at the frequencies ω_1 and ω_2 we have $R(\omega_1) = R(\omega_2) = R$ and at the pump frequency – $R(\omega_P) = 0$. Then, equating the loss and the amplification for the cavity round-trip $I_0 = I_0 \exp(k_{ampl} L) R^2$, we can get the threshold generation condition

$$\exp(k_{ampl} L) R^2 = 1. \qquad (2.4.30)$$

Here we take into account that amplification occurs only for the wave propagating at the frequencies ω_1 and ω_2 in the pump-wave direction – for the back-trip from the output mirror there is no amplification of the waves.

As follows from (2.4.30), the generation is possible if the inequality

$$k_{ampl} L > 2\ln(1/R) \qquad (2.4.31)$$

is fulfilled.

Selecting the above medium and radiation parameters with the amplification factor $k_{ampl} \cong 0{,}6$ см$^{-1}$, we can find from (2.4.31) that at the mirror reflection factor R=0.9 the generation is possible for the crystal length $L > 0.35$ см. These estimates have been obtained for the pump intensity $I_P \cong 1$ MW/cm^2. Selecting a large length of the crystal (e.g., $L = 3.5$ cm), one can lower the required amplification factor by an order. Considering that the amplification factor is proportional to the pump wave amplitude, the threshold pump intensity may be lowered by two orders of magnitude down to $I_P \cong 10$ kW/cm^2.

At the same time, a two-cavity parametric oscillator is frequency instable due to its low threshold. The generation is realized at the frequencies meeting the standing-wave formation condition, at both wave lengths

$$\begin{aligned} n_1(\omega_1)L &= m_1\lambda_1/2 \\ n_2(\omega_2)L &= m_2\lambda_2/2 \end{aligned}, \qquad (2.4.32)$$

where n_1 and n_2 - refractive indices at the frequencies ω_1 and ω_2, respectively; m_1, m_2 - integers, L - cavity base, for simplicity considered equal to the nonlinear crystal length.

It is convenient to use (2.4.32) for the frequencies ω_1 and ω_2

$$\begin{aligned} \omega_1 &= m_1 \pi c/n_1 L \\ \omega_2 &= m_2 \pi c/n_2 L \end{aligned}, \qquad (2.4.33)$$

but we should take into account that $\omega_1 + \omega_2 = \omega_P$. Using the above expressions, we easily can show that, when the cavity base is arbitrary varied as $\Delta L << L$, the generation frequency is varied as $\Delta\omega = \omega_P n_2 \Delta L/(n_2 - n_1)L$, and we have $\omega_1' = \omega_1 + \Delta\omega$, $\omega_2' = \omega_2 - \Delta\omega$. Considering that $n_2/(n_2 - n_1) \sim 100$, just this factor is responsible for higher frequency instability of a two-cavity oscillator as compared to the single-cavity laser for which $\Delta\omega/\omega = \Delta L/L$.

Due to this frequency instability, single-cavity parametric oscillators (with the mirrors which reflect waves, for example, at the frequency ω_1 and are transparent at the frequencies ω_P and ω_2) are most widely used. However, in the case of single-cavity oscillator the absence of a feedback for the second wave results in growing of the total loss and hence in the higher generation threshold. For the above-mentioned medium and cavity parameters, the generation threshold

of a single-cavity oscillator is increased by a factor of 20 coming to $I_P \cong$ 200 kW/cm^2.

The principal advantage of a parametric oscillator is the possibility for tuning of the generation frequency. The generation frequency is determined by the synchronism condition $\Delta\vec{k} = \vec{k}_P - \vec{k}_1 - \vec{k}_2 = 0$, with $\omega_1 + \omega_2 = \omega_P$.

In this way, for propagation of the waves in the same direction, we have the following condition:

$$\begin{aligned} &\omega_1 + \omega_2 = \omega_P \\ &\omega_1 n_1 + \omega_2 n_2 = \omega_P n_P \end{aligned} . \qquad (2.4.34)$$

This condition is automatically fulfilled for dispersion-free media. In real dispersion media it is required to use an ordinary and an extraordinary wave in uniaxial birefringent crystals: for example, in the case of «*ooe*» and «*oee*» interactions in negative crystals **(**$n_o > n_e$**)**, where a higher value of the refractive index for the pump wave is compensated by passing to the extraordinary wave.

We consider two main frequency tuning methods: angular tuning and temperature tuning.

The angular frequency tuning is associated with rotation of a nonlinear crystal within the cavity. To illustrate, consider «*ooe*» interactions in negative birefringent crystals. We assume that for some orientation of the crystal we have

$$\omega_P n_P^e(\omega_P, \Theta) = \omega_1 n_1^o(\omega_1) + \omega_2 n_2^o(\omega_2), \qquad (2.4.35)$$

where Θ - angle between the crystal axis and the cavity axis. When the crystal is rotated by the angle $\Delta\Theta$, the synchronism condition is fulfilled for the frequencies $\omega_1 + \Delta\omega$ and $\omega_2 - \Delta\omega$:

$$\omega_P n_P^e(\Theta + \Delta\Theta) = (\omega_1 + \Delta\omega) n_1^o(\omega_1 + \Delta\omega) + (\omega_2 - \Delta\omega) n_2^o(\omega_2 - \Delta\omega). \qquad (2.4.36)$$

This expression may be transformed for low values of the angular and frequency detuning ($\Delta\Theta \ll \Theta$ and $\Delta\omega \ll \omega$)

$$\begin{aligned} &\omega_P \left(n_P^e(\Theta) + \frac{\partial n_P^e}{\partial \Theta} \Delta\Theta \right) = (\omega_1 + \Delta\omega) \left(n_1^o(\omega_1) + \frac{\partial n_1^o}{\partial \omega} \Delta\omega \right) + \\ &+ (\omega_2 - \Delta\omega) \left(n_2^o(\omega_2) - \frac{\partial n_2^o}{\partial \omega} \Delta\omega \right) \end{aligned} . \qquad (2.4.37)$$

Simultaneous solution of equations (2.4.35) and (2.4.37) makes it possible to derive the relationship between $\Delta\omega$ and $\Delta\Theta$ as follows:

$$\Delta\omega \cong \frac{\omega_P \dfrac{\partial n_P^e}{\partial \Theta}\Delta\Theta}{n_1^o(\omega_1) - n_2^o(\omega_2) + \omega_1 \dfrac{\partial n_1^o(\omega_1)}{\partial \omega_1} - \omega_2 \dfrac{\partial n_2^o(\omega_2)}{\partial \omega_2}}. \qquad (2.4.38)$$

For uniaxial negative crystals we have

$$\frac{\partial n_P^e}{\partial \Theta} = -\frac{n_{Pe}\varepsilon_P^2 \sin 2\Theta}{2(1-\varepsilon_P^2 \cos^2\Theta)^{3/2}}, \qquad (2.4.39)$$

where $\varepsilon_P = \sqrt{1-(n_{Pe}/n_{Po})^2}$ - eccentricity of ellipse, n_{Po}, n_{Pe} - major and minor semiaxes. For the typical parameters of electro-optical crystals, e.g. for lithium niobate, the frequency detuning is ~1000 cm^{-1} (or 100 nm at $\lambda = 1\,\mu$m) when the crystal is rotated by an angle of 1 g (grade) close to the point of degeneracy ($\omega_1 \approx \omega_2$) and ~100 cm^{-1} – far from the point of degeneracy.

Now consider the temperature frequency detuning. It is associated with a change in the phase-matching conditions due to the temperature dependence of the refractive indices for the ordinary and extraordinary waves. Let at some temperature the following synchronism condition for the frequencies ω_1 and ω_2 be fulfilled:

$$\omega_P n_P^e(\omega_P, T) = \omega_1 n_1^o(\omega_1, T) + \omega_2 n_2^o(\omega_2, T). \qquad (2.4.40)$$

It is assumed that at the temperature $T + \Delta T$ the synchronism condition is realized for the frequencies $\omega_1 + \Delta\omega$ and $\omega_2 - \Delta\omega$

$$\begin{aligned}\omega_P\left(n_P^e + \frac{\partial n_P^e}{\partial T}\Delta T\right) &= (\omega_1 + \Delta\omega)\left(n_1^o + \frac{\partial n_1^o}{\partial T}\Delta T + \frac{\partial n_1^o}{\partial \omega}\Delta\omega\right) + \\ &+ (\omega_2 - \Delta\omega)\left(n_2^o + \frac{\partial n_2^o}{\partial T}\Delta T - \frac{\partial n_2^o}{\partial \omega}\Delta\omega\right)\end{aligned}. \qquad (2.4.41)$$

By simultaneous solution of equations (2.4.40) and (2.4.41) we obtain

$$\Delta\omega \cong \frac{\omega_P \dfrac{\partial n_P^e}{\partial T} - \omega_1 \dfrac{\partial n_1^o}{\partial T} - \omega_2 \dfrac{\partial n_2^o}{\partial T}}{n_1^o(\omega_1) - n_2^o(\omega_2) + \omega_1 \dfrac{\partial n_1^o(\omega_1)}{\partial \omega_1} - \omega_2 \dfrac{\partial n_2^o(\omega_2)}{\partial \omega_2}}\Delta T. \qquad (2.4.42)$$

For the typical parameters of electro-optic crystals, the frequency detuning comes to ~300 cm^{-1} when temperature varies by 1 K on the operation close to the point of degeneracy ($\omega_1 \approx \omega_2$) and is lower by an order of magnitude far from the point of degeneracy.

Using the angular and frequency detuning, we can realize the generation over a wide spectral range covering the visible and near IR regions. To illustrate, when pump is the third harmonic of an yttrium aluminum garnet laser (wave length 355 nm) the generation is in the range from 400 nm to 2400 nm. The parametric generation features a relatively wide spectral line that is on the order of 1 nm in the visible spectral region. For its estimation we use the same approach as for analysis of the angular synchronism width in the case of the second-harmonic generation. From the condition for the amplification band width $|\Delta k|L = 2\pi$ it follows that

$$\begin{aligned} &\frac{\omega_P}{c} n_P^e - \frac{(\omega_1 + \delta\omega)}{c}(n_1^o(\omega_1) + \frac{\partial n_1^o(\omega_1)}{\partial\omega}\delta\omega) - \\ &- \frac{(\omega_2 - \delta\omega)}{c}(n_2^o(\omega_2) - \frac{\partial n_2^o(\omega_2)}{\partial\omega}\delta\omega) = \frac{2\pi}{L} \end{aligned}, \qquad (2.4.43)$$

and $\omega_P n_P = \omega_1 n_1 + \omega_2 n_2$.

From this it follows that the angular synchronism width is given by

$$\delta\omega = \frac{2\pi c/L}{\left| n_1^o(\omega_1) - n_2^o(\omega_2) + \omega_1 \frac{\partial n_1^o(\omega_1)}{\partial\omega} - \omega_2 \frac{\partial n_2^o(\omega_2)}{\partial\omega} \right|}. \qquad (2.4.44)$$

As demonstrated by the numerical estimates, the angular synchronism width and hence the spectral generation-line width is on the order of 10 cm^{-1}. To attain a narrower spectral generation line, one can use a dispersion cavity, for example, a diffraction grating and, in special cases, a Fabry-Perot interferometer enabling the generation line widths ~ 0.1 – 0.01 cm^{-1}.

2.5. Parametric processes in media with cubic nonlinearity

2.5.1. Frequency summation and subtraction in media with cubic nonlinearity

Generally, the interaction of light waves in a medium with cubic nonlinearity is described by nonlinear polarization as follows:

$$P^{(3)} = \chi^{(3)} E^3, \qquad (2.5.1)$$

where $E = \frac{1}{2}(E_1 \exp(-i\omega_1 t) + E_2 \exp(-i\omega_2 t) + E_3 \exp(-i\omega_3 t) + E_1^* \exp(i\omega_1 t) + E_2^* \exp(i\omega_2 t) + E_3^* \exp(i\omega_3 t))$ - total field of the interacting waves,

$E_1 = A_1 \exp(ik_1 z)$, $E_2 = A_2 \exp(ik_2 z)$, $E_3 = A_3 \exp(ik_3 z)$ - spatial sections of the electric field strength of light waves at the corresponding frequencies.

It is easily seen that by solving expression (2.5.1) we can obtain 216 components associated with different frequency combinations for the interacting waves. Similar to the media with quadratic nonlinearity, the efficiency of realization of some or other process is determined by the fulfillment of the phase-matching condition. Let us consider several examples.

2.5.2. Third-harmonic generation

Third-harmonic generation may be observed in crystals, liquids, and gasses when using cubic nonlinearity. Such a situation is described by the polarization determined by (2.5.1) on propagation of the wave $E = \frac{1}{2}(E_1 \exp(-i\omega t) + E_1^* \exp(i\omega t))$. The polarization associated with the third harmonic generation may be represented as

$$P_3(3\omega) = \frac{1}{8}\chi^{(3)}\left(E_1^3 \exp(-i3\omega t) + E_1^{*3} \exp(i3\omega t)\right). \qquad (2.5.2)$$

Substituting (2.5.2) into the wave equation

$$\frac{\partial^2 E_3(3\omega)}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2 E_3(3\omega)}{\partial t^2} = \frac{4\pi}{c^2}\frac{\partial^2 P_3(3\omega)}{\partial t^2} \qquad (2.5.3)$$

and using the approximation of slowly varying amplitudes (the light field amplitude is insignificantly varying during the period of time on the order of the oscillation period $dA/dt << \omega A$ or at the distance on the order of the light-wave length: $dA/dz << kA$), in the stationary interaction mode we get the following reduced wave equation for plane waves:

$$\frac{\partial A_3}{\partial z} = i\frac{3\pi\omega}{2cn}\chi^{(3)} A_1^3 \exp(-i\Delta k z), \qquad (2.5.4)$$

where $\Delta k\, z = (k_3 - 3k_1)\, z$ – phase mismatch between the waves at the fundamental and at the triple frequency.

Equation (2.5.4) is solved in the approximation of a minor interaction efficiency when a decrease in the wave amplitude at the frequency ω can be neglected ($A_1 \cong \text{const}$)

$$I_3 = \frac{144\pi^4\omega^2}{c^4 n^4}\left|\chi^{(3)}\right|^2 I_1^3 L^2 \frac{\sin^2(\Delta kL/2)}{(\Delta kL/2)^2}\ . \qquad (2.5.5)$$

Note that the derived equation (2.5.5) is similar to equation (2.3.14) for the third harmonic generation (Section 2.3). An important feature is the cubic dependence of the third-harmonic intensity on the intensity at the fundamental frequency. The effective third-harmonic generation is possible on fulfillment of the phase-matching condition $k(3\omega) = 3k(\omega)$ that can be realized in birefringent crystals with the use of an ordinary and an extraordinary wave, e.g., for negative crystals **(**$n_o > n_e$**),** when the condition $n_e(3\omega) = n_o(\omega)$ is met. But it should be noted that the cubic susceptibility $\chi^{(3)} \sim 10^{-12}$ - 10^{-14} e.s.u. is considerably lower than the quadratic susceptibility $\chi^{(2)} \sim 10^{-8}$ - 10^{-9} e.s.u. Consequently, the factor of conversion to the third harmonic (~ 10^{-3} - 10^{-6}) is significantly lower than that in the case of the second harmonic ($\geq 0{,}1$).

The conversion factors were much higher with the use of resonant nonlinearity in metal vapors, for which $\chi^{(3)} \sim 10^{-9}$ e.s.u. In a mixture of rubidium and xenon vapors the conversion factor was about 50%. Of interest is the fulfilled phase-matching condition: $k(3\omega) = 3k(\omega)$. For the selected wavelength at the fundamental frequency 1064 nm, the third harmonic is associated with the wavelength 355 nm. The absorption peak of rubidium is located between these two wave lengths. Owing to the use of the region of abnormal dispersion, the refractive index of rubidium at the frequency ω is higher than that at the frequency 3ω ($n_{Rb}(\omega) > n_{Rb}(3\omega)$). For xenon, the dispersion $n_{Xe}(\omega) < n_{Xe}(3\omega)$ is normal. Selecting the appropriate concentration of rubidium vapors (by variations in the cell temperature), we can attain the fulfillment of the phase-matching condition: $n_{Rb+Xe}(\omega) = n_{Rb+Xe}(3\omega)$. However, this method is not widely used due to the difficulties in using the thermally stabilized cells filled with an inert gas and metal.

In practice, the third harmonic is generated with the use of birefringent crystals by means of cascade conversions at quadratic nonlinearity. First, the second harmonic $A(2\omega) \sim \chi^{(2)} A^2(\omega)$ is generated and then, due to the summation of the waves at the fundamental and double frequency, we obtain a wave with the triple frequency $A(3\omega) \sim \chi^{(2)} A(\omega) A(2\omega)$. Such a cascade process enables one to achieve the conversion efficiency at a level of 20 – 30 %.

2.5.3. Wave generation at the sum frequency on four-wave interaction

Let in a medium with cubic nonlinearity three waves be propagating $E_j = \frac{1}{2}(E_j \exp(-i\omega_j t) + E_j^* \exp(i\omega_j t))$, where $j = 1, 2, 3$. Then the nonlinear medium susceptibility is described by

$$P = \frac{1}{8}\chi^{(3)}(E_1 e^{-i\omega_1 t} + E_2 e^{-i\omega_2 t} + E_3 e^{-i\omega_3 t} + E_1^* e^{i\omega_1 t} + E_2^* e^{i\omega_2 t} + E_3^* e^{i\omega_3 t})^3, \quad (2.5.6)$$

from where there is a possibility for generation of the waves at different frequencies $\omega_4 = \omega_1 \pm \omega_2 \pm \omega_3$. Consider a variant of the wave generation at the sum frequency $\omega_4 = \omega_1 + \omega_2 + \omega_3$. In this case polarization of the medium takes the form

$$P_4 = \frac{3}{4}\chi^{(3)}(E_1 E_2 E_3 \exp(-i\omega_4 t) + E_1^* E_2^* E_3^* \exp(i\omega_4 t)). \quad (2.5.7)$$

Next, in the assumption of slowly varying amplitudes, a reduced wave equation is of the following form:

$$2ik_4 \frac{\partial A_4}{\partial z}\exp(i\vec{k}_4\vec{r}) = -\frac{4\pi\omega_4^2}{c^2}P_4, \quad (2.5.8)$$

where z – propagation direction of the wave A_4.

Substituting expression (2.5.7) into equation (2.5.8), we have

$$\frac{\partial A_4}{\partial z} = i\frac{3\pi\omega_4}{cn}\chi^{(3)}A_1A_2A_3\exp(-i\Delta kz), \quad (2.5.9)$$

where $\Delta\vec{k} = \vec{k}_4 - \vec{k}_1 - \vec{k}_2 - \vec{k}_3$.

An analytical solution of the derived equation is possible in the approximation of a weak energy exchange of the interacting waves when attenuation of the waves A_1, A_2, A_3 may be neglected ($A_{1,2,3} >> A_4$). In this case we have

$$A_4 = \frac{3\pi\omega_4}{cn}\chi^{(3)}A_1A_2A_3\frac{1-\exp(-i\Delta kL)}{\Delta k} \quad (2.5.10)$$

and

$$I_4 = \frac{cn}{8\pi}|A_4|^2 = \frac{576\cdot\pi^4\omega_4^2}{c^4n^4}\left|\chi^{(3)}\right|^2 I_1I_2I_3\frac{\sin^2(\Delta kL/2)}{(\Delta kL/2)^2}L^2. \quad (2.5.11)$$

As earlier, the effective generation of a wave at the sum frequency necessitates the fulfillment of the phase-matching condition ($\Delta k = 0$), for which the intensity has a quadratic dependence on the length of a nonlinear medium (

$I_4 \sim L^2$). As demonstrated previously for the third harmonic generation (Section 2.5.2), this condition is met with the use of abnormal dispersion in resonant media or in birefringent crystals. Both methods have been realized experimentally but they are not in common use. In the first case this is associated with the difficulties of using thermally stabilized cells which are filled with an inert gas and metal; in the second case – with a minor conversion efficiency (~ 10^{-3} - 10^{-6}).

2.5.4. Wave generation at the difference frequency on four-wave mixing

Among various frequency components of nonlinear polarization (2.5.6), we select the components at the frequency $\omega_4 = \omega_1 + \omega_2 - \omega_3$.

$$P_4 = \frac{3}{4}\chi^{(3)}(E_1 E_2 E_3^* \exp(-i\omega_4 t) + E_1^* E_2^* E_3 \exp(i\omega_4 t)) . \quad (2.5.12)$$

Then the reduced equation for a wave at the difference frequency is as follows:

$$\frac{\partial A_4}{\partial z} = i\frac{3\pi\omega_4}{cn}\chi^{(3)} A_1 A_2 A_3^* \exp(-i\Delta kz) , \quad (2.5.13)$$

where $\Delta\vec{k} = \vec{k}_4 - \vec{k}_1 - \vec{k}_2 + \vec{k}_3$.

To demonstrate different types of analytical solutions, we consider the interaction of waves in the approximation $A_{2,3} >> A_{1,4}$, assuming that $A_{2,3} \cong \text{const}$. Then, along with variations of the wave at the frequency $\omega_4 = \omega_1 + \omega_2 - \omega_3$, we should also consider the wave at the frequency $\omega_1 = \omega_4 + \omega_3 - \omega_2$. Similar to equations (2.5.13), we have

$$\frac{\partial A_1}{\partial z} = i\frac{3\pi\omega_1}{cn}\chi^{(3)} A_4 A_3 A_2^* \exp(i\Delta kz) . \quad (2.5.14)$$

The effective interaction takes place when the phase-matching condition ($\Delta k = 0$) is met. Apart from the possibility to use media with abnormal dispersion or birefringent crystals, on generation of the difference frequency a new geometrical method enables one to realize the phase-matching condition. As seen in Fig. 2.5.1, the synchronism condition may be easily achieved due to variations in the propagation direction of the interacting waves. Changing the angle between the wave vectors, one can attain equality of the vector sums and offer the fulfillment of the condition $\vec{k}_1 + \vec{k}_2 = \vec{k}_3 + \vec{k}_4$.

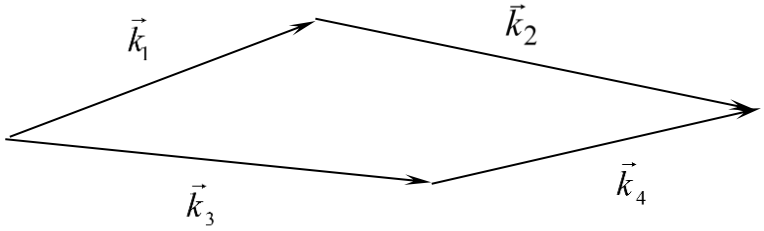


Fig. 2.5.1. Positions of the wave vectors for realization of the phase-matching condition

In this case equations (2.5.13), (2.5.14) may be of the following forms:

$$\begin{cases} \partial A_1/\partial z = i\sigma_1 A_4 \\ \partial A_4/\partial z = i\sigma_4 A_1 \end{cases}, \tag{2.5.15}$$

where $\sigma_1 = \dfrac{3\pi\omega_1}{cn}\chi^{(3)}A_3A_2^*$, $\sigma_4 = \dfrac{3\pi\omega_4}{cn}\chi^{(3)}A_2A_3^*$.

A solution for a system of equations (2.5.15), with the boundary condition $A_1(z=0) = A_{10}$, is obtained as

$$A_1 = A_{10}\cos\gamma z, \tag{2.5.16}$$

$$A_4 = iA_{10}\sqrt{\sigma_4/\sigma_1}\sin\gamma z, \tag{2.5.17}$$

where $\gamma = \sqrt{\sigma_1\sigma_4}$.

The spatial intensity distribution of the waves I_1 and I_4 is described by squares of the trigonometric functions

$$I_1 = I_{10}\cos^2\gamma z, \tag{2.5.18}$$

$$I_4 = I_{10}\sqrt{\omega_4/\omega_1}\sin^2\gamma z. \tag{2.5.19}$$

Relations (2.5.18), (2.5.19) are illustrated in Fig. 2.5.2.

As seen, in the presence of the high-intensity waves A_2 and A_3 in a medium with cubic nonlinearity the energy exchange between the waves A_1 and A_4 takes place periodically. The spatial period of such exchange is determined by the parameter $\gamma = \sqrt{\sigma_1\sigma_4}$. For example, when using waves with the amplitudes $A_2 \sim A_3 \sim 100$ e.s.u. and the cubic susceptibility $\chi^{(3)} \sim 10^{-12}$ e.s.u. (birefringent crystal), this parameter is $\gamma \sim 10^{-2}$ cm^{-1}, i.e., the characteristic period of oscillations is on the order of 1 m. The oscillation period may be reduced significantly on going to resonant media associated with $\chi^{(3)} \sim 10^{-9}$e.s.u. In this case the oscillation period is on the order of several millimeters.

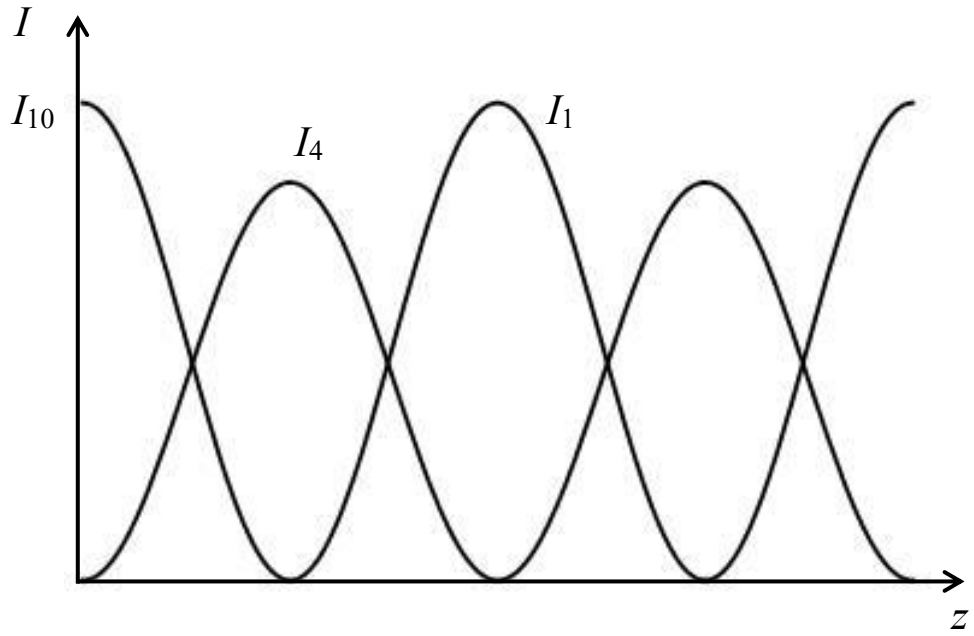


Fig. 2.5.2. Spatial intensity distributions of light beams on generation of the difference frequency in a medium with cubic nonlinearity

2.5.5. Parametric amplification on four-wave counter-interaction

To demonstrate the potentialities of parametric amplification on four-wave counter-interaction, let us select the components of nonlinear polarization (2.5.6) associated with the generation of a wave at the frequency $\omega_4 = \omega_3 + \omega_2 - \omega_1$

$$P_4 = \frac{3}{4}\chi^{(3)}(E_1^* E_2 E_3 \exp(-i\omega_4 t) + E_1 E_2^* E_3^* \exp(i\omega_4 t)). \qquad (2.5.20)$$

Similar to Section 2.5.4, we use the approximation $A_{2,3} >> A_{1,4}$, i.e., we assume that the amplitudes of the high-intensity waves are actually invariable in the process of interaction ($A_{2,3} \cong \text{const}$). Then, along with variations of the wave at the frequency $\omega_4 = \omega_2 + \omega_3 - \omega_1$, we should consider a wave at the frequency $\omega_1 = \omega_2 + \omega_3 - \omega_4$. In this case we take into account that a wave at the frequency ω_4 is propagating in the counter direction to the wave at the frequency ω_1.Then the reduced wave equation for the waves at the frequencies ω_1 and ω_4 is written as

$$\frac{\partial A_1}{\partial z} = i\frac{3\pi\omega_1}{cn}\chi^{(3)} A_2 A_3 A_4^* \exp(-i\Delta kz). \qquad (2.5.21)$$

$$\frac{\partial A_4}{\partial z} = -i\frac{3\pi\omega_4}{cn}\chi^{(3)} A_1^* A_2 A_3 \exp(-i\Delta kz), \qquad (2.5.22)$$

where $\Delta\vec{k} = \vec{k}_4 + \vec{k}_1 - \vec{k}_2 - \vec{k}_3$. Note that the condition of counter propagation allows for the minus sign in equation (2.5.22).

The effective interaction takes place when the phase-matching condition ($\Delta k = 0$) is met, this condition is easily fulfilled due to variations in the propagation direction of the interacting waves. As seen in Fig. 2.5.3, the possibility of changing the interaction geometry for the interacting waves provides fulfillment of the condition $\vec{k}_1 + \vec{k}_4 = \vec{k}_2 + \vec{k}_3$.

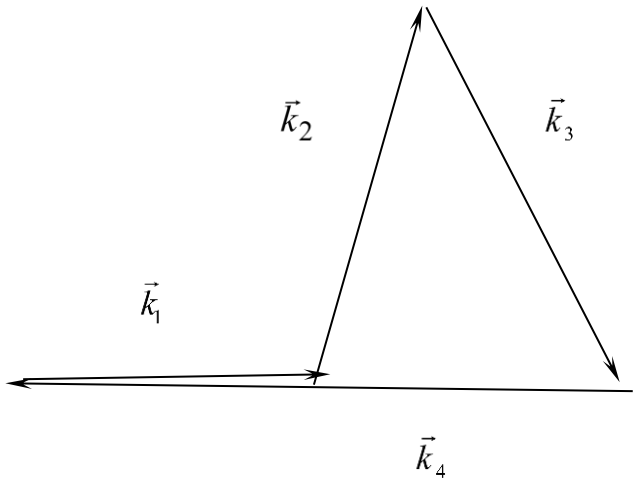


Fig. 2.5.3. Position of the wave vectors for realization of the phase-matching condition

In this case equations (2.5.21), (2.5.22) are given as follows:

$$\begin{cases} \partial A_1/\partial z = i\sigma_1 A_4^* \\ \partial A_4/\partial z = -i\sigma_4 A_1^* \end{cases}, \tag{2.5.23}$$

where $\sigma_1 = \dfrac{3\pi\omega_1}{cn}\chi^{(3)}A_2A_3$, $\sigma_4 = \dfrac{3\pi\omega_4}{cn}\chi^{(3)}A_2A_3$.

A system of the two first-order differential equations is transformed to the second-order differential equation

$$\frac{\partial^2 A_1}{\partial z^2} + \sigma_1\sigma_4 A_1 = 0, \tag{2.5.24}$$

with the solution taking the form

$$A_1 = C_+ \exp(\gamma_1 z) + C_- \exp(\gamma_2 z), \tag{2.5.25}$$

where $\gamma_{1,2}$ - solutions for the characteristics equation $\gamma^2 + \sigma_1\sigma_4 = 0$

$$\gamma_{1,2} = \pm i\sqrt{\sigma_1\sigma_4}\,. \tag{2.5.26}$$

As follows from equation (2.5.26), coefficients of the characteristic equation are pure imaginary, as distinct from the previously considered case of parametric

amplification in a medium with quadratic nonlinearity (Section 2.4). This prevents from the introduction of the amplification factor in the form of a real part of the factor in the characteristic equation. Besides, particular specificity is associated with imposition of the boundary conditions at the opposite sides of the nonlinear medium

$$\begin{aligned} A_1(z=0) &= A_{10} \\ A_4(z=L) &= A_{4L} \end{aligned} . \qquad (2.5.27)$$

Solving a system of equations (2.5.23), we can find the light filed amplitudes after the propagation in the non linear medium

$$\begin{aligned} A_1(z=L) &= A_{10}\cos^{-1}\gamma L + i\sqrt{\omega_1/\omega_4}\,A_{4L}^{*}tg\gamma L \\ A_4^{*}(z=0) &= A_{4L}^{*}\cos^{-1}\gamma L - i\sqrt{\omega_4/\omega_1}\,A_{10}tg\gamma L \end{aligned} , \qquad (2.5.28)$$

where $\gamma = \sqrt{\sigma_1\sigma_4}$.

In equations (2.5.28) one should pay attention to $\cos\gamma L$ in the denominator. The value of $\cos\gamma L$ tends to zero when $\gamma L \to \pi/2$, enabling realization of light wave amplification by means of the energy transfer from two high-intensity pump waves. On fulfillment of the condition $\gamma L = \pi/2$, we expect switching to the generation mode when noise photons at the entrance to the nonlinear medium support generation of the counter-propagating waves. Their directions and frequencies are determined by the phase-matching condition $\vec{k}_1 + \vec{k}_4 = \vec{k}_2 + \vec{k}_3$, as shown in Fig. 2.5.3.

To calculate the length of a nonlinear medium, for which the parametric generation is possible, we transform the equation for $\gamma = \sqrt{\sigma_1\sigma_4}$ with due regard for the expressions for σ_1 and σ_4 (2.5.23)

$$\gamma = \frac{6\pi^2\chi^{(3)}A_2A_3}{n\sqrt{\lambda_1\lambda_4}} . \qquad (2.5.29)$$

Taking the typical parameters of interaction (pump wave amplitudes $A_2 \sim A_3 \sim 100$ e.s.u., wave length ~ 1 μm, cubic susceptibility $\chi^{(3)} \sim 10^{-9}$ e.s.u. (resonant medium), we can get $\gamma \sim 3 - 4$ cm^{-1}. This means that, when a nonlinear medium is only a few millimeters long, the waves are amplified markedly, and the generation conditions are realized when the length is about 0.5 cm.

2.5.6. Phase conjugation on four-wave interaction

The above scheme of parametric amplification of the counter-propagating waves is extensively used for realization of the phase conjugation effect discovered by the Belarusian physicists A.S.Rubanov, E.V.Ivakin, B.I.Stepanov in 1970. The effect is associated with the formation of a wave that has the so-called conjugate wave front. The conjugate-wave amplitude and phase distribution is similar to that of the initial wave but the conjugate wave is propagating in the counter direction. Such a situation is observed on the frequency-degenerate four-wave interaction when all the involved waves are of the same frequency ($\omega_1 = \omega_2 = \omega_3 = \omega_4 = \omega$). Indeed, on the assumption that the two high-intensity counter-propagating pump waves A_2, A_3 and the weak wave A_1 (at a certain angle) are incident on a nonlinear medium, according to the phase-matching condition, we get the formation of the wave A_4 with the wave vector $\vec{k}_4 = \vec{k}_2 + \vec{k}_3 - \vec{k}_1$ (Fig. 2.5.4). Considering that on counter-propagation of the pump waves $\vec{k}_2 + \vec{k}_3$=0, the wave vectors for the waves A_1 and A_4 also conform to the condition $\vec{k}_4 = -\vec{k}_1$. This means that the phase-matching condition in the case of degenerate four-wave interaction and counter directions of pump waves is fulfilled automatically. For a wave with the arbitrary wave vector $\vec{k}_1$, the induced wave is propagating in the counter direction $\vec{k}_4 = -\vec{k}_1$. As follows from (2.5.28), in this case the wave amplitude A_4 at the exit from a nonlinear medium is a complex conjugate of the wave A_1 ($A_4(z=0) \sim A_{10}^*$). Conjugation of the light-wave amplitudes in turn points to the fact that their phases have the opposite signs, i.e., we have $\varphi_4(z=0) = -\varphi_{10}$.

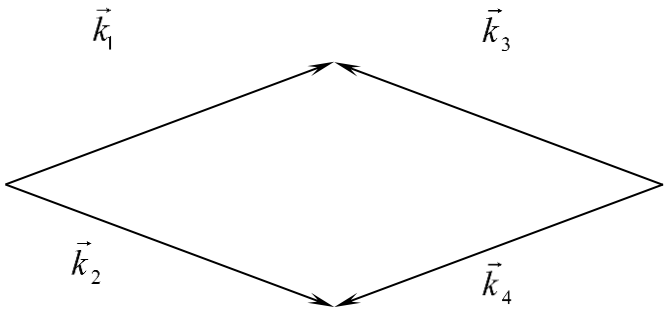


Fig. 2.5.4. Diagram of the wave vectors on the degenerate four-wave interaction

In this way we obtain the following relations for A_4:

$$\omega_4 = \omega_{10}$$
$$\vec{k}_4(z=0) = -\vec{k}_{10}\ . \qquad (2.5.30)$$
$$\varphi_4(z=0) = -\varphi_{10}$$

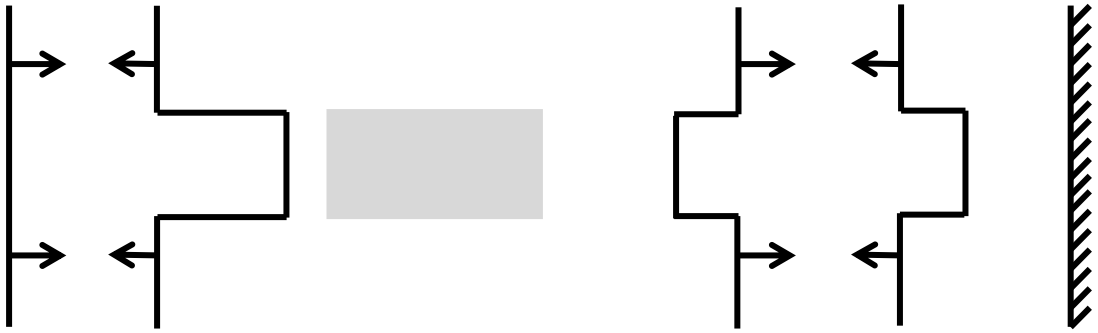
Next we use the representation of light fields in the complex form

$$E_{10} = \frac{1}{2}\left(a_{10}\exp(i(k_{10}z - \omega t + \varphi_{10})) + a_{10}\exp(-i(k_{10}z - \omega t + \varphi_{10}))\right). \qquad (2.5.31)$$

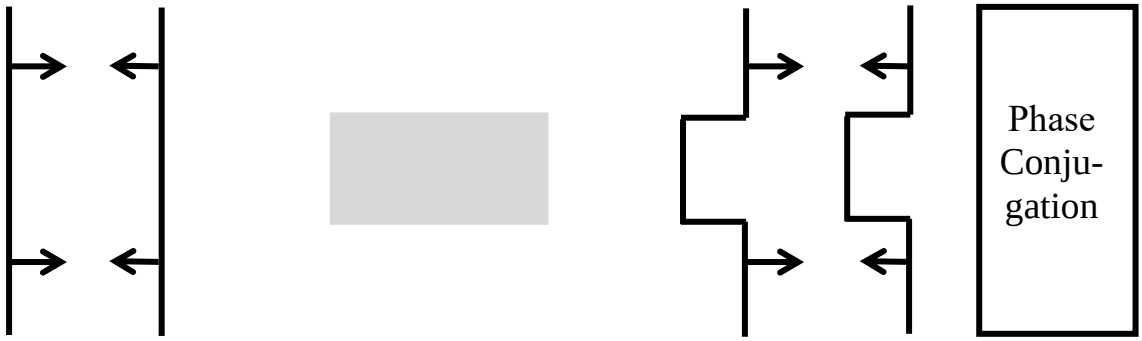
$$E_4 = \frac{1}{2}\left(a_4\exp(i(k_4 z - \omega t + \varphi_4)) + a_4\exp(-i(k_4 z - \omega t + \varphi_4))\right). \qquad (2.5.32)$$

Considering relations (2.5.30), comparison of equations (2.5.31) and (2.5.32) indicates that the wave fronts are coincident with simultaneous substitution of $-t$ for t. In fact, such a substitution leads to the same exponents $\exp(i(k_{10}z - \omega t + \varphi_{10})) = \exp(-i(k_4 z - \omega(-t) + \varphi_4))$ and $\exp(-i(k_{10}z - \omega t + \varphi_{10})) = \exp(i(k_4 z - \omega(-t) + \varphi_4))$. This means that the wave E_4 reproduces a state of the wave E_1 in earlier instants of time.

The essence of this phenomenon is illustrated in Fig. 2.5.5 by comparison of an ordinary mirror reflection with the reflection by a nonlinear medium on four-



a



b

Fig. 2.5.5. Compensation of phase distortions on phase conjugation

wave interaction. When a plane wave of the phase object (e.g., glass cylinder) is transmitted through a medium, the wave front is distorted, forming a step. In the case of an ordinary mirror reflection (Fig. 2.5.5, *a*) the wave front is inverted, the repeated transmission of the phase object leads to the doubled wave-front distortion. Reflection from a nonlinear medium, when the phase-matching condition effect is realized, makes it possible to have the same distribution of the wave front as for the incident wave. As a result, the backward passing of the phase object allows for compensation of the phase distortions and recovery of the initial plane wave (Fig. 2.5.5, *b*). So, we can attain a state of the light field as it was before the propagation of the phase object.

This property of a conjugate light wave has enabled the development of different ways to compensate for the phase distortions when radiation is propagating in optically inhomogeneous media and to solve the problem of laser radiation focusing at the targets of small area. High-power laser systems with a phase-conjugate mirror have been designed to compensate for optical distortions in active elements and to generate radiation characterized by the diffraction-limited divergence and high spectral brightness.

2.6. Stimulated Raman scattering. Stimulated Brillouin scattering.

2.6.1. Stimulated Raman scattering

In 1928 the Indian physicists C.V. Raman and K.S. Krishnan, and independently the Soviet physicists L.I. Mandelshtam and G.S. Landsberg, discovered that a particular small portion of radiation ($\sim 10^{-8}$) incident on the samples under study is subjected to a significant frequency shift in the process of scattering. This phenomenon is termed the Raman scattering of light. The researchers demonstrated that the frequency shifts are equal to the frequencies of natural molecular vibrations and came to the conclusion that polarization of a medium is modulated by atomic vibrations. As a result, the light field is reemitted at the frequencies combining those of the incident radiation and the characteristic frequencies of atomic vibrations. In the general case the process of radiation formation at the Raman frequencies is described by the following relation:

$$P = \alpha E = \alpha E_0 \cos \omega_0 t\,, \qquad (2.6.1)$$

where P – electric dipole moment of a molecule arising under the effect of exciting radiation, α - molecular polarizability tensor, E - electric field strength of exciting radiation, E_0 - amplitude of the electric field strength, ω_0 - frequency of exciting radiation (pump wave).

In the case of an absolutely rigid molecule α is constant in value. In a vibrating molecule α is a function of the normalized coordinate of a molecular shift from the equilibrium position X. Considering molecular vibrations of a material, the polarizability tensor is represented as a sum of the two first expansion terms with respect to the normal coordinate

$$\alpha = \alpha_0 + \left(\frac{\partial \alpha}{\partial X}\right)_0 X. \qquad (2.6.2)$$

Here α_0 - molecular polarizability in the equilibrium state, $\left(\frac{\partial \alpha}{\partial X}\right)_0$ is the so-called differential polarizability.

For a molecule vibrating at its normal vibrational frequency Ω, expression (2.6.1) takes the following form:

$$P = \left(\alpha_0 + \left(\frac{\partial \alpha}{\partial X}\right)_0 X_0 \cos\Omega t\right) E_0 \cos\omega_0 t, \qquad (2.6.3)$$

where X_0 - amplitude of the molecular shift from the equilibrium position.

As follows from (2.6.3), at the output of a nonlinear medium (apart from the initial radiation at the frequency ω_0) two additional luminous fluxes are emerging with the frequencies $\omega_0 - \Omega$ and $\omega_0 + \Omega$. Radiation with the frequency $\omega_0 - \Omega$ is called the Stokes component of Raman light scattering and radiation with the frequency $\omega_0 + \Omega$ - anti-Stokes component. Note that the Stokes $(\omega_0 - \Omega)$ and anti-Stokes $(\omega_0 + \Omega)$ components in turn can serve as the initial radiation generating the frequencies $(\omega_0 - 2\Omega)$ and, and their scattering results in the frequencies $(\omega_0 - 3\Omega)$ and $(\omega_0 + 3\Omega)$, etc.

Fig. 2.6.1, *a* illustrates the origination of the first Stokes component.

Light quanta with the energy $\hbar\omega_0$ are incident on the molecules in the ground vibrational state, and, being subjected to scattering, the molecules go to the excited vibrational state, and photons of scattered light have the energy $\hbar\omega_{1s} = \hbar\omega_0 - \hbar\Omega$.

Fig. 2.6.1, *b*. shows the first anti-Stokes component origination. Light quanta with the energy $\hbar\omega_0$ are incident on the molecules in the excited vibrational state

and, being subjected to scattering, the molecules go to the ground state, whereas photons of scattered light have the energy $\hbar\omega_{1as} = \hbar\omega_0 + \hbar\Omega$.

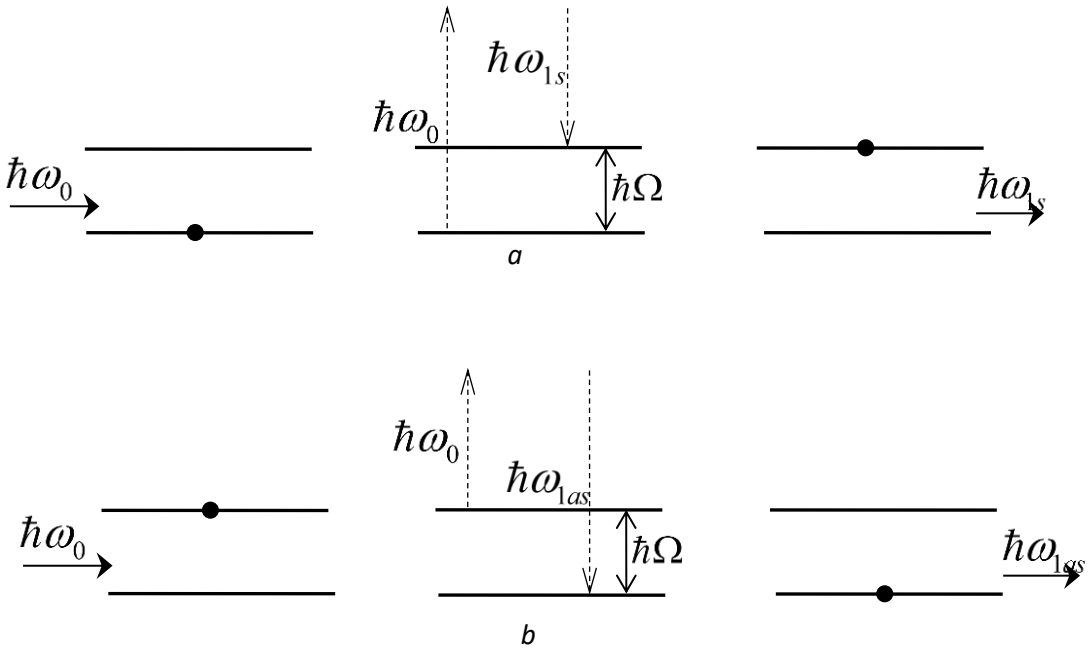


Fig 2.6.1. Origination of the stokes and anti-stokes radiation components on raman scattering

The radiation intensity of the Raman components I is given by the expression

$$I = I_0 \left(\frac{\partial \sigma}{\partial \theta} \right)_0 N d\theta dz , \qquad (2.6.4)$$

where I_0 – intensity of exciting radiation (pump wave), $\left(\frac{\partial \sigma}{\partial \theta} \right)_0$ - differential scattering cross-section per unit volume of a material, N – number of the molecules in a unit volume in the ground (for the Stokes component) and excited vibrational (for the anti-Stokes component) states, $d\theta$ - solid scattering angle, dz - scattering volume length. As in the normal conditions the number of particles N in the ground state is considerably greater than that in the excited state, the power of a Stokes component is significantly higher than that of the anti-Stokes component.

In the case of spontaneous Raman scattering radiation is emitted isotropically in all directions, every molecule acts independently of all others – there is no coherence between radiations form different emitters. The scattering efficiency is low: $10^{-6} \div 10^{-8}$ scattered photons per one incident photon. The line width of the output radiation $\Delta\nu_{RS}$ corresponds to the width of the vibration level for a medium.

An interest to Raman scattering has been greatly increased since the advent of lasers. One of the reasons is that the effect became easily observable and hence more convenient as a method of material studies. Another reason is that in the case of stimulated processes the shifted spectral lines may be very similar in their characteristics to the lines of laser radiation and may be used as new sources of coherent optical radiation. This effect is termed the stimulated Raman scattering (SRS).

The effect of SRS was discovered in 1962 by E. Woodbury and A. Ng in the process of the experiments with a ruby laser, where high-power single pulses were generated by the Q-switching method with the use of a nitrobenzene-based Kerr cell. As SRS offers the efficient conversion of the energy of laser radiation to the energy of the first Stokes component η_{1s} at a level of several tens of percent (theoretical limit $\eta_{1s} = \frac{\hbar\omega_{1s}}{\hbar\omega} = \frac{\hbar\omega_0 - \hbar\Omega}{\hbar\omega_0} = 1 - \frac{\hbar\Omega}{\hbar\omega_0}$), it is a very effective means for the creation of the visible and IR coherent optical-radiation sources. Emission of photons by the molecules in the process of Raman scattering is both phase and direction matched. As distinct from the spontaneous effect, in the case of Raman scattering a positive feedback is realized for the Stokes and anti-Stokes components. A high intensity of radiation generated in the case of Raman scattering is attained due to a mechanism of the molecular vibrations amplification on the interaction of light fields at the frequencies ω_0 and $\omega_{1s} = \omega_0 - \Omega$, the difference of which is equal to the frequency of molecular vibrations. In this case the chaotic molecular vibrations (fluctuation in their character) are overlapped by the forced regular vibrations. Forced vibrations are phased for all the molecules by external fields. Because of this, the contributions from individual molecules are added together, according to the amplitude rather than the intensity as in the spontaneous case, and hence the energy exchange between the waves on stimulated scattering is much higher than with spontaneous scattering. In this way Raman scattering is the process initiated by the optical excitation of intramolecular vibrations; resonance of the excitation is caused by Stokes scattering of a high-power laser wave. Fig. 2.6.2 shows that the process of stimulated Raman scattering results in amplification of the radiation associated with the first Stokes component and molecules of the material are going to the excited vibrational state. An increase in the number of photons for the Stokes component leads to the increased population of vibrational levels. As this takes place, the conditions are created for the generation of different Stokes and anti-Stokes components. As amplification is greatly dependent on frequency,

stimulated scattering occurs mainly at the frequency associated with a maximum amplification. As a result, a width of the spectral line for the output radiation of stimulated Raman scattering $\Delta\nu_{SRS}$ is significantly smaller that for spontaneous Raman scattering.

The process of stimulated Raman scattering may be described qualitatively by a macroscopic theory but it gives no complete pattern of the developed scenario in this case. From the viewpoint of a quantum theory, Raman scattering to a Stokes signal – two-photon process when one photon ($\hbar\omega_0$) is absorbed and one photon ($\hbar\omega_{1s}$) is emitted, the medium going from the initial to the excited vibrational state with the energy $\hbar\Omega = \hbar\omega_0 - \hbar\omega_{1s}$. Note that then the emitted photon $\hbar\omega_{1s}$ may play a role of pump, initiating emission of a photon of the second Stokes component with the frequency $\omega_{2s} = \omega_{1s} - \Omega$ and this photon in turn generates a photon of the third Stokes component at the frequency $\omega_{3s} = \omega_{2s} - \Omega$, and so on. This means that a sequential cascaded generation of the Stokes components takes place at high-power pumping.

According to the quantum and classical theories, there is no origination of high-intensity anti-Stokes components within the scope of two-photon processes. But there is a possibility for the concurrent four-photon process when two photons of pump radiation break into photons of the Stokes and anti-Stokes emission components, and we have

$$2\omega_0 = \omega_s + \omega_{as}. \tag{2.6.5}$$

It is known that such processes are well described by a model of the four-wave interaction (FWI) realized in a medium with the third-order nonlinear susceptibility $\chi^{(3)}(\omega)$. Using the expansion formalism for the nonlinear medium polarization related to $\chi^{(3)}(\omega)$ in terms of nonlinear susceptibilities for each of the frequency components of the interacting waves, we can write expressions for the polarization components associated with the formation of Stokes and anti-Stokes components of scattered radiation as follows:

$$P_s = \chi^{(3)}(\omega_s) E_0{}^2 E_{as}{}^*, \tag{2.6.6}$$

$$P_{as} = \chi^{(3)}(\omega_s) E_0{}^2 E_s{}^*. \tag{2.6.7}$$

For simplicity of further calculations, we use the approximation of the slowly varying amplitudes and obtain the following system of reduced wave equations:

$$\frac{\partial A_s}{\partial z} = i\frac{3\pi\omega_s}{2cn}\chi^{(3)} A_0{}^2 A_{as}{}^* \exp(i\Delta kz), \tag{2.6.8}$$

$$\frac{\partial A_{as}}{\partial z} = i\frac{3\pi\omega_{as}}{2cn}\chi^{(3)}A_0{}^2 A_s{}^*\exp\left(\mathrm{i}\Delta kz\right), \qquad (2.6.9)$$

where $\Delta\vec{k} = 2\vec{k}_0 - \vec{k}_s - \vec{k}_{as}$ - phase mismatch of the wave fronts.

This system of equations is solved in the approximation of the little depleted pump wave when the wave amplitude A_0 may be considered constant. First, the system of equations (2.6.8), (2.6.9) takes the form

$$\frac{\partial A_s}{\partial z} = i\sigma_s A_{as}{}^*\exp\left(i\Delta kz\right), \qquad (2.6.10)$$

$$\frac{\partial A_{as}}{\partial z} = i\sigma_{as} A_a{}^*\exp\left(i\Delta kz\right), \qquad (2.6.11)$$

where $\sigma_{s,as} = \frac{3\pi\omega_{s,as}}{2cn}\chi^{(3)}A_0{}^2$. Then it is transformed to the second-order differential equation

$$\frac{\partial^2 A_s}{\partial z^2} - i\Delta k\frac{\partial A_s}{\partial z} - \sigma_s\sigma_{as} = 0. \qquad (2.6.12)$$

We seek for a solution of this equation as follows:

$$A_s = A_{s1}\exp\left(\alpha_1 z\right) + A_{s2}\exp\left(\alpha_2 z\right), \qquad (2.6.13)$$

where $\alpha_{1,2} = \frac{1}{2}\left(\left(i\Delta k\right) \pm \sqrt{4\sigma_s\sigma_{as} - \left(\Delta k\right)^2}\right)$ - roots of the characteristic equation.

The sign «-» is associated with damped oscillations of a Stokes wave and the sign «+» is related to a wave with a growing amplitude. The amplitude amplification factor of the wave with the growing amplitude is determined by

$$\alpha_{ampl} = \sqrt{\sigma_s\sigma_{as} - \left(\Delta k/2\right)^2}. \qquad (2.6.14)$$

Amplification is at maximum in the propagation direction of the waves meeting the synchronism condition $\Delta\vec{k} = 0$ that is fulfilled when the Stokes and anti-Stokes components are propagating at an angle to the pump wave. As this takes place, the amplification factor is equal to

$$\alpha_{ampl} = \sqrt{\sigma_s\sigma_{as}} = \frac{12\pi^2\sqrt{\omega_s\omega_{as}}}{c^2n^2}\chi^{(3)}I_0. \qquad (2.6.15)$$

The Stokes wave intensity in this case is given as

$$I_s = I_{s0}\exp\left(2\alpha_{ampl}z\right), \qquad (2.6.16)$$

where I_{s0} - intensity of a Stokes wave for $z = 0$.

Stimulated Raman scattering is a threshold phenomenon, i.e. its efficiency is greatly dependent on the power of pump radiation. For its particular value

(threshold power) an intensity of Stokes radiation is sharply increased. A threshold value of the power of pump radiation is easily estimated when we represent a nonlinear medium in the form of a cavity with low mirror-reflection factors $R \sim 10^{-2} - 10^{-6}$. Such reflection is almost always realized by means of radiation scattering within the volume or at the edges of a nonlinear medium. The threshold generation condition is governed by the requirement that Raman amplification should compensate the cavity loss

$$R^2 \exp\left(2\alpha_{ampl} L\right) = 1, \tag{2.6.17}$$

where L - medium length.

The above expression (2.6.17) takes into account that, during the round trip of the cavity, radiation is twice reflected from the mirrors but it is amplified only on concurrent propagation of all the waves involved, in accordance with the phase-matching condition $\Delta\vec{k} = 2\vec{k}_0 - \vec{k}_s - \vec{k}_{as} = 0$.

To find a threshold power of pump radiation, we introduce the normalized amplification factor

$$g(\omega) = 2\alpha_{ampl}(\omega) / I_0. \tag{2.6.18}$$

Then from (2.6.17) we can derive the expressions for the threshold intensity

$$I_0^{th} = -2\ln R / gL \tag{2.6.19}$$

and for the threshold power of a light beam

$$P_0^{th} = -2S\ln R / gL, \tag{2.6.20}$$

where S – cross-sectional area of the pump beam.

As follows from (2.6.20), the threshold pump power is proportional to the beam cross-section and inversely proportional to the interaction length L. For the typical experimental parameters (e.g., classical experiment with benzene in a glass cell 30 cm in length) the threshold intensity is about 100 MW/cm^2.

To illustrate the effective realization of the stimulated Raman scattering process, let us consider the use of an optical fiber offering both large interaction lengths (dozens and hundreds of meters) and sufficiently high power densities of exciting radiation due to smallness of the fiber core diameter (several tens of micron). Stimulated Raman scattering in an optical fiber has certain peculiarities. As distinct from the majority of Raman-active media, where radiation scattering takes place at the specific characteristic frequencies of molecular or lattice vibrations, scattering in optical fibers is possible at many frequencies as a fiber is based on fused rather than crystalline quartz. The radiation amplification factor $\alpha_{ampl}(\omega)$ for fused quartz is appreciably distinct from zero at numerous frequencies. This is due to the fact that in amorphous materials, such as fused

quartz, the frequency bands of molecular vibrations are overlapping to form the continuum. Fig. 2.6.3 shows the profile of the normalized amplification factor $g(\omega) = 2\alpha_{ampl}(\omega) / I_0$ for fused quartz.

As seen in Fig. 2.6.3, the amplification factor is at maximum for the

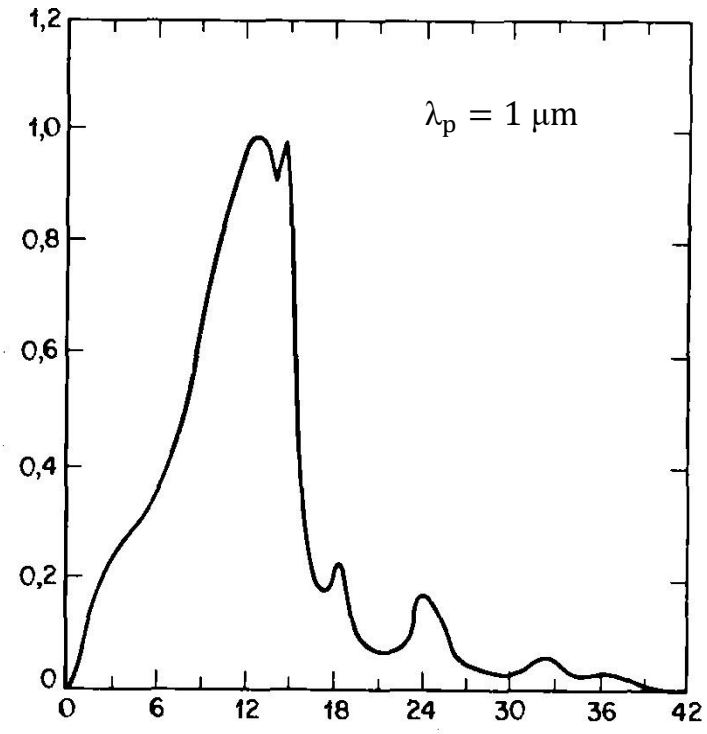


Fig. 2.6.3. Spectral dependence of the normalized amplification factor $g(\omega) = 2\alpha_{ampl}(\omega) / I_0$

frequency shifted with respect to the excitation frequency approximately by 12 – 14 THz. At the initial stage of interaction, when spontaneous Raman scattering is excited by pump radiation, within optical fiber the photons are generated at all the frequencies in the amplification band. However, the frequency component associated with a maximal amplification factor is growing more rapidly than all other components. At a threshold power of exciting radiation the intensity of this component is growing almost exponentially. In this case spontaneous Raman scattering goes to the stimulated Raman scattering stage accompanied by the generation of a Stokes wave, with the frequency determined by the Raman amplification peak for fused quartz. Using optical fibers of great length (hundreds of meters), one is enabled to lower the threshold intensity by several orders of magnitude.

Owing to the use of a Raman-active crystal in an optical cavity, one can lower the threshold of stimulated Raman scattering and increase the conversion factor. A system comprising a cavity with a Raman-active element within is referred to as the stimulated Raman scattering (SRS) laser. The input mirror of the cavity is transparent for pump radiation, completely reflecting the Stokes radiation. On the contrary, the output mirror is semitransparent to provide exit of stimulated Raman scattering radiation into the exterior. In the case of a SRS laser we can realize different generation modes of the Stokes components by selection of the appropriate spectral dependence for the quality factor of the cavity. For example, mirrors may be manufactured with a very low reflection factor (~1 %) for the second Stokes component. Actually, this results in suppression of the second Stokes component generation, allowing for the efficient generation of the first Stokes component only. If the output mirror is opaque for the first Stokes component and semitransparent for the waves with greater lengths, a SRS laser radiates the higher-order Stokes components. When the output mirror completely reflects pump radiation, the so-called round-trip pumping mode is realized, that leads to lowering of the stimulated Raman scattering threshold because the converted radiation makes its roundtrip in the SRS medium.

So, with the use of a long optical fiber or of a laser cavity, we can lower the threshold power of pump considerably (by several orders) down to kilowatt or even watt levels.

2.6.2. Stimulated Brillouin scattering

Light scattering from acoustic waves (inhomogeneities of the medium density) was discovered by Brillouin in 1922. At the same time, light scattering in solids was independently studied theoretically by the Soviet physicist Mandelshtam.

The phenomenon of stimulated Brillouin scattering – an acoustic wave is created by light that is subsequently scattered – was discovered in 1964. An alternating electric field induces, due to electrostriction, a variable deformation of a liquid or of a crystal with excitation of an acoustic wave. On the other hand, the acoustic wave modulates the dielectric medium permittivity and this can lead to the energy exchange between the waves, the difference in the frequencies of which is equal to the acoustic wave frequency. This phenomenon is similar to stimulated Raman scattering but, instead of molecular vibrations, an acoustic wave is involved. In analogy with stimulated Raman scattering, stimulated Brillouin scattering has a threshold, the threshold intensity sometimes exceeding

that of SRS. The self-focusing effect, resultant in increase of the light beam intensity beyond the threshold, in some media may also contribute to stimulated Brillouin scattering.

The characteristics feature of stimulated Brillouin scattering is a minor frequency shift of the scattered wave ω_{SBS} with respect to the pump frequency ω_0 because speed of sound in a medium is considerably lower than speed of light in the same medium. To find a value of this shift, we write the momentum and energy conservation laws in the following form:

$$\omega_{SBS} = \omega_0 - \omega_{acoustic}, \tag{2.6.21}$$

$$\Delta\vec{k} = \vec{k}_0 - \vec{k}_{SBS} - \vec{k}_{acoustic} = 0. \tag{2.6.22}$$

Stimulated Brillouin scattering is most effective when the scattered wave is propagating counter to the pump wave direction ($\vec{k}_0 \uparrow\downarrow -\vec{k}_{SBS}$). Considering that $\omega_{SBS} \approx \omega_0$, from (2.6.22) we get

$$k_{acoustic} \approx 2k_0. \tag{2.6.23}$$

From this it follows that

$$\omega_{acoustic}/v_{acoustic} \approx 2\omega_0 n/c. \tag{2.6.24}$$

Substituting the typical values of speed for acoustic waves in liquids $v_{acoustic} \approx 10^5$ cm/s, we obtain the frequency of the light-induced acoustic wave that is in the order of a few gigahertz (10^9 – 10^{10} Hz). In this case the efficiency of stimulated Brillouin scattering may be as high as 90 %, exceeding the efficiency of SRS. Stimulated Brillouin scattering in the counter direction features the possibility of observing the wavefront conjugation effect as with the four-wave interaction considered in Section 2.5. The backward reflected wave has the same spatial amplitude and phase distribution as the incident wave. Earlier it has been demonstrated that this property of the reflected wave enables one to compensate for the phase distortions when light beams are propagating within inhomogeneous media, e.g., in active elements of high-power lasers and optical amplifiers.

References

1. D.L. Mills. Nonlinear Optics. Basic Concepts. Springer-Verlag Berlin Heidelberg, 1998.

2. A. Newell, J. Moloney. Nonlinear Optics. CRC Press, 2003.

3. R.W. Boyd. Nonlinear Optics. Academic Press, 2008.

4. B.B. Laud. Lasers and Non-Linear Optics. New Age International Publishers, 2011.

5. Y.V.G.S. Murti, C.Vijayan. Essentials of Nonlinear Optics. Wiley, 2014.

6. P.E. Powers, J.W. Haus. Fundamentals of Nonlinear Optics. CRC Press, 2017.

Chapter 3. Coherent Optics and Holography

Introduction. Holography – development stages.

Holography (from Greek *Όλος — holos* «whole» + *γραφή — graphe* «writing, drawing») – technique for recording and subsequent reconstruction of a wave field based on recording the interference pattern of the object wave, containing information about the object, and of the reference wave, coming directly from the radiation source. This interference pattern recorded in a light-sensitive medium is termed the **hologram**. When a light wave similar to the reference wave is directed onto a hologram, a complete reconstruction of the amplitude and phase takes place for the reference light wave coming from the recorded object.

D.Gabor (Great Britain) is a founder of holography – he was the first to record a hologram in 1947. Discovery of holography (the Nobel Prize in Physics in 1971) has been made in the process of the experiments in effort to improve resolution of an electron microscope. The technique proposed by Gabor is based on recording the resultant interference of the wave field coming from the object with a plane reference wave that is coherent to it. This makes it possible to have complete optical information (amplitude, phase, and polarization of a light field) about the involved object, as distinct from photography recording only the amplitude distribution of a light field.

Two fundamental optical phenomena form the basis for recording and reconstruction of a holographic image: ***light interference*** and ***diffraction***. When in a light-sensitive medium we record an interference structure formed by the object wave and by the reference wave coherent with it, and then illuminate this structure by the reference wave, we can reconstruct the object wave due to light diffraction from the recorded interference structure. Reconstruction of a

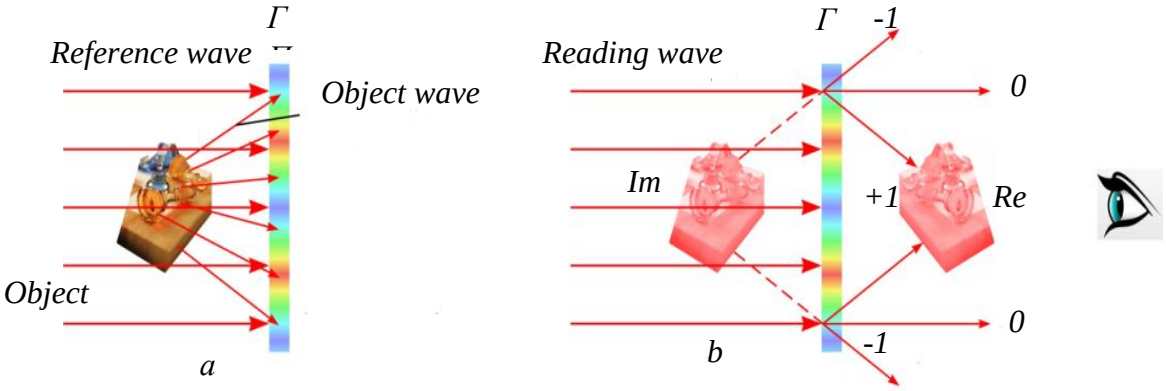


Figure 3.0.1- Gabor scheme for recording (*a*) and reconstruction (*b*) of a hologram.

Re and Im – real and imaginary images of the object.

hologram in the diffracted beams is associated with reconstruction of the wave field containing information about the spatial structure of an object – we can visually percept its volume.

The Gabor scheme is a uniaxial scheme with the object positioned in a field of the reference wave. A part of the light beam scattered from a transparency (object) creates an object wave, whereas the forward transmitted (nonscattered) light represents a reference wave (Fig.3.0.1, *a*). A limitation of this scheme is that, on reconstruction of a hologram, the light beams forming real (*Re)* and virtual (*Im*) images of the object, and also the forward transmitted light are propagating in the same direction. This prevents perception of the image and leads to lower resolution of the hologram (Fig. 3.0.1, *b*).

Since the advent of lasers in 1960, holography has acquired practical significance. E.Leith and Yu.Upatnieks (USA) were the first to form a volume transmission hologram (1962) reconstructed in laser light, thus initiating image holography. They suggested to use a two-beam scheme (scheme with an oblique reference beam) that makes it possible to carry out the reconstructed images from the propagation path of a reading beam (Fig.3.0.2). With this hologram recording scheme, an object is illuminated by a separate coherent beam and hence the possibility of recording holograms for opaque and three-dimensional objects is offered.

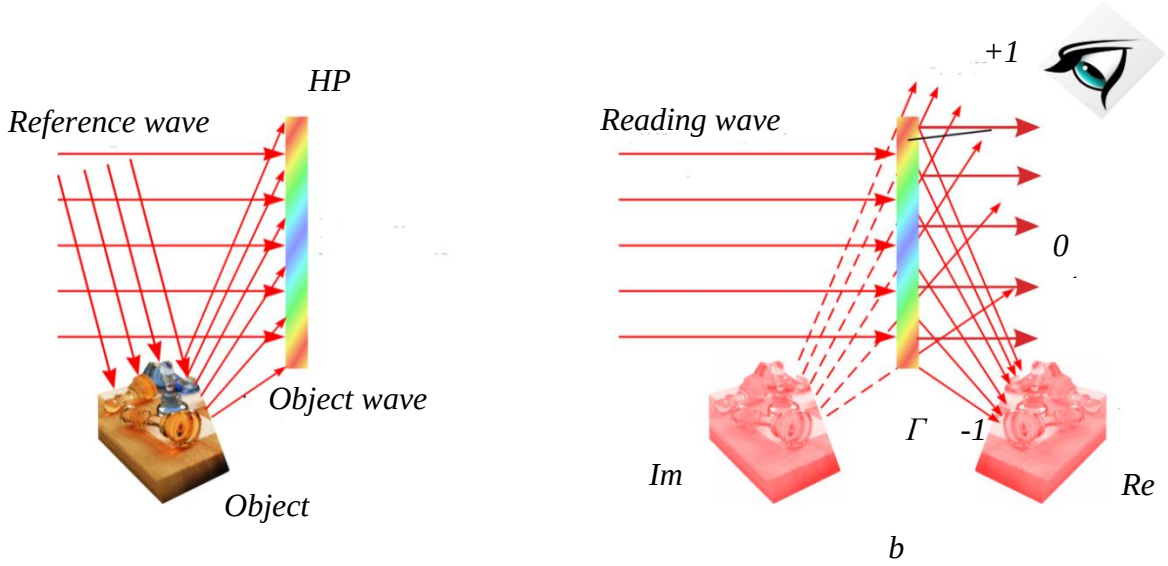


Figure 3.0.2 – Leith-Upatnieks scheme for recording (*a*) and reconstruction (*b*) of a hologram.

The schemes in Figs. 3.0.1 and 3.0.2 are used for recording of thin transmission holograms (thin diffraction gratings). In this case two images are reconstructed: virtual image is reconstructed at the place, where an object is

positioned during the hologram recording, and a real image is at the other side of the hologram. The reconstruction process of a holographic image is associated with diffraction of the hologram-reconstructing light from the interference structure recorded in the hologram. In fact, this interference structure is a diffraction grating. Diffraction from a thin sine diffraction grating may be of two orders: +1 order diffraction and -1 order diffraction (see Fig. 3.0.1). A virtual image is reconstructed in the course of divergent beams, while convergent beams form a real image of the object. A real image may be recorded when we place a photoplate or photodetector at the image location. A virtual image may be observed or recorded with the help of an objective optical lens. Besides, reconstruction of thin holograms requires illumination of a hologram with coherent radiation. There is no possibility of reconstructing such holograms by white light because different spectral components diffract at different angles, making images blurred. The problem was solved by the Soviet physicist Yu.N. Denisyuk (Russia) who in 1962 put forward the idea of recording volume reflection holograms with their reconstruction in white light. When holograms are recorded according to the Denisyuk scheme, the object and reference beams come from different sides of the holographic plate. Their interference within the volume of a light-sensitive medium, due to variations in the refractive index and/or absorption factor, leads to recording of a system of the layers having different physical properties and to the formation of a volume diffraction hologram. On reconstruction of a volume hologram in white light, only the spectral component meeting a particular condition (Bragg diffraction condition) is diffracted.

In 1969 S.Benton (USA) recorded a thin transmission hologram visible in an ordinary white light. Such holograms are termed the rainbow holograms because they are iridescent as a rainbow when their position is changed. This investigation has laid the basis for mass production of inexpensive holograms by forming interference patterns on plastic. The holograms of this type are used against counterfeiting of documents and products.

The first mechanically embossed hologram was manufactured in 1974 by the transformation of interference lines of the rainbow hologram into the surface relief.

In 1979 the technology for mass replication of relief holograms was developed.

In 1982 Master Card adds a hologram to the credit card for protection against forgery.

In 1988 DuPont produced a photopolymeric material to record volume reflection holograms.

A digital holography debuts in 1991 in the dot-matrix form.

In 2005 the InPhase Co. jointly with Hitachi/Maxxell developed the technology to produce holographic data-storage discs, HVD, holographic versatile discs.

In 2006 Smart Holograms created volume holograms for identification of different substances contained in materials, the so-called sensor holograms.

3.1. Coherence

The advent of coherent light sources opened great potentialities of holography. The notion of ***coherence*** that is associated with concurrent (correlated) proceeding in time of several vibrational or wave processes has originated from a classical vibration theory. The waves are characterized by their ability to coherent (i.e. correlated, orderly in time and space) summation. Lasers differ from chaotic sources by their coherence degree.

Coherence of a common thermal source is qualitatively different from that of a laser. Chaotic sources of optical radiation are composed of numerous oscillators (molecules, atoms, ions). Their emission acts are following separately in a statistical disorder. Phases of the emitted waves are also distributed chaotically. The interaction of optical radiations from two chaotic sources is observed at the screen as a statistical distribution of the total intensity without interference. Radiation of chaotic sources is incoherent, there is no correlation between the phase, amplitude, polarization of separate electromagnetic waves. To observe interference of such sources, the wave front is divided, and then we can observe the interference of two waves within a single emission act – the point is that coherence involves two beams of one and the same ***spontaneous*** emission act. The interference and diffraction phenomena based pricisely on coherence were found in the 60-ies of the XVII century. A simple law of interference was put forward by Young in 1801. The first formulation of coherence also belongs to him: «Two parts of one and the same wave can interfere». The classical Young's experiment with interference at two slits, performed to demonstrate a wave nature of light, is still used nowadays to check a degree of coherence for both chaotic and laser sources.

Laser radiation is generated due to ***stimulated*** (induced) emission. Coherence of the induced emission is determined by correlation between all the physical characteristics of stimulating and induced emissions. According to the generally accepted knowledge, real light fields can be considered as a superposition of electromagnetic waves (with differing amplitudes, phases, and frequencies) emitted by the excited atoms of a material. As an electromagnetic wave represents an oscillation occurring in time and space, the resultant summation of a group of the waves (spatial radiation-intensity distribution) depends on the spectral width of the involved frequencies and on the space-time correlation (coherence) of the wave phases.

We distinguish between spatial and temporal coherence of light beams. The temporal coherence is understood as a correlation degree of the electric vector oscillations for separate waves of the light beam arriving from the source to the specified space point with a delay relative each other. Similarly, spatial coherence of a light field is understood as a degree of correlation between oscillations of the waves at two points in space and at the fixed instant of time. The coherence of radiation established at two spatiotemporal points is termed the ***second-order coherence***. Such optical phenomena as interference and diffraction of light waves represent the exhibited second-order coherence and are used to estimate a degree of coherence for the light beams.

3.1.1. Mutual coherence function and complex degree of coherence

For convenience and for better understanding of the obtained (theoretical and experimental) results, the phenomenon of coherence is considered for light beams emitted as a train of the waves with a limited temporal frequency spectrum. The light beams formed by laser sources are good examples of the spectral-limited fields.

In the scalar approximation, the spatiotemporal state of the electric intensity $E(\vec{r},t)$ for an electromagnetic field of this-type radiation at some space point with the radius vector $\vec{r}$ in the instant of time t may be described by the following equation for a quasi-plane quasi-monochromatic wave:

$$E(\vec{r},t) = a(\vec{r},t)\cos\left[\left(\omega t - \vec{k}\cdot\vec{r}\right) + \phi(\vec{r},t)\right], \qquad (3.1.1)$$

where $a(\vec{r},t)$ and $\phi(\vec{r},t)$ – real-values functions slowly varying, with respect to the function $\cos(\omega t - \vec{k}\cdot\vec{r})$,; $\omega = 2\pi/T$ – average frequency of the field oscillations (T – average oscillation period of an electric field), $\vec{k} = (2\pi/\lambda)\vec{n}$ – wave vector, $\vec{n}$ – unit vector for the propagation direction of a wave, λ – average wavelength. To simplify our calculations, we relate the field $E(\vec{r},t)$ to the so-called analytical signal

$$V(\vec{r},t) = g(\vec{r},t)\exp\left[i\left(\omega t - \vec{k}\cdot\vec{r}\right)\right], \tag{3.1.2}$$

where $g(\vec{r},t) = a(\vec{r},t)\exp\left[i\phi(\vec{r},t)\right]$ – slowly varying complex amplitude of the wave-train field, $|g(\vec{r},t)| = |a(\vec{r},t)|$ – absolute value of the amplitude, $\phi(\vec{r},t)$ – its phase. Comparing (3.1.1) and (3.1.2), we can see that

$$E(\vec{r},t) = \mathrm{Re}\{V(\vec{r},t)\}, \tag{3.1.3}$$

i.e., the electric field strength $E(\vec{r},t)$ is equal to a real part of the signal $V(\vec{r},t)$. Therefore, instead of the real function $E(\vec{r},t)$, the complex function $V(\vec{r},t)$ may be used, taking after the mathematical transformations a real part of the result.

In the general case, the functions $a(\vec{r},t)$ and $\phi(\vec{r},t)$ and hence the field strength $E(\vec{r},t)$ as well as the associated signal $V(\vec{r},t)$ are random variables. Note that either $a(\vec{r},t)$ or $\phi(\vec{r},t)$ or both of them may be random. These functions are varying rather rapidly as compared to the observation or field recording time T_0 but considerably slower than the average field period T.

Radiation detectors record not instant values of the field oscillations $E(\vec{r},t)$ at the frequencies ~ 10^{14} Hz but only the intensity averaged over several realizations $\langle I(\vec{r},t)\rangle_e$, where $I(\vec{r},t) = 2E^2(\vec{r},t) = |V(\vec{r},t)|^2$ – instant intensity of the wave train, $\langle\ \rangle_e$ denotes the averaging operation over an ensemble of instantaneous intensity realizations. The factor 2 before $E^2(\vec{r},t)$ is associated with the fact that a real signal includes both positive and negative frequencies. This is obvious if (3.1.1) is given in the complex form. In this case the amplitude of a signal is identically distributed between positive and negative frequencies. As follows from (3.1.2), the analytical signal involves positive frequencies only. Since radiation detectors do not respond to negative branches of the frequencies,

the energy of a real signal is determined by the signal part including positive branches of the frequencies with the amplitude coming to a half of that for the analytical signal in our notation.

For random fields with the average values of fluctuating parameters invariable in time, that is for stationary fields, when a correlation interval of fluctuations is fairly narrow (ergodicity of the fields) and when the intensity $\langle I(\vec{r},t)\rangle_e$ averaged over an ensemble of the independent realizations is equal to the time-averaged intensity of one long realization, we have

$$\langle I(\vec{r},t)\rangle_e = \langle I(\vec{r},t)\rangle_T = \lim_{T_0\to\infty}\left\{\frac{1}{2T_0}\int_{-T_0}^{T_0} 2E^2(\vec{r},t)\,dt'\right\} = \lim_{T_0\to\infty}\left\{\frac{1}{2T_0}\int_{-T_0}^{T_0}\left|V(\vec{r},t')\right|^2 dt'\right\} \quad (3.1.4)$$

Considering that $T_0 >> T$, we assume $T_0 \to \infty$. Note that, as a rule, the second averaging variant is realized. Consequently, further we treat the light field under study as meeting the stationarity and ergodicity condition, and the averaging operation is given in broken brackets $\langle\ \rangle$ without any index.

Let us consider the averaged result of summation at some spatial point *P* sufficiently distant from the plane *R* , with the radius-vector $\vec{s}$ of two fields of the type given by (3.1.1) which are resultant from division of a particular initial source filed *D* by the screen *R* having two holes (Fig. 3.1.1).

Let light oscillations arrive to the point *P* from points of the holes $\vec{r}_1$ and $\vec{r}_2$ (secondary sources) with a delay $\tau = t_1 - t_2$ relative each other. Then a light field in the instant of time *t* at the point $\vec{s}$ is represented as follows:

$$\begin{aligned} V(\vec{s},t) &= V(\vec{r}_1,t) + V(\vec{r}_2,t-\tau) = g(\vec{r}_1,t)\exp\left[i\left(\omega t - \vec{k}_1\cdot\vec{s}\right)\right] + \\ &+ g(\vec{r}_2,t-\tau)\exp\left[i\left(\omega(t-\tau) - \vec{k}_2\cdot\vec{s}\right)\right], \end{aligned} \quad (3.1.5)$$

where $g(\vec{r}_1,t)$ and $g(\vec{r}_2,t-\tau)$ – complex amplitudes of the initial field at the points $\vec{r}_1$ and $\vec{r}_2$, respectively. Next, for the averaged intensity $\langle I(\vec{s},t)\rangle = \left\langle\left|V(\vec{s},t)\right|^2\right\rangle$ we get

$$\begin{aligned} \langle I(\vec{s},t)\rangle &= \left\langle\left|a(\vec{r}_1,t)\right|^2\right\rangle + \left\langle\left|a(\vec{r}_2,t)\right|^2\right\rangle + \\ &+ 2\,\mathrm{Re}\left\{\left\langle a(\vec{r}_1,t)\,a^*(\vec{r}_2,t-\tau)\right\rangle\exp\left[-i\left(\omega\tau + \left(\vec{k}_1-\vec{k}_2\right)\cdot\vec{s}\right)\right]\right\} = \\ &= \langle I(\vec{r}_1,t)\rangle + \langle I(\vec{r}_2,t)\rangle + \\ &+ 2\left|\tilde{A}_{12}(\vec{r}_1,\vec{r}_2,\tau)\right|\cos\left[\omega\tau + \left(\vec{k}_1-\vec{k}_2\right)\cdot\vec{s} + \delta(\vec{r}_1,\vec{r}_2,\tau)\right]. \end{aligned} \quad (3.1.6)$$

Here $I(\vec{r}_1,t)=\left\langle\left|a(\vec{r}_1,t)\right|^2\right\rangle$ and $I(\vec{r}_2,t)=\left\langle\left|a(\vec{r}_2,t)\right|^2\right\rangle$ – average radiation intensities independently created at the point $\vec{s}$ by each of the summated fields $\tilde{A}_{12}(\vec{r}_1,\vec{r}_2,\tau)=\left\langle a(\vec{r}_1,t)a^*(\vec{r}_2,t-\tau)\right\rangle$, $\delta(\vec{r}_1,\vec{r}_2,\tau)=\phi(\vec{r}_1,t)-\phi(\vec{r}_2,t-\tau)$.

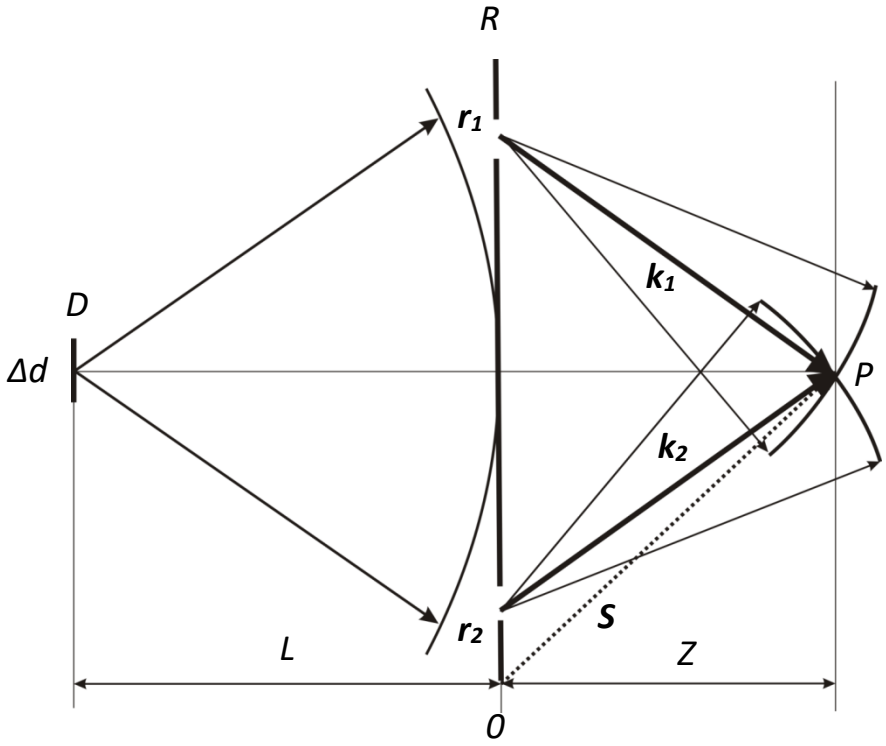


Figure 3.1.1 - Addition of two light fields

As seen from (3.1.5), the total intensity averaged over numerous realizations of the wave components created at the point $\vec{s}$ by the fields coming from the secondary sources which are positioned at the points $\vec{r}_1$ and $\vec{r}_2$ in the general case is not equal to the sum of the intensities created at the same point by the fields arriving from the points $\vec{r}_1$ and $\vec{r}_2$. A structure of the third term in (3.1.6) points to periodical modulation of the total light-field intensity. Such an intensity distribution created by two light beams is termed the interference pattern. The function $\tilde{A}_{12}$ is called the mutual coherence function of oscillations at the points $\vec{r}_1$ and $\vec{r}_2$; its argument $\arg\{\tilde{A}_{12}\}=\omega\tau+\delta(\vec{r}_1,\vec{r}_2,\tau)$ is termed the *phase* of this function. A form of the function Γ_{12} and the values it can take in the general case are determined by the position of the secondary sources on the screen and by a delay time of the waves arriving from these sources to the point $\vec{s}$. When oscillations of the waves at the points $\vec{r}_1$ and $\vec{r}_2$ are not correlated (radiation is spatially incoherent) or the delay time τ is so that oscillations are summed at the

point $\vec{s}$ without correlation (radiation is incoherent in time), a modulus of the function Γ_{12} is equal to zero and there is no interference pattern. Thus, interference is an exhibition of coherence of the initial light field formed by the source *D*. When radiation of the source is totally coherent, then for any positions of the secondary sources on the screen and for any delay time in arrival of the waves to the point $\vec{s}$, the function $\left|\tilde{A}_{12}\left(\vec{r}_1,\vec{r}_2,t-t_1,t-t_2\right)\right|$ is equal to 1, and the radiation intensity distribution within the plane passing through the point $\vec{s}$ is periodic.

In nature there are no totally coherent or totally incoherent fields. As a rule, they are partially coherent. To analyze a coherence degree of radiation, we use the normalized function

$$\gamma\left(\vec{r}_1,\vec{r}_2,\tau\right)=\frac{\tilde{A}_{12}\left(\vec{r}_1,\vec{r}_2,\tau\right)}{\sqrt{\left\langle I\left(\vec{r}_1,t\right)\right\rangle}\sqrt{\left\langle I\left(\vec{r}_2,t\right)\right\rangle}}, \tag{3.1.7}$$

that is called the complex coherence degree of a field. Considering (3.1.7), the expression of (3.1.6) takes the form

$$\begin{aligned}&\left\langle I\left(\vec{s},t\right)\right\rangle=\left\langle I\left(\vec{r}_1,t\right)\right\rangle+\left\langle I\left(\vec{r}_2,t\right)\right\rangle+\\&+2\sqrt{\left\langle I\left(\vec{r}_1,t\right)\right\rangle}\sqrt{\left\langle I\left(\vec{r}_2,t\right)\right\rangle}\left|\gamma\left(\vec{r}_1,\vec{r}_2,\tau\right)\right|\cos\left[\omega\tau+\left(\vec{k}_1-\vec{k}_2\right)\cdot\vec{s}+\delta\left(\vec{r}_1,\vec{r}_2,\tau\right)\right].\end{aligned} \tag{3.1.8}$$

The intensities at a maximum and at a minimum of the interference pattern are given respectively as follows:

$$\left\langle I_{\max}\left(\vec{s},t\right)\right\rangle=\left\langle I\left(\vec{r}_1,t\right)\right\rangle+\left\langle I\left(\vec{r}_2,t\right)\right\rangle+2\sqrt{\left\langle I\left(\vec{r}_1,t\right)\right\rangle}\sqrt{\left\langle I\left(\vec{r}_2,t\right)\right\rangle}\left|\gamma\left(\vec{r}_1,\vec{r}_2,\tau\right)\right|, \tag{3.1.9}$$

$$\left\langle I_{\min}\left(\vec{s},t\right)\right\rangle=\left\langle I\left(\vec{r}_1,t\right)\right\rangle+\left\langle I\left(\vec{r}_2,t\right)\right\rangle-2\sqrt{\left\langle I\left(\vec{r}_1,t\right)\right\rangle}\sqrt{\left\langle I\left(\vec{r}_2,t\right)\right\rangle}\left|\gamma\left(\vec{r}_1,\vec{r}_2,\tau\right)\right|. \tag{3.1.10}$$

The quantity

$$V=\frac{I_{\max}-I_{\min}}{I_{\max}+I_{\min}} \tag{3.1.11}$$

is termed the contrast of the interference pattern. It is a measure of sharpness of the interference fringes. In the case, when $\left\langle I\left(\vec{r}_1,t\right)\right\rangle=\left\langle I\left(\vec{r}_2,t\right)\right\rangle$, we have $V=\left|\gamma\left(\vec{r}_1,\vec{r}_2,\tau\right)\right|$.

Constructing a curve for the interference-pattern fringe visibility as a function of relative positions of the secondary sources and of delay time for the

light oscillations arriving to the point $\vec{s}$, one can find from (3.1.7) a modulus of the mutual coherence function Γ_{12}. To obtain complete information, we should know the phase $\arg\{\tilde{A}_{12}\}$. As follows from (3.1.6), a phase of the mutual coherence function determines positions of maxima (peaks) for the interference fringes. Indeed, at the interference-pattern point with the coordinates $(s_x = 0, s_y = 0)$ maxima of the interference pattern are observed when $\arg\{\tilde{A}_{12}\} = 2m\pi$ ($m = 0, \pm 1,$ $\pm 2\ldots$). If $\arg\{\tilde{A}_{12}\}$ takes other values, a maximum of the interference pattern is shifted with respect to the point $(s_x = 0, s_y = 0)$. When a delay time for the waves arriving to the point $(s_x = 0, s_y = 0)$ is 0 ($\tau = 0$), it is clear that a shift of the interference pattern maximum is determined by the phase difference of oscillations at the points $\vec{r}_1$ and $\vec{r}_2$: $\delta(\vec{r}_1, \vec{r}_2, \tau) = \phi(\vec{r}_1, t) - \phi(\vec{r}_2, t - \tau)$. It is known that illumination of the holes with monochromatic light, due to the 2π phase lag of oscillations at these points, results in the interference pattern shifted by $|\vec{r}_2 - \vec{r}_1|\bar{\lambda} / Z$, where $\bar{\lambda}$ – average wavelength. For the arbitrary phase difference $\delta(\vec{r}_1, \vec{r}_2, \tau)$, the fringes formed in quasi-monochromatic light are shifted (relative to those which would be formed under in-phase illumination of the holes with a monochromatic light having the same wave length) by

$$\Delta y = \frac{|\vec{r}_2 - \vec{r}_1|\bar{\lambda}}{2\pi Z}\delta(\vec{r}_1, \vec{r}_2, \tau) = \frac{|\vec{r}_2 - \vec{r}_1|\bar{\lambda}}{2\pi Z}\left(\arg\{\tilde{A}_{12}\} - \omega\tau\right). \qquad (3.1.12)$$

Finding Δy, with the use of (3.1.12) we can derive $\arg\{\tilde{A}_{12}\}$.

3.1.2. Time coherence

As follows from (1.7), a modulus of the complex coherence degree in the general case is dependent on relative positions of the secondary sources (coordinates $\vec{r}_1$ and $\vec{r}_2$) and on the delay time τ of the waves arriving to the observation point $\vec{s}$. If the initial radiation from a light source in some time *t* would be divided into two beams so that they were of the same initial phases to meet the following condition

$$\phi(\vec{r}_1, t) = \phi(\vec{r}_2, t) = \phi(\vec{r}, t), \qquad (3.1.13)$$

and then would be brought together in a certain observation plane, the contrast of the formed interference pattern would be dependent on τ only. Measuring contrast of this pattern for different values of τ (different relative time shifts of the beams), we can construct a curve for the modulus coherence degree as a function of this parameter only. This function is indicative of the so-called radiation time coherence, showing a degree of timing for the wave oscillations.

A classical Michelson interferometer is applicable to study time coherence of radiation. Its optical scheme is shown in Fig. 3.1.2.

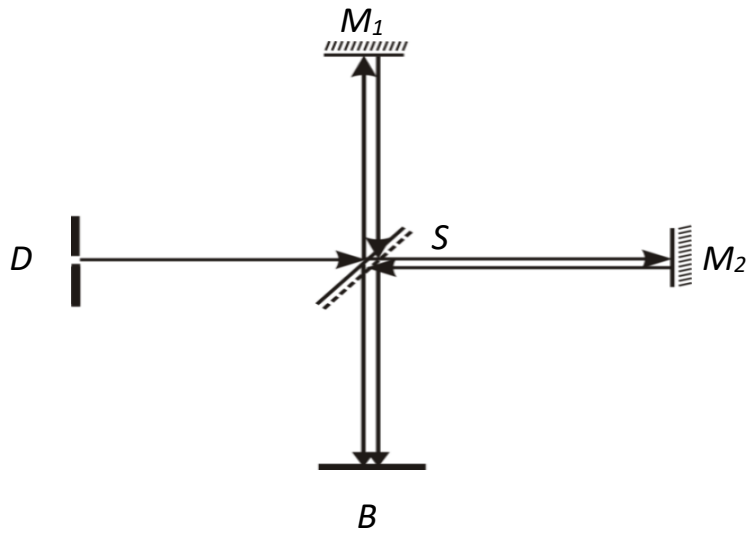


Figure 3.1.2 – Optical scheme of a Michelson interferometer

A light beam from the source *D* is directed to the semitransparent mirror *S*. The mirror divides the beam into two beams of approximately equal intensities, which are incident on the mirrors M_1 and M_2. The mirror-reflected beams are summed up in the plane B. To form sharp interference patterns in this plane, the optical paths of these beams SM_1B and SM_2B should be equalized carefully. One of the interferometer mirrors is tilted at a small angle to the normal wave front of the incident beam to allow for the formation of equal- thickness fringes in the plane B. In this plane an array of photodetectors (CCD linear detector) is aligned perpendicular to the direction of interference fringes to record an interference pattern. To estimate the time coherence degree for radiation, we should determine the visibilities, or contrasts, of interference patterns *V* for different delay times τ or the associated delay lengths $l = c\tau$ of one beam with respect to the other. A delay is attained by moving the nontilted mirror along the normal to the beam. Then the curve is constructed for the modulus time coherence γ or for the equal interference pattern visibility *V* as a function of the delay time τ or of the parameter

l. The width $\Delta\tau$ of the function $|\gamma(\tau)|$ at the half-height is called the coherence time of radiation. The associated path-length difference $\Delta l = c\Delta\tau$ (c – speed of light) is termed the coherence length, or longitudinal coherence «scale».

It is interesting which of the characteristics of a light field are responsible for the coherence scale, $\Delta\tau$ or Δl ? To answer this question, let us consider another – spectral – model for the representation of light beams. According to this model, atoms emit the light waves which are time-limited rather than infinite in time. An electric field of the train of such waves may be described by the function $F(\vec{r},t)$ meeting the conditions

$$F(\vec{r},t) = F_T(\vec{r},t) \text{ for } |t| \le T,$$

$$F(\vec{r},t) = 0 \text{ for } |t| > T, \qquad (3.1.14)$$

where T – time of the wave train.

Despite the fact that the function $F(\vec{r},t)$ is space and time random, due to its limited nature, we can use the Fourier transform operation, i.e., we can find some function $f(\vec{r},\omega)$ representing a spectrum of the temporal perturbation frequencies $F(\vec{r},t)$. In the case of a direct Fourier transform we can write

$$f_T(\vec{r},\omega) = \int_{-\infty}^{\infty} F_T(\vec{r},t)\exp(i\omega t)\,dt, \qquad (3.1.15)$$

where ω – frequency of the spectral component. An inverse Fourier transform is as follows:

$$F_T(\vec{r},t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f_T(\vec{r},\omega)\exp(-i\omega t)\,d\omega. \qquad (3.1.16)$$

With the use of (3.1.15) and (3.1.16), for the correlation function we get

$$\tilde{A}(\vec{r},\tau)=\lim_{T\to\infty}\left\{\frac{1}{2T}\int_{-\infty}^{\infty}F_T(\vec{r},t)F_T^{*}(\vec{r},t-\tau)\,dt\right\}=$$
$$=\frac{1}{2\pi}\lim_{T\to\infty}\left\{\frac{1}{2T}\int_{-\infty}^{\infty}F_T(\vec{r},t)\left(\int_{-\infty}^{\infty}f_T^{*}(\vec{r},\omega)\exp\left[i\omega(t-\tau)\right]d\omega\right)dt\right\}=$$
$$=\frac{1}{2\pi}\lim_{T\to\infty}\left\{\frac{1}{2T}\int_{-\infty}^{\infty}f_T^{*}(\vec{r},\omega)\exp(-i\omega\tau)\left(\int_{-\infty}^{\infty}F_T(\vec{r},t)\exp(i\omega t)\,dt\right)d\omega\right\}= \tag{3.1.17}$$
$$=\frac{1}{2\pi}\lim_{T\to\infty}\left\{\frac{1}{2T}\int_{-\infty}^{\infty}\left|f_T(\vec{r},\omega)\right|^2\exp(-i\omega\tau)\,d\omega\right\}.$$

Smoothing the calculation result for (3.1.17), i.e., taking an ensemble average of the perturbations $F_T(\vec{r},t)$, we derive

$$\tilde{A}(\vec{r},\tau)=\frac{1}{2\pi}\int_{-\infty}^{\infty}G(\omega)\exp(-i\omega\tau)\,d\omega=\frac{1}{2\pi}\int_{-\infty}^{\infty}G(\omega)\exp(i\omega\tau)\,d\omega, \tag{3.1.18}$$

where $G(\omega)=\lim_{T\to\infty}\left\{\overline{\left|f_T(\vec{r},\omega)\right|^2}/(2T)\right\}$ – spectral radiation density. As follows from (3.1.18), the mutual coherence function is related to the spectral density (energy spectrum) by the Fourier transform. By an inverse Fourier transform we obtain the following:

$$G(\omega)=\int_{-\infty}^{\infty}\Gamma(\vec{r},\tau)\exp(-i\omega\tau)\,d\tau. \tag{3.1.19}$$

Correlation of the effective spectral widths for these functions may be given by the relation

$$\Delta\tau\Delta\omega\geq\frac{1}{2}\quad, \tag{3.1.20}$$

that in optics is known as reciprocity relation. In (3.1.20) $\Delta\tau$ – time coherence ratio (coherence time) and $\Delta\omega$ – effective spectral width of the temporal radiation frequencies. When approximately the spectral density $G(\omega)$ is of a Gaussian form and its average frequency $\bar{\omega}$ is high as compared to the effective spectral width $\Delta\omega$, the inequality sign in (3.1.20) is replaced by the order of magnitude sign as follows:

$$\Delta\tau\Delta\omega\sim\frac{1}{2}. \tag{3.1.21}$$

From (3.1.21) it follows that the radiation coherence time is one-to-one determined by the spectral width of its temporal frequencies. Knowing the average wave length of radiation $\bar{\lambda}$ and its spectral width $\Delta\lambda$, we can estimate a coherence width by the formula

$$\Delta l = \frac{\bar{\lambda}^2}{\Delta\lambda}. \tag{3.1.22}$$

3.1.3. Spatial coherence

As light fields are formed by electromagnetic waves emitted by the excited particles of a material distributed within a unit volume, at a given time the phases of these waves may differ from each other at different spatial points. A degree of phase correlation for the train of waves at a specific instant of time at two points of space is characteristic for its spatial coherence. In the process of this laboratory work, we can study the transverse spatial coherence of a light field, i.e., the light coherence at two points positioned within the plane perpendicular to the field propagation vector.

An analysis of the light-field spatial coherence degree is based on examination of the interference pattern contrast formed by the light waves arriving to the observation point without a delay relative each other from two points of the wavefront. Such an interference pattern can be formed with the help of a Young interferometer, see Fig. 3.1.3. This scheme is similar to that used to study the basic relations in a theory of the second-order coherence.

In this way the intensity distribution in the Young interference pattern is described by the expression similar to that of (3.1.18) derived using the calculation model, except of the fact that in this case the delay time τ of the waves is zero. Then we can obtain

$$\begin{aligned}
&\left\langle I\left(\vec{s},t\right)\right\rangle = \left\langle \left|a\left(\vec{r}_1,t\right)\right|^2 \right\rangle + \left\langle \left|a\left(\vec{r}_2,t\right)\right|^2 \right\rangle + \\
&+2\,\mathrm{Re}\left\{ \left\langle a\left(\vec{r}_1,t\right) a^*\left(\vec{r}_2,t\right)\right\rangle \exp\left[-i\left(\vec{k}_1 - \vec{k}_2\right)\cdot\vec{s}\,\right]\right\} = \\
&= \left\langle I\left(\vec{r}_1,t\right)\right\rangle + \left\langle I\left(\vec{r}_2,t\right)\right\rangle + 2\left|\Gamma_{12}\left(\vec{r}_1,\vec{r}_2,0\right)\right| \cos\left[\left(\vec{k}_1 - \vec{k}_2\right)\cdot\vec{s} + \delta\left(\vec{r}_1,\vec{r}_2,0\right)\right].
\end{aligned} \tag{3.1.23}$$

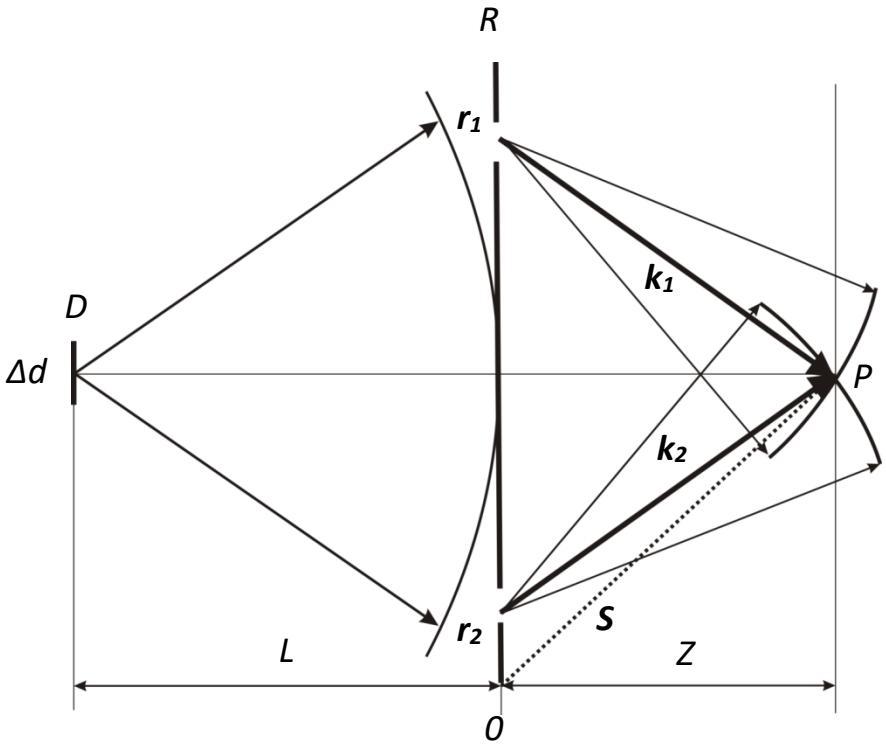


Figure 3.1.3 - Addition of quasiharmonic fields

In (3.1.23) the quantities $I(\vec{r}_1,t)=\left\langle\left|a(\vec{r}_1,t)\right|^2\right\rangle$ и $I(\vec{r}_2,t)=\left\langle\left|a(\vec{r}_2,t)\right|^2\right\rangle$ – average radiation intensities created independently by each of the summed fields at the observation point with the radius-vector $\vec{s}$. The function $\tilde{A}_{12}(\vec{r}_1,\vec{r}_2,0)=\left\langle a(\vec{r}_1,t)a^*(\vec{r}_2,t-\tau)\right\rangle$ is called the ***mutual spatial-coherence function*** of the oscillations at the points $\vec{r}_1$ and $\vec{r}_2$, and its argument $\delta(\vec{r}_1,\vec{r}_2,0)$ is named the ***phase*** of this function.

The normalized function

$$\gamma(\vec{r}_1,\vec{r}_2,0)=\frac{\tilde{A}_{12}(\vec{r}_1,\vec{r}_2,0)}{\sqrt{\left\langle I(\vec{r}_1,t)\right\rangle}\sqrt{\left\langle I(\vec{r}_2,t)\right\rangle}} \tag{3.1.24}$$

is a complex degree of the spatial field coherence.

The function

$$V=\frac{I_{\max}-I_{\min}}{I_{\max}+I_{\min}}, \tag{3.1.25}$$

where

$$\left\langle I_{\max}(\vec{s},t)\right\rangle=\left\langle I(\vec{r}_1,t)\right\rangle+\left\langle I(\vec{r}_2,t)\right\rangle+2\sqrt{\left\langle I(\vec{r}_1,t)\right\rangle}\sqrt{\left\langle I(\vec{r}_2,t)\right\rangle}\left|\gamma(\vec{r}_1,\vec{r}_2,0)\right|, \tag{3.1.26}$$

$$\left\langle I_{\min}\left(\vec{s},t\right)\right\rangle = \left\langle I\left(\vec{r}_1,t\right)\right\rangle + \left\langle I\left(\vec{r}_2,t\right)\right\rangle - 2\sqrt{\left\langle I\left(\vec{r}_1,t\right)\right\rangle}\sqrt{\left\langle I\left(\vec{r}_2,t\right)\right\rangle}\left|\gamma\left(\vec{r}_1,\vec{r}_2,0\right)\right|, \qquad (3.1.27)$$

is called the ***visibility (contrast) function*** of an interference pattern. When $\left\langle I\left(\vec{r}_1,t\right)\right\rangle = \left\langle I\left(\vec{r}_2,t\right)\right\rangle$, we have $V = \left|\gamma\left(\vec{r}_1,\vec{r}_2,0\right)\right|$.

The visibility of an interference pattern *V* and the modulus complex spatial-coherence degree $|\gamma|$ are characterizing coherence of the wave oscillations at the points $\vec{r}_1$ and $\vec{r}_2$. Actually, when oscillations of the waves at these points are noncorrelated, there is no periodic intensity modulation within the observation plane and hence *V* and $|\gamma|$ are equal to zero. In the other limiting case, when the oscillations at the points $\vec{r}_1$ and $\vec{r}_2$ are totally correlated, *V* and $|\gamma|$

The fields completely coherent in time or in space are nonexistent in nature. As a rule, they are partially coherent. To estimate a degree of the spatial coherence of a light field, we construct the curve for $|\gamma|$ as a function of different positions of the points $\vec{r}_1$ and $\vec{r}_2$ at the wave front. Note that the fields created by the majority of real sources are spatially inhomogeneous (though in some cases they may be considered homogeneous) and the modulus spatial coherence degree of such fields is dependent only on the difference in positions of the points $\vec{r}_1$ and $\vec{r}_2$, i. e. on the quantity $\left|\vec{r}_1 - \vec{r}_2\right|$. The so-called coherence radius – quantity that is equal to a width of the function $|\gamma|$ at its half-height – serves as a measure of the radiation spatial-coherence degree. To determine the radiation coherence radius, we should measure contrasts of the interference patterns obtained for different distances $\left|\vec{r}_1 - \vec{r}_2\right|$ between the points at the wave front of a light field and construct a curve for the observed contrast as a function of these distances.

Which characteristics of a radiation field are decisive for its coherence radius? To answer this question, we can find the mutual spatial coherence function at the given distance *L* from some extended source of the size ΔS (Figure 3.1.4).

Let us assume that the source S emits quasi-monochromatic light and its size is $\Delta S << L$. We determine the spatial coherence of radiation Γ at the two points of its wave front $P\left(\vec{r}_1\right)$ and $P\left(\vec{r}_2\right)$positioned within the plane of a screen *D*. By definition, we have

$$\tilde{A}\left(\vec{r}_1,\vec{r}_2,0\right)=\lim_{T\to\infty}\left\{\frac{1}{2T}\int_{-T}^{T}V\left(\vec{r}_1,t\right)V^{*}\left(\vec{r}_2,t\right)dt\right\}, \tag{3.1.28}$$

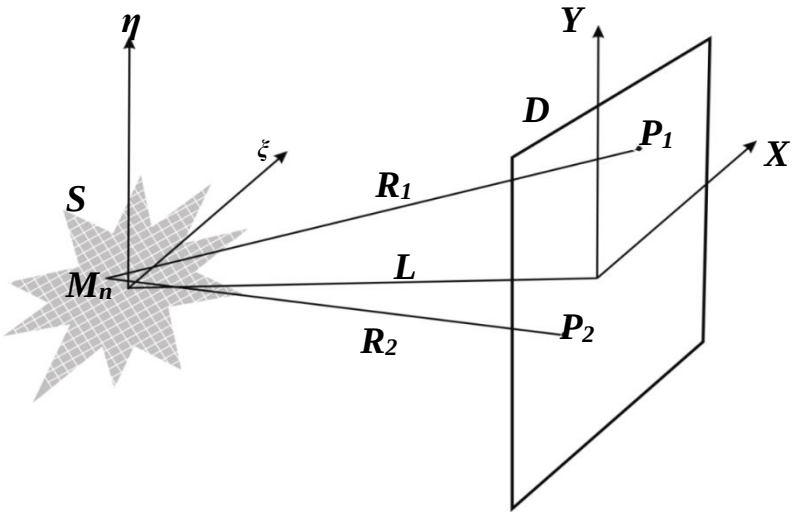


Figure. 3.1.4 – To determine the coherence radius

where $V\left(\vec{r}_1,t\right)$ and $V\left(\vec{r}_2,t\right)$ are the electric field strengths at the points of its wave front with the radius vectors $\vec{r}_1$ and $\vec{r}_2$. The electric field created by the atom positioned at the point $M_n\left(\vec{s}\right)$ of the source may be described by the analytical signal

$$V_n\left(\vec{s},t\right)=a_n\left(\vec{s},t\right)\exp\left(i\omega t\right), \tag{3.1.29}$$

where a_n - amplitude of the field emitted by the atom, ω – radiation frequency, $\vec{r}$ – radius-vector of the atom within the source plane. The oscillations at the points $P\left(\vec{r}_1\right)$ and $P\left(\vec{r}_2\right)$ created by the atom $M_n\left(\vec{s}\right)$ are respectively equal to

$$V_n\left(\vec{r}_1,t\right)=a_n\left(\vec{s},t-R_1/c\right)\frac{\exp\left[i\omega\left(t-R_1/c\right)\right]}{R_1}, \tag{3.1.30}$$

$$V_n\left(\vec{r}_2,t\right)=a_n\left(\vec{s},t-R_2/c\right)\frac{\exp\left[i\omega\left(t-R_2/c\right)\right]}{R_2}, \tag{3.1.31}$$

where R_1 and R_2 – distances from the point $M_n\left(\vec{s}\right)$ to the points $P\left(\vec{r}_1\right)$ and $P\left(\vec{r}_2\right)$, respectively; c – speed of light. The total fields at the points $P\left(\vec{r}_1\right)$ and $P\left(\vec{r}_2\right)$ are

due to all the emitters of the extended source *S*. Provided the transverse dimensions of the source are significantly smaller than the distance *L*, we have

$$V(\vec{r}_1,t)=\sum_n a_n(\vec{s},t-R_1/c)\frac{\exp\left[i\omega(t-R_1/c)\right]}{R_1}, \tag{3.1.32}$$

$$V(\vec{r}_2,t)=\sum_n a_n(\vec{s},t-R_2/c)\frac{\exp\left[i\omega(t-R_2/c)\right]}{R_2}. \tag{3.1.33}$$

The mutual coherence of oscillations at the points $P(\vec{r}_1)$ and $P(\vec{r}_2)$ may be found by the formula

$$\tilde{A}(\vec{r}_1,\vec{r}_2,0)=\left\langle V(\vec{r}_1,t)V^*(\vec{r}_2,t)\right\rangle=\sum_{m,n}\left\langle V_m(\vec{r}_1,t)V_n^*(\vec{r}_2,t)\right\rangle. \tag{3.1.34}$$

Note that, with the time-averaging operation within the broken brackets $\langle\ \rangle$, all the cross terms with $m\neq n$ go to zero due to mutual incoherence of different sources. Because of this,

$$\tilde{A}(\vec{r}_1,\vec{r}_2,0)=\sum_n\left\langle a_n(\vec{s},t-R_1/c)a_n^*(\vec{s},t-R_2/c)\right\rangle\frac{\exp\left[i\omega(R_2-R_1)/c\right]}{R_1R_2}. \tag{3.1.35}$$

As the delay $\tau=(R_2-R_1)/c$ is small according to the assumption that the transverse dimensions of the source are small, we have

$$\Gamma(\vec{r}_1,\vec{r}_2,0)=\sum_n I_n(\vec{s})\frac{\exp\left[i\omega(R_2-R_1)/c\right]}{R_1R_2}, \tag{3.1.36}$$

where $I_n(\vec{s})=\left\langle a_n(\vec{s},t)a_n^*(\vec{s},t)\right\rangle$ – intensity of the atom-emitted filed. Substituting the integral for the sum over *n* in (1.36), we obtain

$$\tilde{A}(\vec{r}_1,\vec{r}_2,0)=\int_{\Delta S} I(\vec{s})\frac{\exp\left[ik(R_2-R_1)\right]}{R_1R_2}dS, \tag{3.1.37}$$

where $k=\omega/c$ – wave number, $I(\vec{s})$ – spatial distribution of the light intensity within the source.

Formula (1.37) represents the Van Cittert – Zernike theorem by which the radiation cross-correlation function is a Fourier transform of the function describing the source intensity.

The complex coherence degree $|\gamma(\vec{r}_1,\vec{r}_2,0)|$ is equal to (see (3.1.24))

$$\gamma\left(\vec{r}_1,\vec{r}_2,0\right)=\frac{1}{\sqrt{I\left(P_1\right)}\sqrt{I\left(P_2\right)}}\int_{\Delta S} I\left(\vec{s}\right)\frac{\exp\left[ik\left(R_2-R_1\right)\right]}{R_1R_2}dS\,, \qquad (3.1.38)$$

where $I\left(P_1\right)=\int_{\Delta S}\left[I\left(\vec{s}\right)/R_1^{\,2}\right]dS$, $I\left(P_2\right)=\int_{\Delta S}\left[I\left(\vec{s}\right)/R_2^{\,2}\right]dS$ – intensities at the points $P\left(\vec{r}_1\right)$ and $P\left(\vec{r}_2\right)$. Expressing the distances R_1 and R_2 in terms of the relations

$$R_1\approx L\left(1+\frac{\left(x_1-\xi\right)^2+\left(y_1-\eta\right)^2}{2L}\right), \qquad (3.1.39)$$

$$R_2\approx L\left(1+\frac{\left(x_2-\xi\right)^2+\left(y_2-\eta\right)^2}{2L}\right), \qquad (3.1.40)$$

and introducing the angular coordinates $p=\left(x_1-x_2\right)/L$ and $q=\left(y_1-y_2\right)/L$, we can transform the expression of (1.38) as follows:

$$\gamma\left(\vec{r}_1,\vec{r}_2,0\right)=\exp\left(i\psi\right)\frac{\int_{\Delta S} I\left(\xi,\eta\right)\exp\left[-ik\left(p\xi+q\eta\right)\right]d\xi d\eta}{\int_{\Delta S} I\left(\xi,\eta\right)d\xi d\eta}, \qquad (3.1.41)$$

where $\psi=k\left[\left(x_1^{\,2}+y_1^{\,2}\right)-\left(x_2^{\,2}+y_2^{\,2}\right)\right]/L$; ($\xi$, η) and (x, y) – positions of the points within the planes S and, respectively. As follows from (3.1.41), $\left|\gamma\right|$ equals the absolute value of the normalized Fourier transform of the function describing the source intensity.

For a source in the form of a circle with the radius ρ and with a homogeneous brightness distribution, when the following conditions are met

$$\begin{aligned} &I\left(\xi,\eta\right)=1; \quad \left(\xi,\eta\right)\le\rho,\\ &I\left(\xi,\eta\right)=0; \quad \left(\xi,\eta\right)>\rho, \end{aligned} \qquad (3.1.42)$$

we get

$$\left|\gamma\left(\vec{r}_1,\vec{r}_2,0\right)\right|=\frac{2J_1\left(v\right)}{v}, \qquad (3.1.43)$$

where $J_1\left(v\right)$ – Bessel function of the 1st kind and $v=k\rho\left(p^2+q^2\right)^{1/2}$. For v = 0, the modulus spatial coherence complex degree $\left|\gamma\right|=1$, whereas at v = 3.83 it is equal to 0. The distance r between the points $P\left(\vec{r}_1\right)$ and $P\left(\vec{r}_2\right)$ that is associated with complete incoherence of radiation is determined as follows:

$$r = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} = \frac{0,61\lambda L}{\rho}. \qquad (3.1.44)$$

The distance

$$r_0 = \frac{0,08\lambda L}{\rho}, \qquad (3.1.45)$$

for which the function $|\gamma| = 2J_1(v)/v = 0,88$ (v = 1), is taken as a radius of the coherence area. From (3.1.45) it follows that the quantity r_0 is determined by the average radiation wave length λ and by the angular dimensions of the source $\alpha = \rho / L$ observed within the plane of a screen *D*. As it has been noted above, a half-width of the modulus spatial coherence degree measured at the half-height is practically used as a measure of the coherence radius. So, by the construction of the visibility function $V(|\vec{r}_1 - \vec{r}|) = |\gamma(\vec{r}_1, \vec{r}_2, 0)|$ within the plane *D* and measuring its half-width, we can estimate the radiation coherence radius of a source. Another characteristic of the spatial coherence is the coherence factor $C = r_0 / a_0$ (a_0 - half-width of a light beam in the observation plane) that is invariant (constant) relative to variations in a position of the plane, where the spatial coherence of a light beam is measured.

3.1.4. Schemes for measurement of the radiation spatial-coherence parameters

The above-mentioned scheme, used for analysis of the radiation spatial coherence with the help of a Young interferometer, is a classical scheme to measure a transverse radius of the radiation coherence. However, it is difficult to construct the modulus spatial-coherence degree according to this scheme because manufacturing of small-size holes of a regular circular form is a very laborious task. At the present time one can use other methods to measure the transverse coherence radius: slit diffraction method, polarization method, holographic method, modified Michelson interferometer method. All these methods are less laborious than the Young method and their spatial resolution is higher (minimal size of the measured radiation spatial-coherence area). The modified Michelson interferometer method enables one to construct the spatial-coherence function in a single stage in the form of the fringe visibility spatial modulation. Modification of a Michelson interferometer is realized by substitution of the total internal reflection (TIR) prism P for one of its mirrors (Fig. 3.1.5).

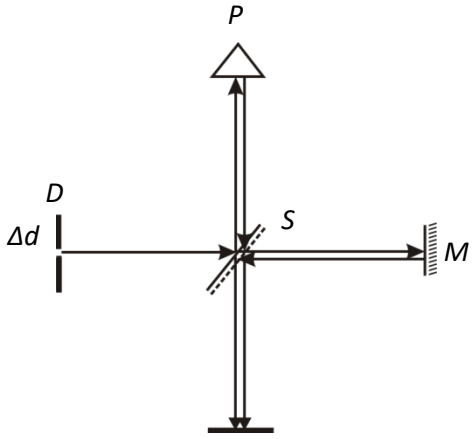


Fig. 3.1.5 – Optical scheme of modified Michelson interferometer

Let us consider the principle of estimation of the modulus radiation spatial coherence by means of such an interferometer. We estimate the spatial coherence of a light field, emitted by a source of rather small spatial dimensions, ΔS within the detector plane B. The total amplitude of the radiation field $E(\vec{r})$ reflected from the mirror M and subjected to the total internal reflection from the prism P is determined at the point $\vec{r}$ of the detector plane B as follows:

$$E(\vec{r},t)=\frac{\exp(ik\vec{n}_1\cdot\vec{r})}{i\lambda L}\int\limits_{\Delta S}A(\xi,\eta,t)\exp\left[ik(\xi x+\eta y)\right]d\xi d\eta+$$
$$+\frac{\exp(ik\vec{n}_2\cdot\vec{r})}{i\lambda L}\int\limits_{\Delta S}A(-\xi,\eta,t)\exp\left[ik(-\xi x+\eta y)\right]d\xi d\eta, \qquad (3.1.46)$$

where $A(\xi,\eta,t)$ – amplitude of the source field subjected to reflection from mirror 5, $A(-\xi,\eta,t)$ – amplitude of the source field subjected to the wavefront rotation by 180^0 about the coordinate axes at total internal reflection from prism 6, (ξ,η) – Cartesian coordinates in the source plane, $k=2\pi/\lambda$ – wave number, $\vec{n}_1$ and $\vec{n}_2$ – normal vectors of the wave fronts formed within the source plane at the point $\vec{r}$. The intensity distribution $I(\vec{r})=\left\langle\left|E(\vec{r},t)\right|^2\right\rangle$ in recoding plane 9 is given by

$$I(\vec{r}) = -\frac{1}{\lambda^2 L^2}\iint\limits_{\Delta S}\left\langle A(\xi_1,\eta_1,t)\,A^*(\xi_2,\eta_2 t)\right\rangle \exp\left(ik\left[(\xi_1-\xi_2)x+(\eta_1-\eta_2)y\right]\right)d\xi_1 d\eta_1 d\xi_2 d\eta_2 -$$

$$-\frac{1}{\lambda^2 L^2}\iint\limits_{\Delta S}\left\langle A(-\xi_1,\eta_1,t)\,A^*(-\xi_2,\eta_2,t)\right\rangle \exp\left(ik\left[(\xi_2-\xi_1)x+(\eta_1-\eta_2)y\right]\right)d\xi_1 d\eta_1 d\xi_2 d\eta_2 -$$

$$-\frac{2}{\lambda^2 L^2}\,\mathrm{Re}\left\{\exp\left[ik(\vec{n}_1-\vec{n}_2)\cdot\vec{r}\right]\times\right.$$

$$\left.\times\iint\limits_{\Delta S}\left\langle A(\xi_1,\eta_1,t)\,A^*(-\xi_2,\eta_2,t)\right\rangle \exp\left(ik\left[(\xi_1+\xi_2)x+(\eta_1-\eta_2)y\right]\right)d\xi_1 d\eta_1 d\xi_2 d\eta_2\right\}.$$

(3.1.47)

We assume that the source field amplitude is of a spatial symmetry, i.e. we have $A(\xi,\eta,t) = A(-\xi,\eta,t)$, and is delta-correlated with respect to the space and we have $\left\langle A(\xi_1,\eta_1,t)\,A^*(\xi_2,\eta_2,t)\right\rangle = I(\xi_1,\eta_1)\,\delta(\xi_1-\xi_2)\,\delta(\eta_1-\eta_2)$. Then expression (1.47) for $I(\vec{r})$ is transformed as

$$I(\vec{r}) = -\frac{1}{\lambda^2 L^2}\left\{2\iint\limits_{\Delta S} I(\xi,\eta)\,d\xi d\eta + \cos\left(k(\vec{n}_1-\vec{n}_2)\cdot\vec{r}\right)\iint\limits_{\Delta S} I(\xi,\eta)\cos(2ik\xi x)\,d\xi d\eta\right\}.$$

(3.1.48)

The first term in expression (1.48) describes a uniform background within the source plane, whereas the second – interference field described by the factor $\cos\left(k(\vec{n}_1-\vec{n}_2)\cdot\vec{r}\right)$, that is modulated for one of the coordinates by a Fourier spectrum of the source intensity distribution. According to the above-mentioned Van Cittert–Zernicke theorem, the Fourier transform of the intensity distribution function for the source is equal to the mutual spatial coherence function of the source field in the observation plane. In this way the interference field is actually modulated by the spatial coherence function in one of its directions. Varying the prism edge position, we can construct spatial coherence functions of the field in other directions. In the process the directions of fringes in the interference pattern should be varied by rotation of the mirror M so that the fringes be parallel to the prism edge.

The observed formation of an envelope of interference fringes in the form of the spatial coherence function of a light field has the following physical meaning. The wave fronts of the beams reflected by a flat mirror and a corner reflector are mirror reflections of each other. On precise alignment of their centers, the fields in central regions of the wave fronts are varying practically in phase, forming in the plane *P* the high-contrast interference patterns little differing in their positions. The fields of the wavefront regions remote from the centers are

phase shifted and create in the plane P the interference patterns shifted relative each other. When interference patterns from all the points of the wave fronts are summed up, the resultant interference pattern is formed in the plane P with spatially inhomogeneous contrast. The contrast of fringes is maximal at the center of the interference pattern, lowering to the edges. A distance from the center of the interference pattern to the point, where the contrast of fringes is only a half of the maximal value, is assumed equal to the radiation coherence radius in the observation plane. To estimate the radiation coherence radius in the source plane, we should take into account the scaling factor for light fields in the given measuring scheme. It is a ratio of the light field diameter in the observation plane to the light field diameter at the output of a laser. To find the coherence radius, we record an interference pattern formed in the plane P using an array of CCD detectors and visualize it by a display. Then we measure maximal and minimal ($I_{\max}$ and $I_{\min}$) intensities in every fringe of the interference pattern, beginning from its center to the edge, and the visibility of fringes V is determined by

$$V(\Delta\rho) = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}, \qquad (3.1.49)$$

where $\Delta\rho$ – distance from the pattern center to the corresponding bright fringe. Then we construct a curve for V as a function of $\Delta\rho$ to determine the spatial coherence radius ρ as a width of the constructed function at the half-height.

\

3.2. Hologram types: thin and volume, amplitude and phase, reflection and transmission

Hologram is an amplitude or phase «*imprint*» of the interference pattern formed on the interaction of coherent object and reference light beams. Let us consider a structure of the interference pattern created by two point sources of coherent light O_1 and O_2. Fig. 3.2.1 shows section of an interference field by the drawing plane.

Within the space surrounding the sources, a system of stationary (standing) waves is formed.

The condition for antinodal surfaces (maxima of the interference field intensity) is given by the following expression:

$$r_1 - r_2 = k\lambda, \qquad (3.2.1)$$

where $k = 0, \pm 1, \pm 2, \ldots$

The condition for nodal surfaces (minima of the interference field intensity) is determined as

$$r_1 - r_2 = \frac{2k+1}{2}\lambda, \qquad (3.2.2)$$

where $k = 0, \pm 1, \pm 2, \ldots$

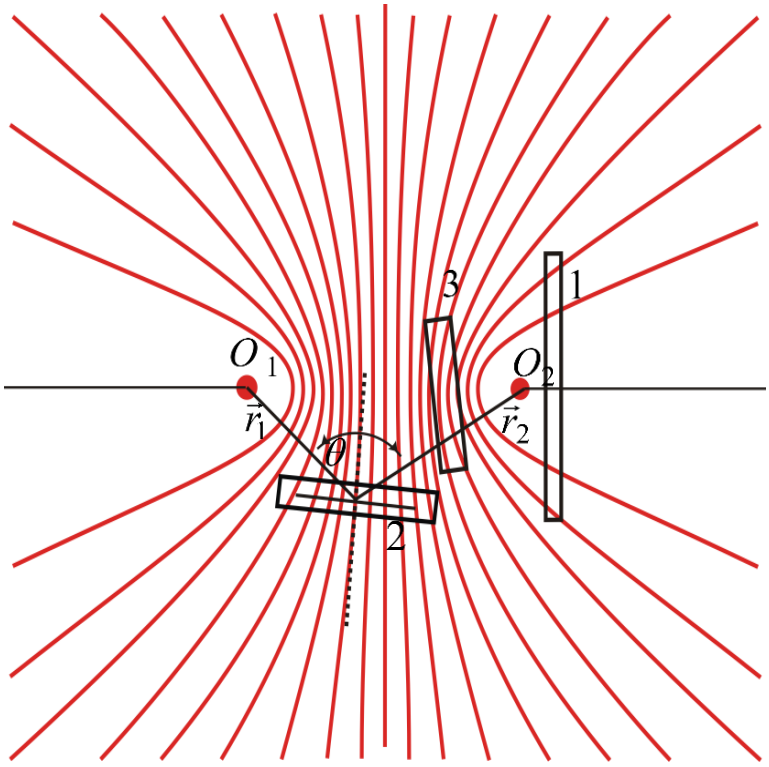


Figure 3.2.1 – Transverse section of the interference pattern formed by two coherent point sources and the geometry of the holographic plate position corresponding to three principal schemes for hologram recording.

The expressions of (3.2.1-3.2.2) are equations for a system of hyperboloids with the rotation axis O_1O_2. The planes, which are tangential to the nodal and antinodal surfaces at each point of the space, bisect the angle θ between the vectors $\vec{r}_1$ and $\vec{r}_2$. So, the interference fringes are directed along the bisector of the angle between the vectors of the interacting coherent light beams.

Also, Fig. 3.2.1 demonstrates a geometry of the holographic plate position corresponding to the three principal schemes for hologram recording. When a holographic plate is in position 1, light from both sources is directed almost

collinearly to record an axial hologram according to the Gabor scheme. When the plate is in position 2, a hologram is recorded according to the Leith-Upatnieks scheme, and in position 3 – according to the Denisyuk scheme.

An interference pattern may be recorded by means of variations in the refractive index or in the thickness of a photosensitive medium, or else by changing its absorption due to the effect of the light interference field. According to the way of the interference field recording, we distinguish ***phase*** and ***amplitude holograms.*** Radiation diffraction from the amplitude hologram is realized due to modulation of the absorption factor of a photosensitive layer. Radiation diffraction from the phase hologram is realized due to modulation of the refractive index or of the holographic medium thickness.

Phase and amplitude holograms are in turn subdivided into thin and volume holograms. Recording of a ***thin (planar, 2D)*** or ***thick (volume, 3D) hologram*** is determined by the recording geometry of a holographic image, by the wavelength of hologram recording light, and by the thickness of a photosensitive layer. Identification of a hologram as ***thin, planar or 2D*** is indicative of the fact that we deal with the surface two-dimensional diffraction grating, whereas in the case of ***thick, volume or 3D holograms*** the diffraction grating is three-dimensional. A hologram is attributed to the thin or volume type according to the relationship between the thickness of a holographic emulsion and the interference pattern (diffraction grating) period. When the photoemulsion thickness d is considerably lower than the interference structure (diffraction grating) period Λ, $d << \Lambda$, a thin (2D) hologram is recorded. When $d << \Lambda$, a volume (3D) hologram is recorded.

In thin holograms a thickness of photoemulsion offers recording of the interference pattern (diffraction grating) period only. In volume holograms it is possible to record not only the period but also the direction of interference fringes within the volume of a holographic medium. Just this feature is decisive for the formation of absolutely different diffraction patterns on thin and volume diffraction gratings (holograms).

A criterion enabling precise attribution of a hologram to the thin or volume type is the ***Klein parameter***

$$Q = \frac{2\pi\lambda d}{n\Lambda^2}, \qquad (3.2.3)$$

where λ – wavelength of hologram recording light; d – thickness of a photosensitive layer; n – refractive index of a photosensitive medium; Λ – period of the recorded interference structure.

It is assumed that in the case, when $Q << 1$, a ***hologram is thin***, when $Q >> 10$ — a ***hologram is* volume**. In the intermediate cases, when the Klein parameter falls within the interval $1<Q<10$, a hologram is attributed to neither planar nor volume type. Diffraction from such structures has no exact analytical description.

Let us find a period of the recorded interference structure. Assume that at a photosensitive layer two plane coherent waves 1 and 2 converge at the normal angles to the photolayer surface φ_1 and φ_2 , respectively, as shown in Fig. 3.2.2.

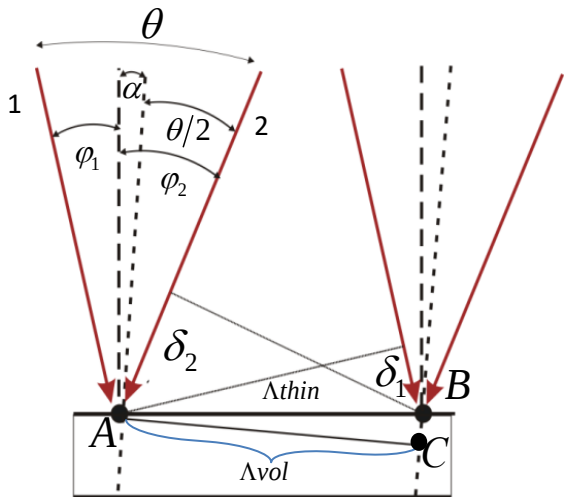


Figure 3.2.2 – To calculate the holographic grating period a

When a thin hologram is recorded, the interference pattern is formed on the surface of a photoplate.

If the points A and B are associated with positions of two adjacent maxima, the path difference of the interfering beams δ_1 and δ_2, on going from the point A to the point B, is changed by λ. So, we can write the following condition:

$$\delta_1 + \delta_2 = \lambda. \tag{3.2.4}$$

As $\delta_1 = AB \sin\varphi_1$ and $\delta_2 = AB \sin\varphi_2$, the interference pattern (hologram) period in the case of a thin hologram is given by the expression

$$\Lambda thin = \frac{\lambda}{\sin\varphi_1 + \sin\varphi_2}. \qquad (3.2.5)$$

When the beams are incident symmetrically, the expression of (3.2.5) takes the form:

$$\Lambda_{thin} = \frac{\lambda}{2\sin\theta/2} \qquad (3.2.6)$$

where θ – angle between the interfering waves.

In the case of volume holograms not only the period but also the direction of interference fringes within a holographic medium is recorded. In other words, the medium is fixing vector ***K*** of the diffraction grating. In this case a distance between two adjacent maxima is determined by the AC length – to find a position of the point C, we should drop a perpendicular from the point A to the line that is a bisector direction of the angle θ. The interference-pattern period for a volume hologram is given by:

$$\Lambda_{vol} = \Lambda_{thin}\cos\alpha \qquad (3.2.7)$$

Considering that $\alpha = \varphi_1/2 - \varphi_2/2$ and finding the period of a thin hologram from the equality of (3.2.5), we can obtain an expression for the volume hologram period as follows:

$$\Lambda vol = \frac{\lambda}{2\sin\theta/2}, \qquad (3.2.8)$$

where θ – convergence angle of the hologram-recording light beams; θ/2 gives an angle between wave vectors of the hologram recording light and the direction of interference fringes (see Fig.3.2.2).

Diffraction from thin and volume holograms is differing qualitatively.

Diffraction from thin (2D) holograms is described by the Raman-Nath diffraction. The direction of the principal diffraction maxima is given by the formula for a thin diffraction grating as follows:

$$\Lambda\left(\sin\alpha_{in}+\sin\alpha_d\right)=\pm k\lambda, \tag{3.2.9}$$

where Λ – interference pattern period determined by (3.2.5), λ – wave length of light, α_{in} – angle of light incidence on the grating, α_d – diffraction angle, k – diffraction order (k=0,±1, ±2,…). When an electromagnetic wave is interacting with such a hologram (on condition that recording is linear), two diffracted waves are formed (*k*=+1 and *k*= -1): +1st and -1st diffraction orders and forward transmitted light (*k*=0). Just this is associated with reconstruction of two images: virtual image (+1 diffraction order) reconstructed at the place, where an object was positioned in the process of hologram recording, provided the hologram is reconstructed by the reference wave, and real image (-1diffraction order), see Fig.3.0.1,b and Fig.3.2,b. As a rule, a real image is positioned from the other side of a hologram. A virtual image may be observed directly or recorded by means of an optical objective. A real image is recorded when we position a photoplate or a photodetector at the image localization place and then its 2D projection is obtained.

We can demonstrate that an elementary thin hologram recorded by two coherent plane waves incident at the angles φ_1 and φ_2 on a photosensitive layer is a diffraction grating. When such a hologram is illuminated by one of the waves involved in its recording, it reconstructs the wave that was absent on reading. Using the formula for a thin grating (3.2.9), we can find the first-order diffraction angle, with regard to the formula for a period of the interference pattern formed by two coherent plane waves (3.2.5), as follows:

$$\sin\alpha_d=\sin\varphi_1+\sin\varphi_2-\sin\alpha_{in}. \tag{3.2.10}$$

The recorded interference pattern of two plane waves in the form of modulation of the transmission factor is an elementary amplitude hologram. To reconstruct the recorded hologram, we use one of the beams involved in its recording. For example, let us select beam 1 that is incident on the photoplate at the angle φ_1 (Fig. 3.2.2). As follows from (3.2.10), the diffraction angle is $\alpha_d=\varphi_2$, i.e., the diffracted light goes at the incidence angle of the second wave involved in the hologram recording process φ_2. When the hologram is illuminated by the wave incident at the angle φ_2, we can reconstruct the beam that during hologram recording was incident on the photoplate at the angle φ_1. Any complex

wave front may be represented as a superposition of plane waves with different direction cosines, and the result may be generalized for a random object wave. Proceeding form the above, we can conclude that the reference and object waves are reciprocal and hologram is a diffraction grating reconstructing the waves involved in its recording. This is demonstrated by the ***basic relation for a thin hologram.*** First, we describe theoretically the hologram recording process. The complex amplitude of an object wave within the hologram plane is given as

$$П = П\ (x,y)e^{i\varphi_n(x,y)}, \tag{3.2.11}$$

where $П = П\ (x,y)$ - amplitude factor of the wave front coming from the object within the plane of a holographic plate, $e^{i\varphi_n(x,y)}$ – phase factor.

The complex amplitude of a plane reference wave within the hologram plane is denoted as

$$O = O_0 e^{i\varphi_o}. \tag{3.2.12}$$

Interference of the reference and object waves results in the total intensity distribution of the wave field in the interference pattern that is found from the expression

$$I = |П + O|^2 = П^2(x,y) + {O_0}^2 + П(x,y)O_0 e^{i(\varphi_П - \varphi_o)} + П(x,y)O_0 e^{-i(\varphi_П - \varphi_o)}. \tag{3.2.13}$$

Expression (3.2.13) gives all the information about the amplitude and phase of a reference wave falling onto a holographic plate.

In the case when a thin amplitude hologram is linearly recorded, the amplitude transmission of a photosensitive medium t is linearly dependent on the light exposure (intensity)

$$t = t_0 - \gamma I, \tag{3.2.14}$$

where t_0 – initial transmission of a photosensitive medium, I – light intensity, γ – proportionality factor.

Substituting the expression for the wave field intensity in the interference pattern (3.2.13) into (3.2.14), we can determine the amplitude transmission for a thin hologram as:

$$t = t_0 - \gamma\left(П^2(x,y) + {O_0}^2 + П^*(x,y)O_0 e^{i(\varphi_n - \varphi_o)} + П(x,y)O_0 e^{-i(\varphi_n - \varphi_o)}\right). \tag{3.2.15}$$

Now let us describe the hologram reconstruction process. The complex amplitude of a reconstructing plane wave within the hologram plane is given by

$$B = B_0 e^{i\varphi_в}, \qquad (3.2.16)$$

where B_0 and $\varphi_в$ - amplitude and phase of the hologram reconstructing wave in the plane of a holographic plate, respectively.

The light field beyond the hologram E_H is found by multiplication of the hologram amplitude transmission (3.2.15) into the complex amplitude of a reconstructing wave (3.2.16) as follows:

$$E_\Gamma = Bt_\Gamma = B_0 e^{i\varphi_в}\left[t_0 - \gamma\left(П^2(x,y) + O^2{}_0\right)\right] - \gamma B_0 П(x,y) O_0 e^{i(\varphi_П - \varphi_О + \varphi_В)} - \\ - \gamma B_0 П_0 O_0 e^{i(\varphi_О - \varphi_П + \varphi_в)}. \qquad (3.2.17)$$

The relation of (3.2.17) is termed the ***principal relation of a thin hologram***. Let us analyze the derived expression. A field on the other side of the hologram, at the interaction of a reconstructing wave having a plane wave front $B = B_0 e^{i\varphi_в}$ with a thin amplitude hologram, is determined by three components. The first component describes a plane wave that retains the direction and form of a reconstructing wave because the phase factor $e^{i\varphi_в}$ involved is determined only by the reconstructing-wave phase. The component $П^2(x,y)$ makes it inhomogeneous. This is the so-called zeroth diffraction order. The phase factor of the second component involves the object-wave phase $e^{i\varphi_п}$. The second component describes a divergent wave propagating in the direction of the +1-order diffraction that is responsible for reconstruction of the virtual image (object). The phase factor of the third component involves $-\varphi_п$. Assuming that a complex amplitude of the hologram reconstructing wave is described by (3.2.12) (i.e., reconstruction is effected by the reference wave), the phase factor of the second component is determined by the object wave phase. To within the amplitude factor, we can reconstruct an image of the object in the position, where it was located on hologram recording. This means that the second component

describes a complex-conjugate wave with respect to the object wave that is converging and propagating in the opposite direction (in the direction of the first-order diffraction), and corresponds to the real image.

Fig. 3.2.3 Fig. 3.2.3This is illustrated by Fig. 3.2.3 that shows recording and reconstruction of a thin hologram of the point source according to the Leith-Upatnieks scheme. It is seen that, on reconstruction of a hologram by the reference wave, a virtual image of the point source is formed in beams of the +1-order diffraction (divergent spherical wave) and «hits» observer's eye that is positioned on the other side of the holographic plate.

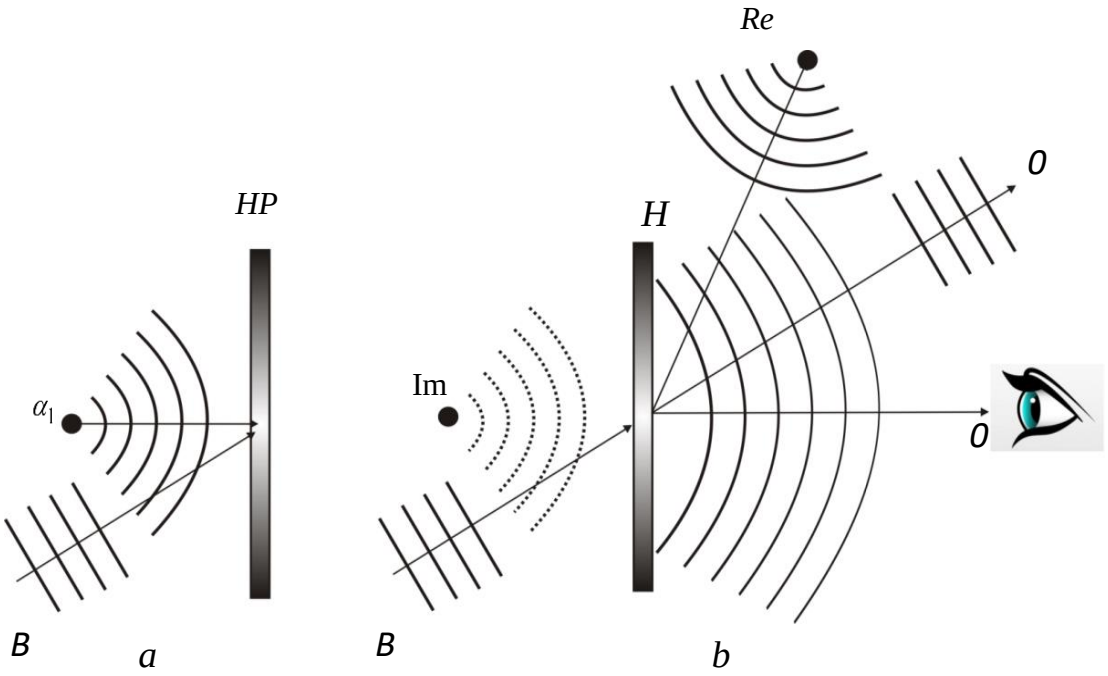


Figure 3.2.3 - Recording (a) and restoration (b) of a thin hologram of a point source according to the Leith-Upatniex scheme

Considering the physical principles which form the basis for recording and reconstruction of holograms, it may be inferred: ***provided that on the light-sensitive surface the interference structure is formed by an arbitrary object wave and by a reference (plane or spherical) wave coherent to it and then illuminated by the reference wave, the light diffraction (+1 diffraction order) results in reconstruction of a virtual image in the extended divergent beams, whereas convergent beams form a real image of the object (-1 diffraction order). Forward transmitted light governs the zeroth order diffraction.***

Reconstruction of two images may be explained in terms of the Huygens-Fresnel principle. According to this principle, the wave front propagation is

described as propagation of spherical waves from each point of the space. Perturbation goes both forwards and backwards: the perturbation that propagates backwards is damped by the antiphase wave field coming from the object. Because of this, the wave front from the object is propagating only in the forward direction. There is no object per se on reconstruction of a hologram and hence the light wave propagates both forwards and backwards, forming real and virtual images.

A virtual image is ***orthoscopic***, i.e., the phase distribution over the image surface corresponds to the phase distribution over the surface of the object itself.

A real image is ***pseudiscopic***, i.e., the surface phase distributions of the image and of the object are identical in their absolute values but have different signs. In this case the observer sees an «inverted» image of the object. A similar effect may be observed with the use of stereophotography: hills become cavities and vice versa.

A real image is directly observed on illumination of the hologram with a wave that is conjugate to the reference one. The waves are ***conjugate*** when their propagation is antiparallel and their amplitudes are complex conjugates in any plane. Instead of the reference wave, we substitute into (3.2.17) the wave that is conjugate to the reference one $B = O_0 e^{-i\varphi_0}$. Then a field on the other side of the hologram is given in the form

$$E_\Gamma = Bt_\Gamma = O_0 e^{-i\varphi_0}\left[t_0 - \gamma\left(\left|\Pi(x,y)\right|^2 + \left|O_0\right|^2\right)\right] - \gamma {O_0}^2 \Pi(x,y) e^{-i(\varphi_n)} - \\ - \gamma {O_0}^2 e^{2i\varphi_0} \Pi_0(x,y) e^{i\varphi_n}. \qquad (3.2.18)$$

In (2.18) the second term, to within an amplitude factor, describes the wave conjugate to the object wave diverging from the object in the process of hologram recording. The reconstructed wave, that is conjugate to the object wave, is converging to a real image of the object. As this wave is conjugate to the object wave, the real image directly observed by eye is orthoscopic (see Fig.3.2.4).

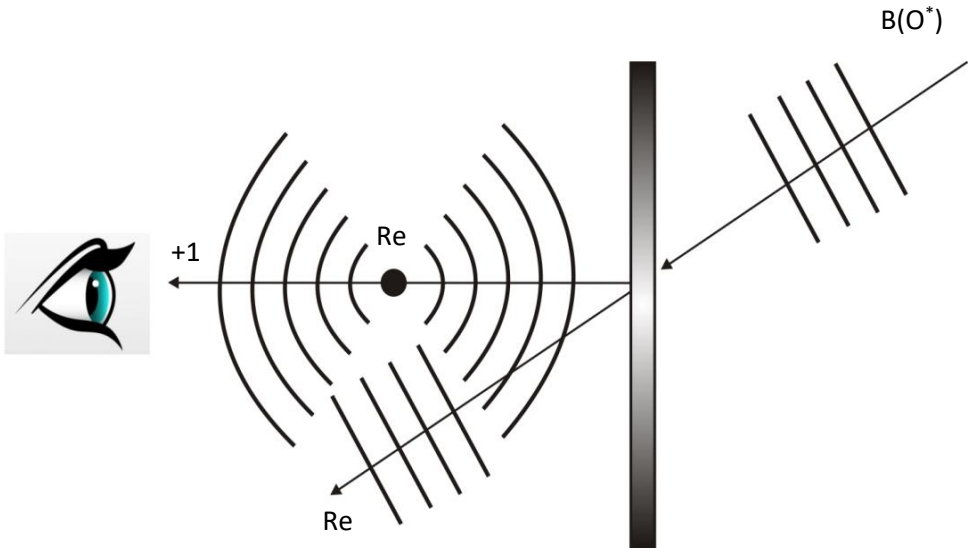


Figure 3.2.4 - Reconstruction of a real pseudoscopic image

Appearance of the two images (virtual and real) on reconstruction of a thin hologram is caused by ambiguity of its recording. As it has been already noted, in thin holograms a thickness of photoemulsion allows for recording of only the period of the interference pattern (diffraction grating) and there is no information about the direction of interference fringes.

Fig.3.2.5 (a,b) shows two symmetric geometries for hologram recording. In both schemes the diffraction structures are formed in a photosensitive medium with identical periods but with different directions of the slope of interference fringes. When we record a thin hologram according to the scheme shown in Fig. 3.2.5 and reconstruct it by beam 1, a field on the other side of the hologram will be represented by three beams: zeroth diffraction order, +1 diffraction order for reconstruction of beam 2 (virtual image), and -1 diffraction order for reconstruction of a real image. Origination of the -1 order diffraction is due to ambiguity in recording of a thin hologram, where the direction of interference fringes is not recorded. In this way, interfering with a thin diffraction grating, light «has no information» about the slope direction of interference fringes and is diffracted «just in case» at the angle α in the direction of beam 2 and at the angle $-\alpha$ in the direction of beam $2'$ (Fig.3.2.5,c).

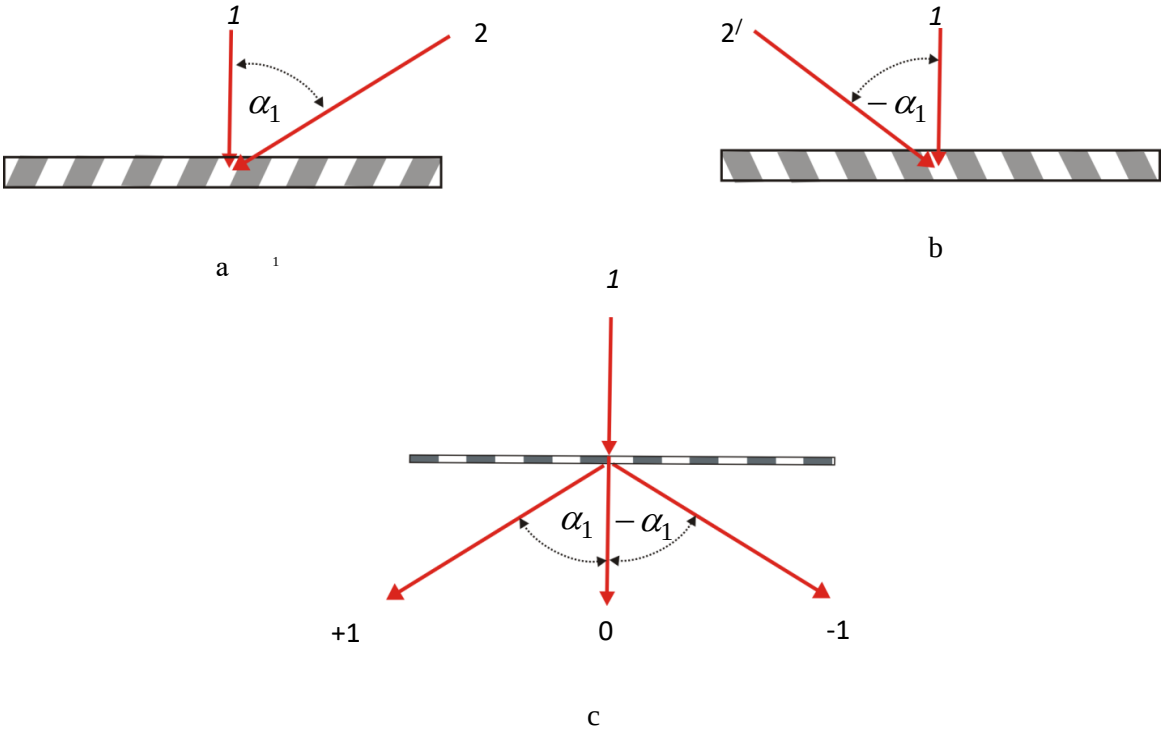


Figure 3.2.5 – Appearance of the virtual and real images on reconstruction hologram

Diffraction from 3D hologram is associated with volume diffraction structures (Bragg diffraction) and is meeting the Wulf-Bragg condition that gives the direction of principal maxima as follows:

$$2\Lambda \sin\alpha_d = k\lambda. \qquad (2.19)$$

When an electromagnetic wave interacts with a volume hologram recorded in the linear mode, only a single diffraction order is formed, $k = +1$, that is responsible for reconstruction of the object wave (virtual image). Changing orientation of the hologram with respect to the reference beam, we can realize reconstruction of the real (pseudoscopic) image only.

Higher orders of diffraction are associated with the nonlinear hologram-recording mode (nonlinear character of the photoresponse of a medium to light exposure) and they lead to noise on reconstruction of both thin and volume holograms.

Holograms may be of the ***transmission*** and ***reflection types***.

Different orientations of a recording medium with respect to recording beams make it possible to record transmission or reflection holograms. When the beams are incident on a photosensitive layer at the same side (Fig.3.2.6,a), recording is realized in concurrent beams.

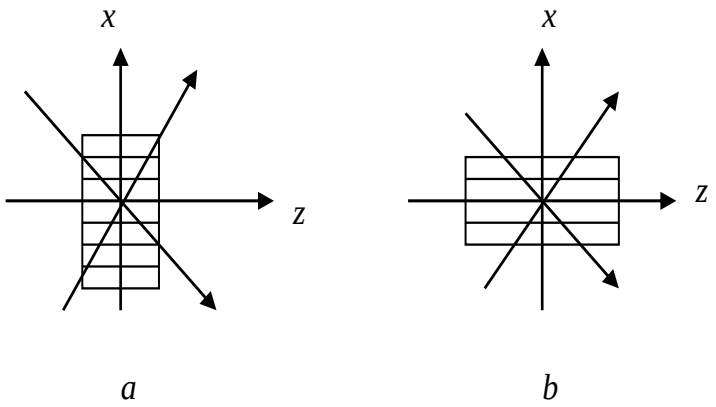


Figure 3.2.6 – Symmetric schemes for hologram recording in concurrent (*a*) and colliding (*b*) beams

As this takes place, orientation of the interference patterns is close to the perpendicular erected with respect to the layer plane (Fig. 3.2.6, *a*). When the beams are incident from different sides (counter-propagation, Fig. 3.2.6, *b*), the interference surfaces are mainly oriented within the layer plane.

Returning to Fig. 3.2.1, we can see that Gabor and Leith-Upatnieks holograms belong to the transmission type, whereas Denisyuk holograms belong to the reflection type.

The difference in reflection and transmission holograms on image reconstruction is that light reconstructing the object wave is either reflected by a hologram (reflection hologram) or transmitted through a hologram (transmission hologram).

3.3. Diffraction efficiency

As noted above, holographic images are reconstructed in diffracted beams. Because of this, the primary characteristic of a hologram that governs the reconstructed image brightness is the ***diffraction efficiency***. The diffraction efficiency η is defined as a ratio of the radiation intensity in the diffracted wave of a given diffraction order (I_D) to the intensity of radiation incident on a hologram in the process of its reading (*Iin*)

$$\eta = \frac{I_d}{I_{in}} \times 100\%. \qquad (3.3.1)$$

Thus, the diffraction efficiency of a hologram is determined by the energy fraction of incident light that is spent on reconstruction of the object beam. Table 3.3.1 lists maximal diffraction efficiencies for different types of holograms.

Table 3.3.1

Amplitude planar (%)	Phase planar (%)	Volume transmission amplitude (%)	Volume reflection amplitude (%)	Volume transmission phase (%)	Volume reflection phase (%)
6.25	3.39	3.7	7.2	100	100

As seen from this table, the diffraction efficiency of thin (planar, 2D) holograms is comparatively low. The diffraction efficiency of a planar amplitude hologram with a sinusoidal profile is at maximum ($\eta_{\pm 1} = 6{,}25\%$) in the +1 and -1 diffraction orders, the efficiency of forward transmitted light being $\eta_0 = 25\%$. The remaining light (72%) is absorbed by the grating. If the grating profile is rectangular (transmission of a half of the grating period is zero and that of the other half is 1), the diffraction efficiency in ±1 order of diffraction comes to

$\eta_{\pm 1} = 10{,}1\%$, while the efficiency of forward transmitted light comes to $\eta_0 = 25\%$.

Note that deviation of the amplitude transmission profile from the sinusoidal form (even insignificant departure of the amplitude transmission function from the sine is exhibited as unevenness of the profile) results in the appearance of higher-order diffraction peaks. They are due to the fact that a grating with a nonsinusoidal profile (rectangular, triangular, etc.) may be represented as a set of sinusoidal diffraction gratings, each of the period $\Lambda_k = \pm\frac{\Lambda}{k}$. The light diffraction angle on normal incidence onto such a grating is different for each of the gratings

$$\varphi_{d_k} = ar\sin\left(\pm\frac{k\lambda}{\Lambda}\right), \qquad (3.3.2)$$

and the field scattered by such a grating represents the totality of light diffracted from k sinusoidal gratings.

The diffraction efficiency of a thin phase hologram having a sinusoidal phase-transmission profile is described by the k-th order Bessel function of the 1st kind. In the case of the zeroth and ±1 diffraction order we have

$$\begin{aligned}\eta_0 &= J_0^2\left(\frac{2\pi}{\lambda}\Delta n d\right)\\ \eta_{+1} &= J_1^2\left(\frac{2\pi}{\lambda}\Delta n d\right),\\ \eta_{-1} &= J_{-1}^2\left(\frac{2\pi}{\lambda}\Delta n d\right)\end{aligned} \qquad (3.3.3)$$

where Δn - modulation amplitude of the refractive index, d – grating thickness.

To illustrate, a maximal efficiency of a thin amplitude hologram with a sine phase profile comes to $\eta_{\pm 1} = 33{,}9\%$ for the Bessel function argument from formula (3.3.3) that is equal to 0.582. An increase in the amplitude modulation Δn or in the recording medium thickness d leads to the appearance of higher-order peaks and to lowering of the ±1 order diffraction efficiencies. Note that just the beams diffracted into ±1 orders are responsible for reconstruction of virtual and real images. The higher diffraction orders result in phantom images in the form of noise blooming around the real images. The diffraction efficiency of a thin phase grating with the rectangular profile is somewhat higher: $\eta_{\pm 1} = 40{,}1\%$.

Theoretically, the diffraction efficiency of volume phase holograms may reach its limiting value $\eta_1 = 100\%$.

The diffraction efficiency of a volume transmission phase hologram $\eta^t{}_{+1}$ is described as follows:

$$\eta^t{}_{+1} = \sin^2\left(\frac{\pi}{\lambda}\frac{\Delta n d}{\cos\frac{\theta}{2}}\right). \qquad (3.3.4)$$

The diffraction efficiency of a volume reflection phase hologram $\eta^t{}_{+1}$ is given by

$$\eta^r{}_{+1} = th^2\left(\frac{\pi}{\lambda}\frac{\Delta n d}{\cos\frac{\theta}{2}}\right), \qquad (3.3.5)$$

where θ - angle between the object and reference waves.

These formulae are valid for a symmetric geometry of the hologram recording normal to the holographic plate surface.

3.4. Spectral and angular selectivity

Hologram selectivity sets the hologram reading conditions for which the diffraction efficiency of radiation reconstructing an object wave falls down to the first minimum $\eta = 0$. We differentiate between ***spectral selectivity*** of a hologram that is associated with variations in a frequency spectrum of the hologram recording wave $\Delta\lambda = \lambda_B - \lambda_O$ and ***angular selectivity*** associated with variations of the hologram reconstruction angle (spatial spectrum) $\Delta\varphi = \varphi_B - \varphi_O$. A value of the angular and spectral selectivity is proportional to the ratio of the diffraction structure period to its thickness $\frac{\Lambda}{d}$. In the case of a thin hologram the grating period is much greater than its thickness ($\Lambda >> d$); a thin hologram has no angular and spectral selectivity. Indeed, according to the formula for a thin grating (2.9), each spectral component diffracts at its own angle α_∂. This means that, with a thin grating, weight light undergoes dispersion into a spectrum.

Variation of the incidence angle α_n of monochromatic light with the wavelength λ results in variations of the diffraction angle α_∂. Diffraction always takes its effect irrespective of the reconstruction conditions of a thin hologram. Because of this, thin holograms may be observed only in coherent monochromatic light over the narrow spectral range $\Delta\lambda$. When a thin hologram is illuminated by a broad spectral beam, each spectral component is diffracted at its own angle and the hologram reconstructs numerous spatially-shifted monochromatic images. The resultant image reconstructed by a thin hologram is broad, looking as a blurred spectrally colored spot.

As distinct from a thin hologram, volume holograms possess spectral and angular selectivity.

An accurate analysis of the selective properties of volume holograms is performed on the basis of a coupled wave theory.

Proceeding from the Kogelnik theory, an amplitude of the wave diffracted from a ***volume transmission phase*** hologram is dependent on the two parameters

$$\xi = \delta \frac{2\pi}{\lambda} \sin\frac{\theta}{2} \ , \tag{3.4.1}$$

$$\nu = \frac{\pi \Delta n d}{\lambda \cos\dfrac{\theta}{2}}, \tag{3.4.2}$$

where δ - angle of deviation from the Bragg angle on hologram reading, Δn - refractive index modulation, θ - angle between a reference and an object wave in a medium on recording of a hologram, λ - light wave length in a medium.

In this way the parameter ξ specifies the geometry for hologram reading, and the parameter ν - phase modulation of light.

An amplitude of the diffracted wave at the output of the hologram with the thickness *d* is given by

$$A(d) = i \frac{\exp(-i\xi)\sin\left(\xi^2 + \nu^2\right)^{\frac{1}{2}}}{\left(1 + \xi^2 / \nu^2\right)^{\frac{1}{2}}}. \tag{3.4.3}$$

When reading of a hologram is at the Bragg angle (δ =0), the parameter ξ = 0. In this case the diffraction efficiency of a volume transmission hologram, with regard to the fact that an amplitude of the wave incident on the hologram is equal to 1, may be found as

$$\eta = \sin^2\left(\frac{\pi\Delta n d}{\lambda\cos\frac{\theta}{2}}\right). \qquad (3.4.4)$$

According to (3.4.4), the efficiency of a volume transmission hologram approximates 100% for the following argument:

$$\frac{\pi\Delta n d}{\lambda\cos\frac{\theta}{2}} = \frac{\pi}{2}. \qquad (3.4.5)$$

From (3.4.5) it follows that even at a minor modulation of the refractive index Δn, meeting the Bragg conditions, we can achieve the diffraction efficiency close to 100% due to an increase in the hologram thickness d. If a hologram is read on detuning from the Bragg angle ($\delta \neq 0$), the parameter ξ is nonzero. In this case the diffraction efficiency could hardly reach 100%. Fig. 3.4.1 presents the relative diffraction efficiency η/η_0 of a transmission phase hologram as a function of the parameter ξ for different values of $\nu = \frac{\pi}{2}; \frac{\pi}{4}; \frac{3\pi}{4}$.

As seen from graph in Fig. 3.4.1, the diffraction efficiency for different values of the parameter $\frac{\pi}{2} \geq \nu \geq \frac{\pi}{4}$ is zero when $\xi \approx$ }. Substituting $\xi \approx$ } into (3.4.1), we can find an approximate formula for estimation of the hologram angular selectivity as follows:

$$\Delta\varphi = \frac{3\lambda}{2\pi\ d\sin\frac{\theta}{2}} \approx \frac{\lambda}{2\,d\sin\frac{\theta}{2}}, \qquad (3.4.6)$$

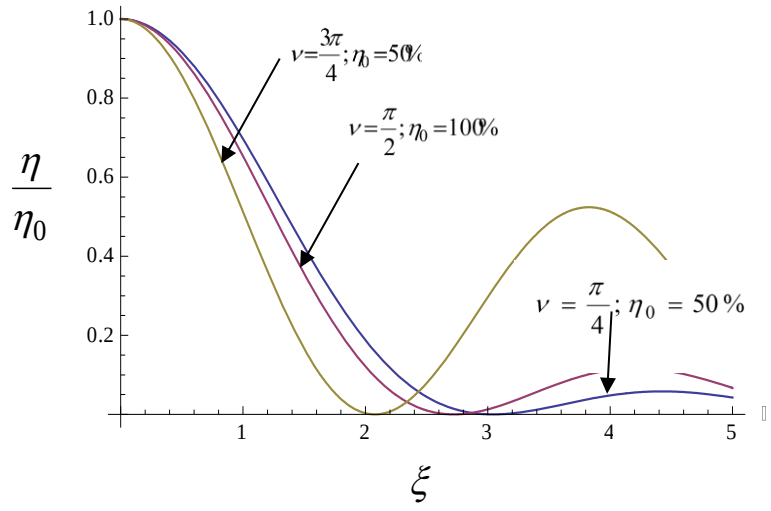


Fig. 3.4.1 – Relative diffraction efficiency η/η_0 of a transmission phase hologram as a function of different values of the parameter $\nu=\frac{\pi}{2};\frac{\pi}{4};\frac{3\pi}{4}$

where λ - radiation wave length on recording a hologram with the thickness d at the angle θ between the reference and object beams. Considering the Bragg condition of (3.2.8), for +1 order diffraction we use (3.4.6) to evaluate the angular selectivity of a volume transmission hologram

$$\Delta\varphi \approx \frac{\Lambda}{d}. \tag{3.4.7}$$

To find an approximate relationship for spectral selectivity, we can write the Bragg condition (3.2.8) with due regard for the fact that the hologram is read using a radiation source with the wave length $\lambda+\Delta\lambda$

$$2\Lambda\sin(\frac{\theta}{2}+\Delta\varphi)=\lambda+\Delta\lambda. \tag{3.4.8}$$

As seen from (3.4.8), a maximal diffraction efficiency of a hologram is attained when the hologram is illuminated at the angle $\frac{\theta}{2}+\Delta\varphi$. In the approximation of small angles $\Delta\varphi(\sin\Delta\varphi=\Delta\varphi,\cos\Delta\varphi=1)$, from (3.4.8) we obtain:

$$\Delta\varphi = \frac{\Delta\lambda}{2\Lambda\cos\dfrac{\theta}{2}}. \qquad (3.4.9)$$

Considering the Bragg condition (3.2.8) and using (3.4.9) we can obtain:

$$\Delta\varphi = \frac{\Delta\lambda}{\lambda} tg\,\frac{\theta}{2}. \qquad (3.4.10)$$

With regard to the expression for the angular selectivity $\Delta\varphi$ (3.4.7), from (3.4.11) we can derive an approximate formula to estimate the spectral selectivity

$$\Delta\lambda \approx \lambda\frac{\Lambda}{d} ctg\,\frac{\theta}{2}. \qquad (3.4.11)$$

Expressions (3.4.7) and (3.4.11) approximately describe the angular $\Delta\varphi$ and the spectral $\Delta\lambda$ selectivity of a volume transmission phase hologram for $\frac{\pi}{2} \geq \nu \geq \frac{\pi}{4}$. Deviations of the angle $\Delta\varphi$ with respect to the reference beam geometry and of the wave length $\Delta\lambda$ on hologram recoding are associated with lowering of the diffraction efficiency to zero.

Proceeding from the Kogelnik theory, an amplitude of the wave diffracted from a ***volume reflection phase*** hologram with the thickness *d* is determined as follows:

$$S(0) = \frac{-i}{\left(i\xi/\nu\right) + \{1-(\xi/\nu)^2\}^{\frac{1}{2}} cth\left(\nu^2-\xi^2\right)^{\frac{1}{2}}}. \qquad (3.4.12)$$

The parameters defining the reading geometry and phase modulation of light may be expressed as

$$\xi = \delta\beta d\cos\theta_0, \qquad (3.4.13)$$

$$\nu = \frac{\pi\Delta n d}{\lambda\cos\dfrac{\theta}{2}}. \qquad (3.4.14)$$

When the Bragg condition is met (ξ=0), the diffraction efficiency of a hologram, with regard to (3.4.12), is determined as follows:

$$\nu = th^2\left(\frac{\pi\Delta n d}{\lambda\cos\frac{\theta}{2}}\right). \qquad (3.4.15)$$

As seen from (3.4.15), the diffraction efficiency of a volume reflection phase hologram is asymptotically approaching the maximal limiting value.

Fig. 3.4.2 shows the relative diffraction efficiency η/η_0 of a volume reflection phase hologram as a function of the parameter ξ calculated for different values of the parameter $\nu = \frac{\pi}{2};\frac{\pi}{4};\frac{3\pi}{4}$.

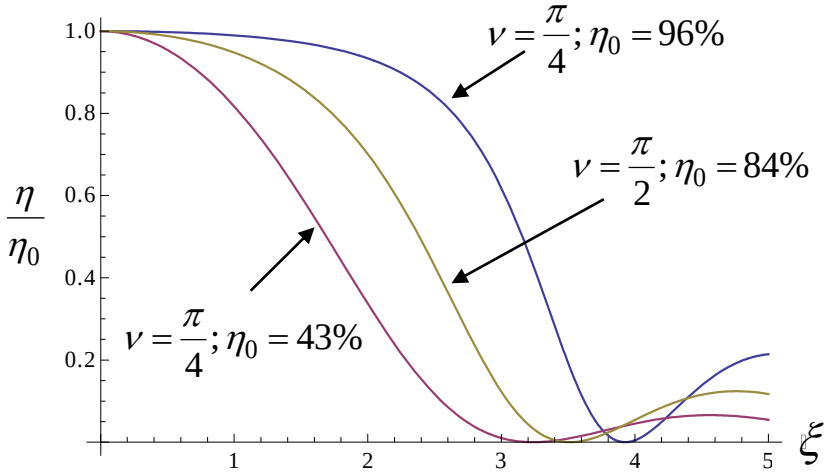


Figure 3.4.2 – Relative diffraction efficiency η/η_0 of a reflection phase hologram as a function of the parameter ξ calculated for different values of $\nu = \frac{\pi}{2};\frac{\pi}{4};\frac{3\pi}{4}$

As seen from Fig. 3.4.2, an increase in ν results in broadening of the first peak of the relationship between the diffraction efficiency of a hologram and the incidence angle of the hologram reading light. Similar to transmission holograms, with the use of Fig. 3.4.2 we can estimate the selective properties of a volume reflection phase hologram. For $\frac{\pi}{2} \geq \nu \geq \frac{\pi}{4}$, the diffraction efficiency decreases

to zero when ξ =3,5. Setting (3.4.13) equal to 3.5, we get the expression for the angular selectivity

$$\delta = \frac{3{,}5\lambda}{2\pi d \cos\theta}. \qquad (3.4.16)$$

Based on (3.4.16), with regard to the Bragg condition (3.2.8), we can write an approximate formula for a volume reflection phase hologram when $\frac{\pi}{2} \geq \nu \geq \frac{\pi}{4}$

$$\delta \approx \frac{\Lambda}{d} tg \frac{\theta}{2}. \qquad (3.4.17)$$

Substituting (4.17) into the expression of (4.10), we derive an approximate expression for the spectral selectivity of a volume reflection phase hologram as follows:

$$\Delta\lambda \approx \lambda \frac{\Lambda}{d}. \qquad (3.4.18)$$

Expressions (3.4.17) and (3.4.18) approximately describe the angular $\Delta\varphi$ and the spectral $\Delta\lambda$ selectivity of a volume reflection phase hologram for $\frac{\pi}{2} \geq \nu \geq \frac{\pi}{4}$.

Analyzing expressions (3.4.7), (3.4.11), (3.4.17), and (3.4.18), we can state that a volume transmission phase hologram has a very good angular selectivity growing with an angle between the reference and object beams. However, the spectral selectivity of such holograms is not high. A volume phase reflection hologram has a very good spectral selectivity, whereas the angular selectivity of such holograms is not high, lowering as the angle between the reference and object beams is increased.

3.5. Denisyuk hologram. Fourier hologram. Rainbow hologram.

3.5.1. Denisyuk hologram

Because a thin hologram has no spectral selectivity, the quality of the observed image of an object is not very good in a white light, because different spectral components of the spatially-shifted reconstructed holographic images are overlapping. Recording of a volume transmission hologram contributes nothing to solving of this problem too. For example, recording of a hologram at the generation wave length of a Nd-YAG laser operating in the second-harmonic generation mode ($\lambda = 0{,}532\mu m$) when the angle between the recording beams is $\theta = 40^{\circ}$, in accordance with (3.4.11), offers the spectral selectivity $\Delta\lambda \approx 97nm$. Such spectral selectivity on reconstruction of the image over a wide wavelength range leads to its spatial blurring with spectral coloration. In 1062 the problem was solved by the Soviet physicist Yu.N. Denisyuk (Russia) who suggested to record volume reflection holograms in counter-propagating beams. Fig. 3.5.1 shows the schemes for recording (a) and reading (b) of a Denisyuk hologram.

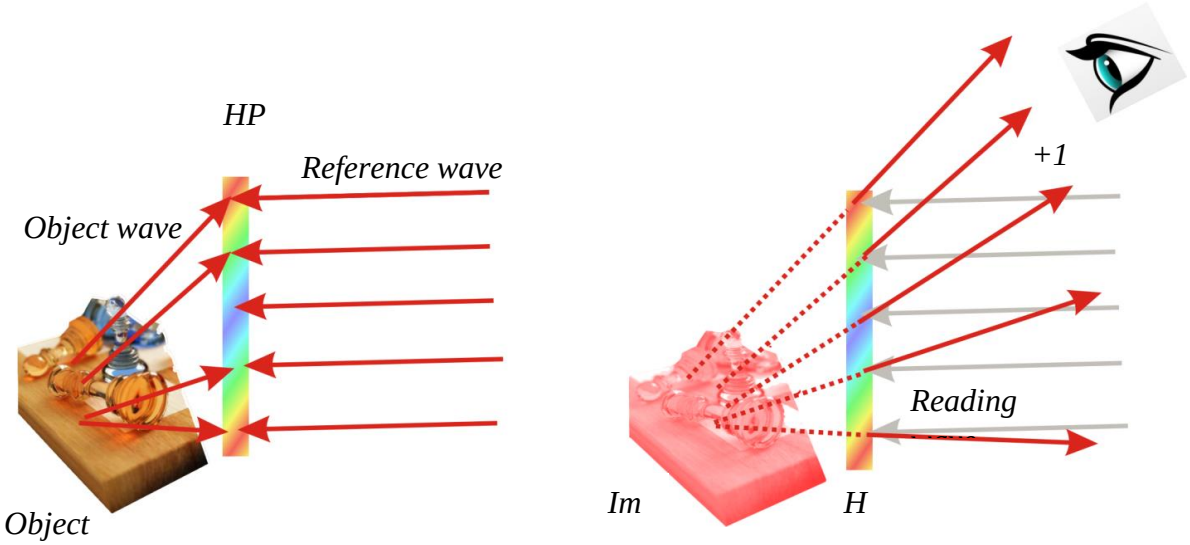


Figure 3.5.1 - Denisyuk scheme for recording (*a*) and reconstruction (b) of holograms.

When a hologram is recorded according to the Denisyuk scheme, the object and reference beams are on different sides of the holographic (Fig. 3.5.1,a).

To evaluate the spectral selectivity of a volume reflection hologram, let us assume that , when the hologram is recorded at the angle between the hologram recording waves $\theta = 180^{\circ}$ and at the wave length $\lambda = 532nm$ in a holographic emulsion with the layer thickness $d = 10\mu m$, its spectral selectivity is on the order of $\Delta\lambda \approx 13nm$. On reconstruction of a volume reflection hologram in a white

light the only spectral component that diffracts is the component meeting the Bragg diffraction condition (3.2.19), and the image is reconstructed in a narrow spectral range. All other spectral components of reconstructing light are transmitted through the hologram without diffraction. In this case the reconstruction involves a single image of the object, virtual or real. Fig. 3.5.1,b shows the case when a virtual image is reconstructed. To form a real image, the hologram should be illuminated by the wave that is conjugate to the reference one.

The method of recording and reconstructing volume images in accordance with the Denisyuk scheme was of crucial importance for the development of commercial picture holography. The use of three wave lengths (red, green, blue) in the process of holographic recording enables the formation of color reflection holograms rendering not only the form of an object but also its color.

3.5.2. Fourier hologram

Spatial frequency.

A complex amplitude of the spatially modulated wave may be considered either in a coordinate region of the plane *a(x,y)* through which light is transmitted or in the spatial frequency region $A(\xi,\eta)$.

The complex wave amplitude $a(x,y)$ in the coordinate region is related to a complex amplitude in the spatial frequency region $A(\xi,\eta)$ by means of the Fourier transform

$$F[a(x.y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(x.y)\exp(2\pi i\xi x)\exp(2\pi i\eta y)dxdy = A(\xi,\eta)$$
$$F^{-1}[A(\xi,\eta)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(\xi,\eta)\exp(-2\pi i\xi x)\exp(-2\pi i\eta y)d\xi d\eta = a(x,y). \quad (3.5.1)$$

The function $A(\xi,\eta)$ is a Fourier transform of the function $a(x,y)$. The transformation operation of the function $a(x,y)$ from the coordinate region to the region of spatial frequencies $A(\xi,\eta)$ is termed the ***direct Fourier transform*** and is mathematically represented as $a(x,y) \supset A(\xi,\eta)$. The transformation operation of the function $A(\xi,\eta)$ from the spatial frequency region to the coordinate region $a(x,y)$ is called the ***inverse Fourier transform***. The function $a(x,y)$ is the inverse Fourier transform of the function $A(\xi,\eta)$. The inverse Fourier transform is mathematically represented as $A(\xi,\eta) \subset a(x,y)$.

Spatial frequencies of a wave field are determined by spatial oscillations of the field along the specified direction. Let us consider a plane wave shown in Fig.3.5.2.

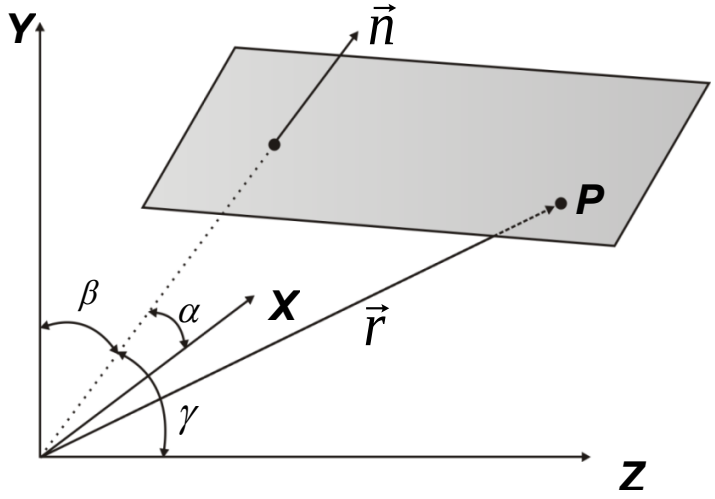


Figure 3.5.2 – Plane wave in the rectangular coordinate system

The plane wavefront condition is given by the expression

$$\vec{r} \cdot \vec{n} = const, \qquad (3.5.2)$$

where $\vec{r}$ - radius-vector of the space point, $\vec{n}$ – unit vector normal the wave surface. Let a complex electric field be of the form

$$V(x,y,z,t) = a_1 \exp(-ik\vec{r} \cdot \vec{n})\exp(i2\pi\nu t), \qquad (3.5.3)$$

where a_1-wave amplitude, *k*-wave number, ν -wave frequency. The expression of (3.5.3) may be rewritten in terms of the direction cosines $\cos\alpha$, $\cos\beta$, $\cos\gamma$ for the vector $\vec{n}$ corresponding to the coordinate axes (x, y, z) as follows:

$$\begin{aligned}
&V(x,y,z,t) = a_1 \exp(-ik\vec{r} \cdot \vec{n})\exp(i2\pi\nu t) = \\
&a_1 \exp(-ik(x\cos\alpha + y\cos\beta + z\cos\gamma))\exp(i2\pi\nu t) = \\
&a_1 \exp\left(-i2\pi\left(x\frac{\cos\alpha}{\lambda} + y\frac{\cos\beta}{\lambda} + z\frac{\cos\gamma}{\lambda}\right)\right)\exp(i2\pi\nu t) = \\
&a_1 \exp(-i2\pi(x\xi + y\eta + z\zeta))\exp(i2\pi\nu t),
\end{aligned} \qquad (3.5.4)$$

where the quantities $\xi = \frac{\cos\alpha}{\lambda}, \eta = \frac{\cos\beta}{\lambda}, \zeta = \frac{\cos\gamma}{\lambda}$ are referred to as the ***spatial frequencies***. Spatial frequencies are inverse to the spatial periods measured along the axes *x,y,z*.

Sometimes spatial frequencies are expressed in terms of the angles $\theta_1 = \frac{\pi}{2} - \alpha, \theta_2 = \frac{\pi}{2} - \beta, \theta_3 = \frac{\pi}{2} - \gamma$.

In this case the spatial frequencies are of the form

$$\begin{aligned} \xi &= \frac{\sin\theta_1}{\lambda} \\ \eta &= \frac{\sin\theta_2}{\lambda} \\ \varsigma &= \frac{\sin\theta_3}{\lambda}. \end{aligned} \tag{3.5.5}$$

Fig 3.5.3 presents an example of calculations for the spatial frequencies of a plane wave propagating within the plane (Y,Z). As seen in Fig. 3.5.3, the spatial frequency (period) of oscillations of a plane wave propagating within the plane (Y,Z) with respect to the *x* direction is zero. At the same time, the spatial frequencies, measured with respect to the axes *y* and *z*, have identical values

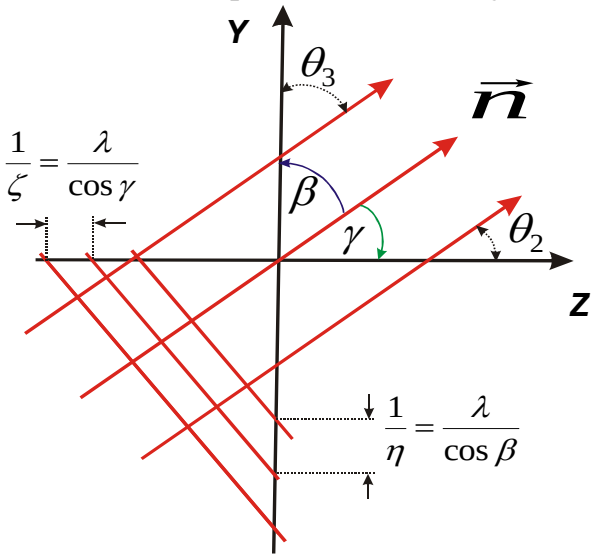


Figure 3.5.3 – Calculation of the spatial frequencies for a plane wave propagating within the plane (y,z)

determined by the wavelength and incidence angles with respect to the *y* and *z* directions.

Fourier hologram.

In the general case the Fourier hologram may be defined as a hologram of a two-dimensional (2D) object written with the help of the reference beam that is within the object plane. There are different types of Fourier holograms depending on the way of their recording: with or without lenses. However, these methods of the Fourier hologram recording are associated only with 2D objects and are not applicable to the objects going beyond the bounds of the hologram plane.

Every holographic image may be decomposed in the form of a 2D spatial frequency spectrum. By this operation, an image is represented as a set of sinusoidal diffraction gratings with different periods and orientations.

Fig. 3.5.4 shows a scheme for recording of the Fourier hologram according to the Vander Lugt method. In this case interference of two waves is recorded, with the complex amplitudes within the hologram plane representing Fourier transforms of the object and of the reference source. It is known that the Fourier transform of a 2D object positioned in the front focal plane of a lens is formed in

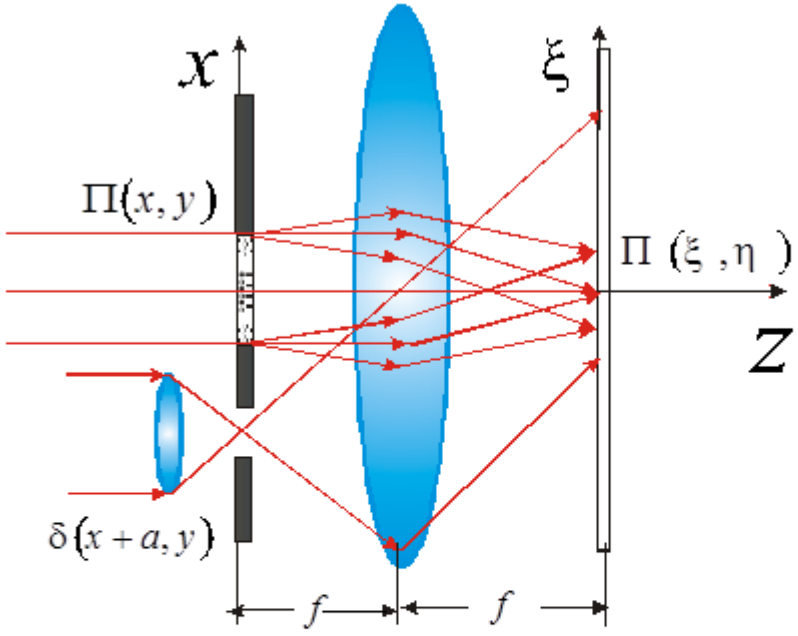


Figure 3.5.4 – Scheme to form Fourier holograms

its back focal plane. When a wave is plane, its Fourier transform is a point at the lens focus. When a wave is spherical, its Fourier transform is represented by the wave with a plane wave front.

Fig. 3.5.4 demonstrates as a reference beam the transparency with the amplitude $t_{tr}(x, y)$ positioned in the front focal plane of the lens and illuminated by the plane wave, with the constant amplitude r_0, that is propagating along the axis z. Passing through the transparency, the plane wave is scattered within the plane X, Y to form an object wave in the coordinate region of the hologram with the complex amplitude distribution $\Pi(x, y)$. An amplitude of the object wave within the hologram plane, at the back focal plane of the lens, is the Fourier transform of the object wave $\Pi(\xi, \eta)$

$$\Pi(\xi,\eta) = \int \Pi(x,y) e^{2\pi i(x\xi + y\eta)} dxdy, \qquad (3.5.6)$$

where ξ and η – spatial frequencies of the object wave measured along the directions x, y.

A point source, with a complex amplitude described by the delta-function $\delta(x + a, y)$, forms a reference wave and is also positioned in the front focal plane of the lens. The reference wave may be represented as a Fourier transform of the point source in the following way:

$$O(\xi,\eta) = \delta(\xi,\eta) = \int \delta(x + a, y) e^{2\pi i(x\xi + y\eta)} dxdy = e^{-2\pi i(\xi a)}. \quad (3.5.7)$$

As seen from (3.5.7), in the back focal plane of the lens (within the hologram plane) the reference wave is a plane wave with the complex amplitude distribution in the spatial frequency ξ measured along the axis X (the spatial frequency η measured along the axis y is equal to 0).

Considering (3.5.6) and (3.5.7), for the interference field recorded by the hologram we can write

$$I = |\Pi|^2 + 1 + \Pi^*(\xi,\eta) e^{-2\pi i(\xi a)} + \Pi(\xi,\eta) e^{+2\pi i(\xi a)}. \qquad (3.5.8)$$

When the Fourier hologram is recorded in the linear mode, the hologram transmission $t(x, y)$ is proportional to the interference pattern intensity

$$t(x,y) \approx I = |\Pi|^2 + 1 + \Pi^*(\xi,\eta) e^{-2\pi i(\xi a)} + \Pi(\xi,\eta) e^{+2\pi i(\xi a)}. \quad (3.5.9)$$

Fig. 3.5.5 shows a scheme for reconstruction of the Fourier hologram.

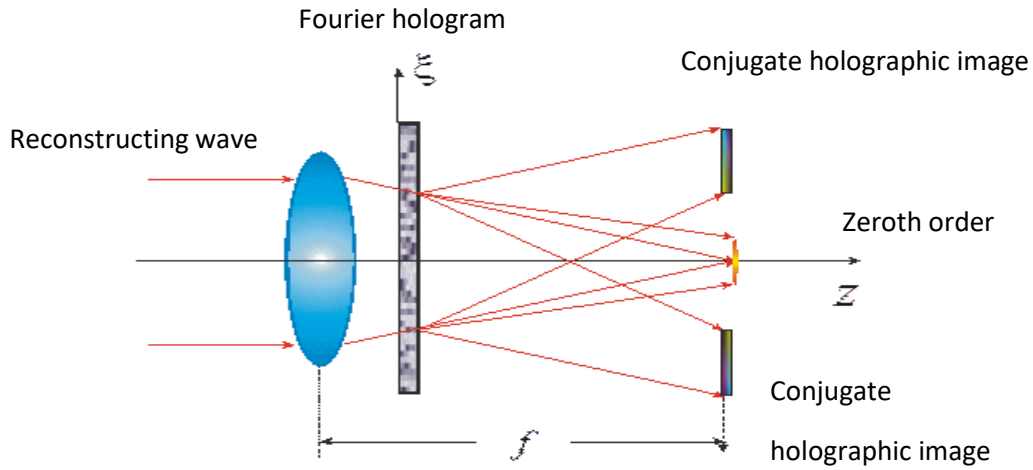


Figure 3.5.5 – Reconstruction of the Fourier hologram

We illuminate the Fourier hologram by a plane wave, with the constant amplitude r_0, propagating along the axis z. Then a field on the other side of the hologram (amplitude of the diffracted light is directly beyond the hologram) may be given by

$$W \approx r_0 t(x,y) \approx 1 + |\Pi|^2 + \Pi^*(\xi,\eta)e^{-2\pi i(\xi a)} + \Pi(\xi,\eta)e^{+2\pi i(\xi a)}. \qquad (3.5.10)$$

The lens positioned before the hologram or beyond the hologram creates at the lens focus the distribution corresponding to the product of the inverse Fourier transform for the function W into the phase factor of a spherical wave. Recording the intensity, we omit the phase factor to obtain the following expressions for the virtual and real images:

$$\begin{aligned} &\Pi\ (\xi,\eta)e^{2\pi i(\xi a)} \Rightarrow \Pi(x-a,y) \\ &\Pi^*(\xi,\eta)e^{-2\pi i(\xi a)} \Rightarrow \Pi^*(-(x+a),-y). \end{aligned} \qquad (3.5.11)$$

In this way we reconstruct two images at one side of the hologram: one image $\Pi(x-a,y)$ is shifted with respect to the initial transmission to the negative region along the axis X (virtual image), whereas the real image $\Pi^*(-(x+a),-y)$ that is conjugate and mirror-symmetric to the initial transmission is shifted by the distance *a* from the origin of coordinates to the negative region. In both cases the diffracted light in both diffraction orders is convergent to form two real images. Fig. 3.5.6 shows a photograph of the image plane for the reconstructed Fourier hologram. The main feature of the Fourier hologram is that both direct and conjugate holographic images are on the same side of the hologram in one plane that contains the reconstructing beam too. Most important for Fourier holograms is their property to retain immobility of the reconstructed image position on transverse shifting of the hologram.

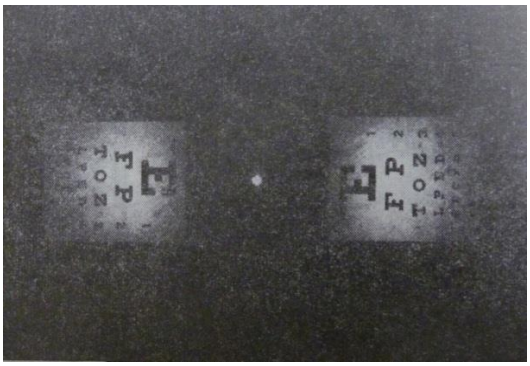


Figure 3.5.6 – Photograph of the image plane for the reconstructed Fourier hologram

3.5.3. Rainbow hologram.

Currently, optical holography has numerous applications in science and engineering: picture holograms for exposition of unique and priceless objects, manufacturing of various diffraction optical elements to be used in scientific research and to solve different applied problems in printing, biology, medicine, etc. Holography is of particular importance in protection of valuable papers and documents. Because of this, it is essential to use simple processes for mass hologram replication. The well-known techniques of hologram replication are optical (holographic) and mechanical replication. In essence, the optical method of replication is similar to hologram recording but the wave front formed by a real object is replaced by the hologram-reconstructed wave front of the object – recording of the hologram is realized using the reconstructed real image from the original hologram. In the process of mechanical replication we produce replicas

of the surface holographic relief. Such replication is mostly applicable for copying of thin phase holograms with information recorded in the form of spatial variations in the spatial relief. These replicas are produced by molding and embossing directly from the original or by embossing the hologram matrix produced with the use of galvanoplastics. Thin holograms formed according to the Leith-Upatnieks scheme are replicated by a fairly simple and inexpensive technology of embossing. But the Leith-Upatnieks holograms recorded in thin photosensitive layers feature no spectral selectivity. Because of this, in a white light from such holograms numerous volume images are reconstructed in all the spectral components. These images are spatially shifted relative each other (see Fig. 3.5.7) and, as a result, we observe a blurred light spot with the spectral coloring at the boundaries rather than a sharp image. The problem was solved in 1969 by the American physicist Stephen Benton. He has proposed an original method to lower the information capacity of a hologram, in fact without any loss in volume (three-dimensionality) of images. On reconstruction of an rainbow hologram, the spectral composition of the image «renders» all the rainbow colors when the reading angle is changed (Fig. 3.5.8). No wonder that such holograms are called the rainbow holograms. For replication of rainbow holograms, it is sufficient once to record a hologram according to the Benton scheme and manufacture on its basis the metal matrix that is used during the holographic replication. This allows for manufacturing of rainbow holograms by the fairly inexpensive and rapid technological procedure.

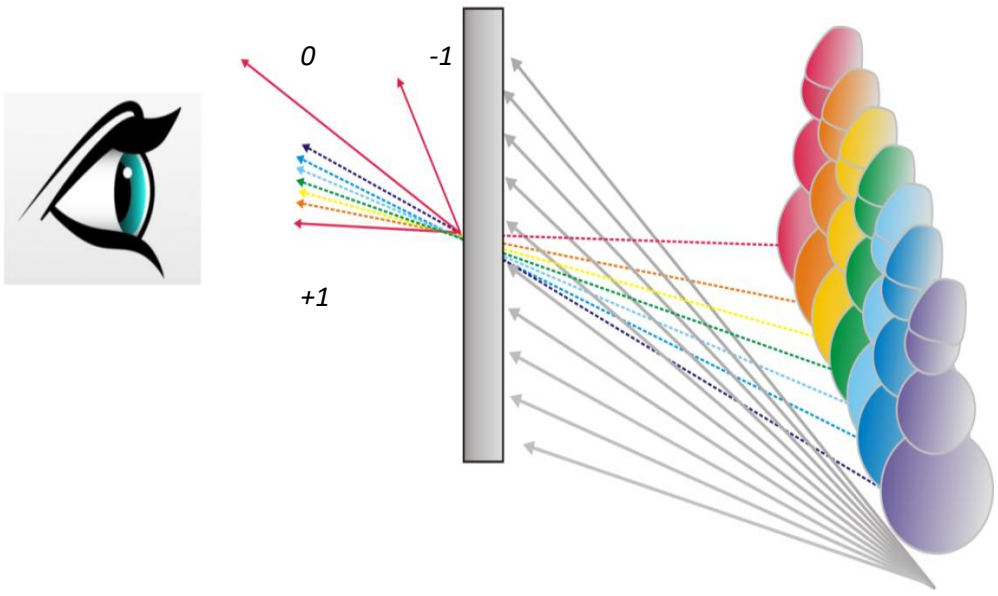


Fig. 3.5.7 – Reconstruction of Leith-Upatnieks hologram by a white light.

Figure 3.5.8- Images reconstructed from an iridescent hologram

Recording of an rainbow hologram by Benton's method includes two successive stages for recording of two holograms according to the Leith-Upatnieks scheme. At the first stage, an ordinary (master) hologram is recorded (see Fig.3.5.9).

An object is usually positioned at the distance 25-30 cm that is considered the distance of best vision.

The second stage is a repeated recording (rerecording) of the already recorded holographic image. To this end, a pseudoscopic volume image is reconstructed on the basis of the hologram with the use of the reconstructing beam that is conjugate to the reference beam in the process of recording (Fig.3.5.10).

In the process a narrow horizontal slit is left on the hologram for reconstruction of the real image. As a rule, orientation of the slit is selected perpendicular to the vertical axis of the object – to retain the horizontal plane

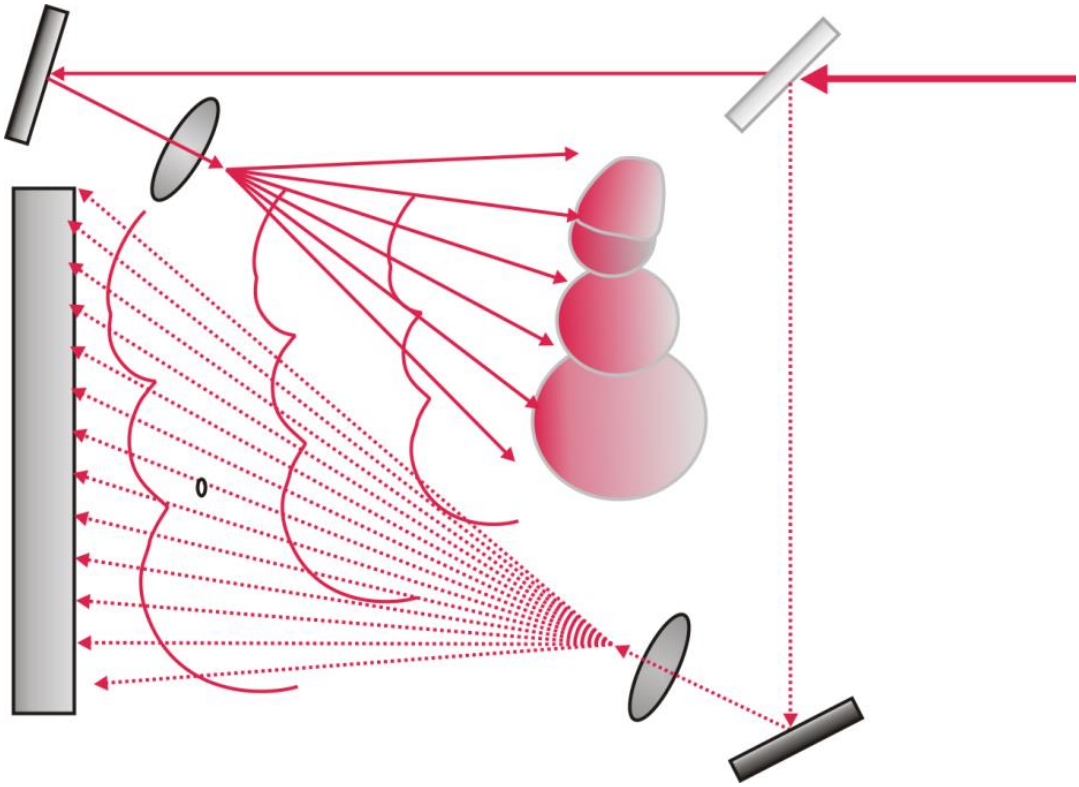


Figure 3.5.9–First stage of recoding the iridescent hologram in accordance with the Leith-Upatnieks scheme.

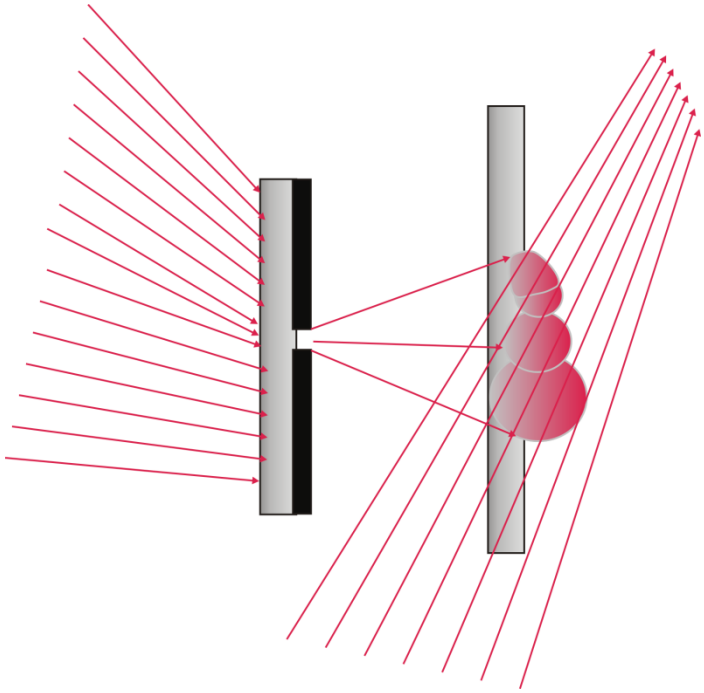


Figure 3.5.10 – Second stage of recording the iridescent hologram in accordance with the Benton scheme.

parallax of the reconstructed image. A light sensitive layer is placed at the localization of the reconstructed real image and subjected to the effect of the reference beam that is coherent to the beam reconstructing an image from the first hologram. At the second stage of an rainbow hologram the image from the master hologram is reconstructed through a narrow slit and recorded.

If the produced hologram is illuminated by the white light beam conjugate to the reference beam, we can observe both the orthoscopic image of the object and image of the slit – strip shimmering in colors of a visible spectrum. A pupil of the observer's eye, if located at the slit region, isolates from the whole spectrum the only color component that determines the percepted reconstructed image (Fig. 3.5.11).

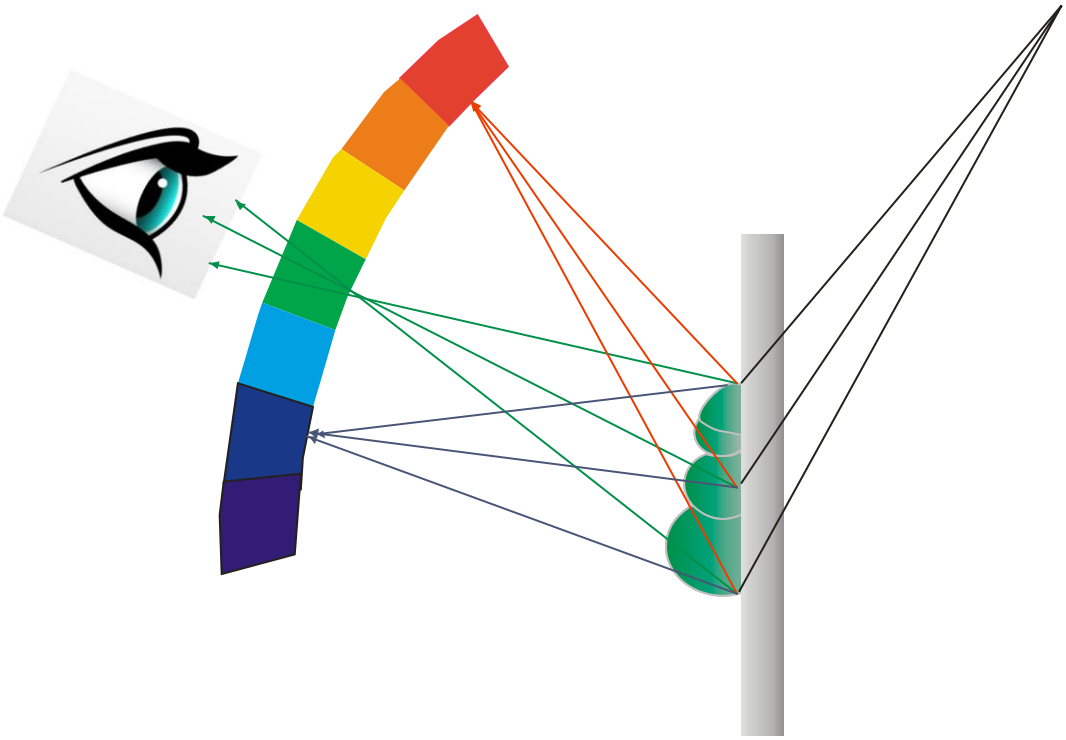


Figure.3.5.11 – Reconstruction of the iridescent hologram according to the Benton scheme.

Each image of the slit in a particular spectral component plays a role of a monochromator for observation of the reconstructed holographic image in the corresponding color. Variations in the angle of view or in the incidence angle of reading light lead to changes in color of the reconstructed image. When an observer moves the head within the limits of one color strip, similar to the ordinary hologram, we observe parallax of the reconstructed volume holographic image. In the direction perpendicular to the slit there is no parallax but this has practically

no effect on three-dimensionality of the percepted image as observer's eyes are horizontally aligned. For a wide horizontal angle of view, a length of each strip (rainbow width) should be large enough due to a great width of the primary hologram.

The angular dimensions of the reconstructing light source mainly determine the scene depth of the image reconstructed from an rainbow hologram. From every point of the extended light source, an independent image is formed, being shifted in the transverse direction with respect to the images reconstructed by other points of the source. This shift is the greater the farther the object image from the hologram and the larger the angular distance between the source points.
References

3.6. Dynamic holography

In the preceding sections we have considered holograms recorded in a photosensitive material and retaining their properties during a long period of time after the recording. Along with the use of holographic materials for irreversible fixation of images, holograms may be recorded in nonlinear-optical media. Such holograms are known as dynamic holograms. Their lifetime is determined by the optical nonlinearity relaxation time. Dynamic holograms are recorded in nonlinear media capable to vary their absorption factor and/or refractive index in real time under the effect of laser radiation. There is no need in postexposure treatment as is the case, e.g., with silver halide photomaterials. The first dynamic holograms were recorded in 1971 in a dye solution by the Belarusian physicists B.I. Stepanov, A.S. Rubanov, and E.V. Ivakin who introduced the notion of dynamic holography.

Selection of a nonlinear medium is dictated by the requirements to the spectral sensitivity region and to the speed of response depending on the parameters of laser radiation used (wavelength, intensity, pulse length). Among nonlinear-optical media used for recording of dynamic holograms, of particular importance are resonant media (atomic, simple and complex molecular media, photochromic materials) which enable operation over the range from infra-red to ultraviolet with the use of both continuous-wave and pulsed laser radiation. In the case of low-intensity luminous fluxes photorefractive crystals are more advantageous: operation is realized at the milli- and microwatt powers of optical radiation.

Dynamic holography arose at the junction of classical holography and nonlinear optics. Owing to combination of the principles of holography and of nonlinear optics, new methods of transforming the spatial-temporal structure of laser radiation have been proposed, the techniques and devices for optical data processing have been developed. The dynamic holography principles provided the basis for the development of holographic correlators of images, RAM systems, and fast image-control systems.

The greatest achievement of dynamic holography is a discovery of the wavefront conjugation effect. A conjugate wave is obtained when on a nonlinear medium we guide two plane reference waves propagating in counter directions and an object wave (Fig. 3.6.1). Each of the reference waves and an object wave are recording a dynamic hologram that is read by another reference wave. The reconstructed conjugate wave is a real image of the recorded object, it has the same amplitude and phase distributions as the object wave but is propagating in the counter direction. The conjugate wave reproduces the state of the object wave in the previous instants of time, offering the possibility to eliminate the wave phase distortions when it returns back through an inhomogeneous scattering medium. This property of a conjugate wave makes it possible to compensate for phase distortions in high-power laser amplifiers and to realize self-guidance of laser radiation to the target that is used for systems of laser fusion synthesis and space-based systems.

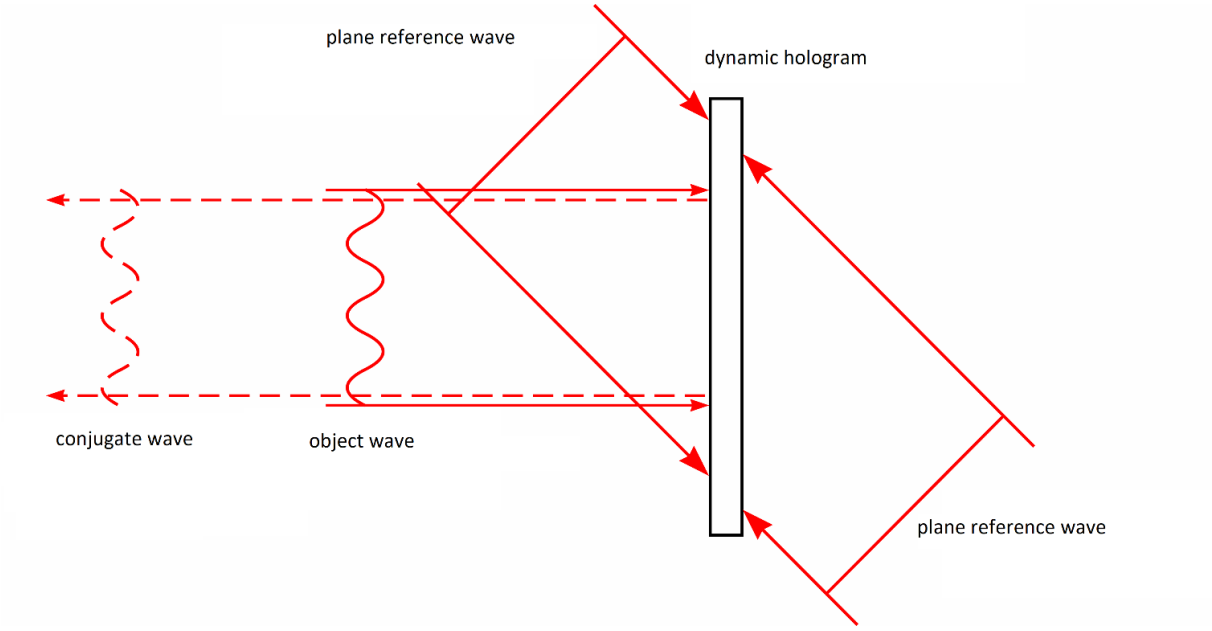


Fig. 3.6 1. Scheme for observation of the wavefront conjugation effect

The above-mentioned holographic approach to description of dynamic holograms and wavefront conjugation is useful for physical interpretation of the observed effects. To construct a quantitative model and to give numerical description for recording and reading processes of dynamic holograms, we can use a nonlinear-optical analysis of the interaction processes. From the viewpoint of nonlinear optic, the scheme given in Fig. 3.6.1 represents a variant of four-wave interaction (see section 2.5). Two counter-propagating reference waves and an object wave are incident on a nonlinear medium. As a result of the interaction between these three waves, the fourth conjugate wave is formed. Such a variant of nonlinear interaction is described with the use of a four-wave interaction scheme in a medium with cubic nonlinearity.

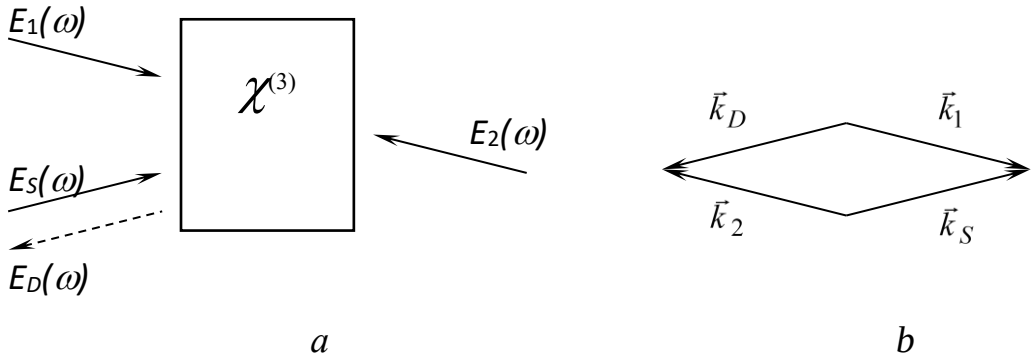


Fig. 3.6 2. Scheme of four-wave interaction and wave vector diagram

Fig. 3.6.2, *a* shows a scheme of the interaction. The counter-propagating reference waves E_1 and E_2, interacting with the signal wave E_S, form the diffracted wave E_D.

As indicated in section 2.5, in the approximation of weak (compared to the reference waves) signal and diffracted waves the reduced wave equations for the waves E_S and E_D take the form

$$\frac{\partial A_S}{\partial z} = i\frac{3\pi\omega}{cn}\chi^{(3)} A_1 A_2 A_D^* \exp(-i\Delta kz). \qquad (3.6.1)$$

$$\frac{\partial A_D}{\partial z} = -i\frac{3\pi\omega}{cn}\chi^{(3)} A_1 A_2 A_S^* \exp(-i\Delta kz), \qquad (3.6.2)$$

where $\Delta\vec{k} = \vec{k}_1 + \vec{k}_2 - \vec{k}_S - \vec{k}_D$. The condition of counter propagation for the waves E_S and E_D allows for «minus» sign in equation (3.6.2).

The effective interaction takes place when the phase-matching condition ($\Delta k = 0$) is fulfilled; it is realized on counter propagation of the reference waves: $\vec{k}_1 + \vec{k}_2 = \vec{k}_S + \vec{k}_D = 0$ (Fig. 3.6.1, *b*).

In this case equations (3.6.1), (3.6.2) may be written as

$$\begin{cases} \partial A_S / \partial z = i\sigma A_D^* \\ \partial A_D / \partial z = -i\sigma A_S^* \end{cases}, \tag{3.6.3}$$

where $\sigma = \dfrac{3\pi\omega}{cn}\chi^{(3)}A_1A_2$.

A system of the first-order differential equations is transformed to the second-order differential equation

$$\frac{\partial^2 A_S}{\partial z^2} + \sigma^2 A_S = 0, \tag{3.6.4}$$

a solution of which may be given in the following form:

$$A_S = C_1 \sin(\sigma z) + C_2 \cos(\sigma z). \tag{3.6.5}$$

A similar equation is the case for the diffracted wave too

$$A_D = iC_1 \cos(\sigma z) - iC_2 \sin(\sigma z). \tag{3.6.5}$$

To find the constants C_1 and C_2 , we use the boundary conditions

$$\begin{aligned} A_S(z=0) &= A_0 \\ A_D(z=L) &= 0 \end{aligned}, \tag{3.6.6}$$

which suggest that a signal wave with the amplitude A_0 is incident on a medium and the diffracted wave is formed within the nonlinear medium.

In this case a solution of a system of equations (2.5.23) enables one to find values of the diffracted wave amplitude at the output from a nonlinear medium as follows:

$$A_D(z=0) = iA_0^* tg(\sigma L)\,. \qquad (3.6.7)$$

Note that in equations (3.6.7) $tg(\sigma L)$ tends to infinity $\sigma L \to \pi/2$, allowing for realization of the amplification due to the energy transfer from two high-power reference waves. Also note that the amplitude of a signal wave is conjugated. Considering counter propagation of signal and diffracted waves, this means that the diffracted wave is conjugated with respect to the signal wave. The wavefront conjugation effect is realized when the signal and the diffracted wave have identical spatial amplitude and phase distributions but are propagating in counter directions. The effective wavefront conjugation was attained in resonant media (atomic, simple and complex molecular media) which are characterized by a unique combination of fast response and sensitivity, enabling operations with continuous-wave laser radiation and ultrashort light pulses over a wide spectral range from infer-red (IR) to ultraviolet (UV) regions. A maximum reflection factor of the conjugate wave $R = |E_D|^2 / |E_S|^2 \approx 700$ on simultaneous reduction of the light pulse length was achieved with the use of sodium vapors as a nonlinear medium.

Considering dynamic holograms in resonant media, we should take into account that, along with cubic nonlinearity, one can observe manifestations of the fifth and higher order nonlinearities. These nonlinearities result in distortions of the holographic grating groove profile, leading to changes in the Bragg reflection condition that may be fulfilled for some fixed frequencies of the reading wave or its propagation directions. This enables one to realize spatial transformations of the wavefront structure (smoothing or amplification) or frequency transformations of volume images for their visualization. Dynamic holograms make possible implementation of the associative holographic memory when the required image can be reconstructed (selected from a set of recorded images) by means of the image fragment. The use of three optical waves during recording or reading of dynamic holograms enables realization of algebraic operations, matrix multiplication including, by the optical methods.

3.7. Holographic interferometry

Classical interferometry involves the interference of two relatively simple optical wavefronts which are formed and directed by optical components. The need for high-quality optical surfaces in classical interferometry is a consequence of the difficulty, using classical optical methods, of generating two separate but identical optical wavefronts of arbitrary shape.

Combining conventional interferometry with holography, one can produce three-dimensional interferograms – images of diffusely-reflecting, three-dimensional objects that are overlaid with interference fringes indicating areas of deformation or displacement in the object.

3.7.1. Holography

Holography is a technique used to record and then later reconstruct the light field scattered by an object. The light wave scattered by the object in question is referred to as the object wave. To reconstruct the object wave, a second light wave, called the reference wave, is created such that it is coherent with the object wave. The two waves are directed onto the same photographic recording material, and the resulting recording of the interference pattern that arises is called a hologram.

Before we examine applications of holographic interferometry, we must of course understand the principles behind holographic interferometry. To that end, we will briefly review holograph and interferometry, and then see how the two sciences are combined to create holographic interferometry.

A common configuration of die holographic setup is shown in Fig 3.7.1.

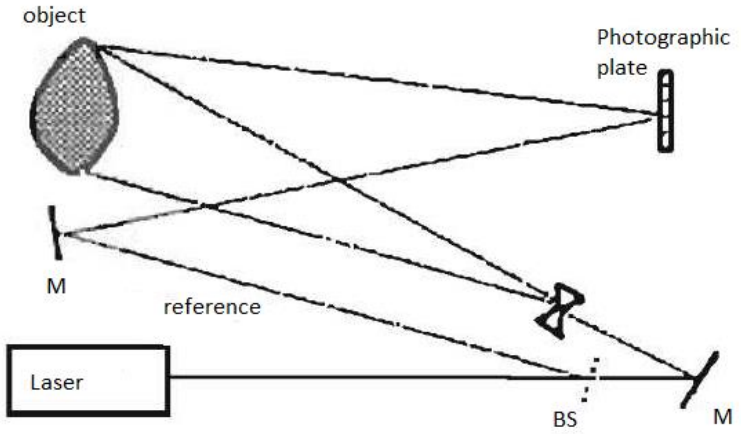


Fig 3.7.1. Common Holographic setup

Expressing the recording step mathematically, we have the object wave, which can be expressed as

$$U_0(x, y) = a_0(x, y) \exp[-i\varphi(x, y)], \tag{3.7.1}$$

the reference wave, which can be expressed as

$$U_R(x, y) = a_R, \tag{3.7.2}$$

and the resulting recorded intensity at the plane of the recording material:

$$l(x,y) = |a_R+U_0(x,y)^2|= a_R^2+|U_0(x,y)|^2 +a_RU_0(x,y) + a_RU^*_0 (x,y) \tag{3.7.3}$$

The developed recording material will have a transmittance proportional to the recorded intensity:

$$t(x,y) = t_b + \beta(a_RU_0(x,y) + a_RU_0^*(x,y)), \tag{3.7.4}$$

where t_b will be nearly uniform over the film, therefore representing a bias level in the exposure. β is the sensitivity parameter of the recording material.

To reconstruct the object wave, the hologram is illuminated by a uniform wave (often simply the reference wave itself):

$$Uc(x,y) = a_c, \tag{3.7.5}$$

leading to the reconstructed wave:

$$U_R(x,y) = Uc(x,y)*t(x,y)= a_c t_b+\beta a_c a_r U_0(x,y) + \beta a_c a_R U_0 (x,y) , \quad (3.7.6)$$

The term $\beta a_c a_r U_0(x,y)$ is proportional to the object wave, assuming the $a_c t_b$ term is uniform. By filtering other components, we have reconstructed a facsimile of the object wave.

Holography is a significantly more thorough (and difficult) method of capturing the image of an object than simply measuring the intensity at a plane some distance from die object (e.g., via photography). In holography, since the object wave itself is reconstructed, not only the magnitude of the light scattered from the object but also the phase of that light is recorded. Thus, when viewed from slightly different angles, the holographic image will appear to have the same three-dimensional characteristics as the original object itself.

3.7.2. Types of Holographic Interferometry

In conventional interferometry, an object wavefront is combined with a reference wavefront to create an interference pattern at the detector which allows for an understanding of the variations in the object wavefront.

Combining conventional interferometry with holography, one can produce three-dimensional interferograms – images of diffusely-reflecting, three-dimensional objects that are overlaid with interference fringes indicating areas of deformation or displacement in the object. Similarly, transparent objects will be overlaid with fringes indicating a change the refractive index of the material as due to a structural deformity or change in thickness. Such interferometry is possible because the light field scattered from an object may be first holographically recorded, then later holographically reconstructed and compared to another light field scattered from the same object under different conditions, all with interferometric precision. This type of interferometry, called holographic interferometry. is defined as interferometry in which at least one of the waves in question is holographically reconstructed. The combination of two or more waves (of which one is a hologram) is referred to as a holographic interferograms (whereas a simple interference pattern recorded on a fiat screen in intensity – as by the eye – is referred to simply as an interferogram, no modifier).

Double-exposure holographic interferometry

In double-exposure holographic interferometry, both the reference hologram and subject hologram are recorded in the same photographic plate. Tire exposures are made using a standard off-axis holographic imaging system as shown in Fig. 3.7.1. Upon development, the plate will then be illuminated using the reference beam, and the resulting hologram observed will consist of a three-dimensional image of the original object, overlaid with a pattern of interference fringes. Notably, the arrangement of the fringes will change as the viewing angle shifts, in similar fashion to tire viewing of a single hologram.

To understand the origin of the fringes, we'll begin by writing down the two fields that are recreated using the double-exposure hologram. The first field – the field of the object captured in the reference hologram – is written as:

$$U_0\,(x, y) = a(x, y)\exp[-i\varphi(x, y)]. \qquad (3.7.7)$$

The second field – the field of the stressed object captured m the subject hologram – is a fleeted primarily by a phase change due to the deformation:

$$U_0^*\,(x,y) = a(x,y)\exp[-j\{\varphi(x,y) + \Delta\varphi(x,y)\}] \qquad (3.7.8)$$

The observed intensity of the reconstructed wave is given by:

$$I(x,y)= |U_0(x,y)+U_0^*\,(x,y)|^2 = 2a^2(x,y)\{1 + \cos[\Delta\varphi(x,y)]\} \qquad (3.7.9)$$

The fact that the fields (and not intensities) of the two object waves add is the crux of holographic interferometry – it is what makes interferometry by holography superior to classical interferometry, particularly when investigating three dimensional effects. It is the linearity of the holographic process that gives rise to this phenomenon of field addition.

As (3.7.9) shows, the observed intensity will consist of the intensity of the original object modulated by the fringe pattern $\{1 + \cos[\Delta\varphi(x,y)]\}$.

Real-time holographic interferometry

In some cases, it is necessary or useful lo observe the response of the object under study to changing excitation in real time. In particular, when mechanical vibrations are the subject of concern – for example, if one wants to identify the

resonant frequencies of a given object by sweeping the excitation frequency over a wide range of values «real-time" holographic interferometry is the best choice. In real-time holographic interferometry, the light scattered by the object interferes by superposition with the holographically reconstructed image of the object itself, recreated using a reference hologram taken under at-rest conditions. The configuration used for real-lime holographic interferometry is identical (o that used for double-exposure holographic interferometry. The reference hologram is recorded, developed, and returned to the recording plane. When the laser is turned back on, the hologram will be illuminated simultaneously by the reference wave and the light scattered by the object under some kind of varying excitation. The observer will then see the interference of the two waves (holographic and object) that will describe not only (the deformity or displacement of the object but that will also indicate vibrational amplitude of the excitation.

Mathematically, if the complex amplitude of the holographically reconstructed wave is $U_0(x,y)$, then the instantaneous object wave can be expressed as $U_0\ (x, y)\ \exp[j\Delta\varphi(x, y,t)]$, where $\Delta\varphi(x, y,t)$ is the phase change due to the vibration at time t. The instantaneous intensity observed during real-time holographic interferometry will then be:

$$I(x,y,t) = |U_0(x,y)|^2+2(1-\cos[\Delta\varphi(x,y,t)]\} \qquad (3.7.10)$$

In real-time holographic interferometry, the fringes arise from the $\{1 – \cos[\Delta\varphi(x,y,t)]\}$ term in (3.7.10).

To constitute real-time holographic interferometry, the image must be recorded in real-time by a CCD sampling at a rate exceeding the vibration frequency.

Time-average holographic interferometry

In time-average holographic interferometry, a single holographic exposure is made during vibrational excitation. Naturally, the exposure period will be much longer than the vibrational period, thus a time-average view of the object field is recorded. Since an object under vibration spends most of its time near the positions of maximum positive or negative displacement (where it has zero velocity), the resulting hologram is qualitatively similar to a double-exposure interferogram, with the difference that is also displays contours of the object during the intermediate displacements between the two maxima.

Practical concerns

In order to achieve successful use of holographic interferometry, extreme care must be taken to reposition die reference hologram (in the case of real-time or time-average interferometry) and perform development of the reference hologram so as to prevent any magnification or demagnification of the reconstructed image by swelling or shrinking of the recording material. In double-exposure holographic interferometry, there are fewer sources of mechanical error, although one must take care not to displace the object other than by the intended source of stress to prevent extraneous fringing effects in the resulting interferogram. These effects are an entire science alone, but for the scope of this paper they will not be examined in great detail.

3.7.3. Applications of Holographic Interferometry

Holographic interferometry has many uses, several of which will be presented here. The major thrust of this overview will consider uses of holographic interferometry in nondestructive testing (NDT). In nondestructive testing, holographic interferometry is used to characterize deformations of various structures under stress. NDT techniques via holographic interferometry are used to measure vibration and stress-response characteristics of structures including jet engines and musical instruments. In addition, digital holographic techniques have enabled the characterization of MEMS structures to assess the fabrication process (by examining residual stress) as well as the testing of MEMS's deformation responses to thermal loading, etc. Finally, holographic interferometry is used for more exotic applications, such as detecting buried, non-metallic antipersonnel mines and evaluating the structural Integrity of aging priceless works of art.

Dynamic vibration analysis of a jet turbine compressor rotor

An aircraft engine is an excellent example of a mechanical component that requires developmental engineering for high-stress applications. To guarantee the engine can tolerate extremes in vibration as well as impulsive or continuous stress, the vibrational modes, common displacements and motion geometries of the pad must be understood. To that end, holographic interferometry has been employed to provide real-time, nondestructive metrology that not only identifies the characteristics of normal operation of a jet turbine compressor (Fig 3.7.2, 3.7.3) but also to search for hidden structural weaknesses or other anomalies/defects which would compromise the integrity of the component.

The techniques used for this study encompassed both real-time and time-

average holographic interferometry. First, using a standard holographic setup, a holographic exposure of the turbine structure was made in a stress-free state. This initial holographic image was then left in place, the structure was excited by a low-level non-contact acoustical source, and the real-time holographic image containing the superposition of the stress-free hologram and excited turbine was collected by a camera. The vibrational frequency was then swept across the range of interest and particularly strong resonances were noted.

Then, using time-average holographic techniques, interferograms of the strong resonances were recorded, and based on interference fringes produced by the vibrational stresses, mappings of the displacement responses of the part could be acquired. Fig 3.7.4, 3.7.5 show samples of interferograms of both the front

Fig 3.7.2. Turbine compressor, front view.

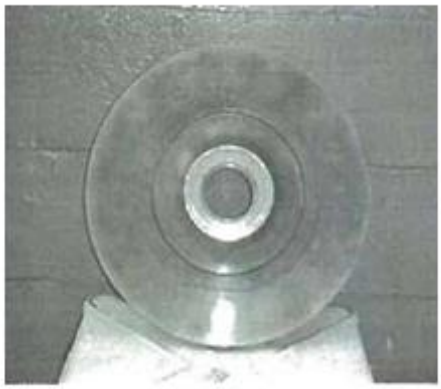


Fig 3.7.3. Turbine compressor, back view.

and back, respectively, of the turbine. The complex shapes of the interference patients demonstrate the strong nodal behavior of the part, with nodal "ridges" that divide the high-amplitude fringe groups indicating a phase change the occurs between them. These pictures also illustrate and define high-stress points in the part that might be especially susceptible to manufacturing defects.

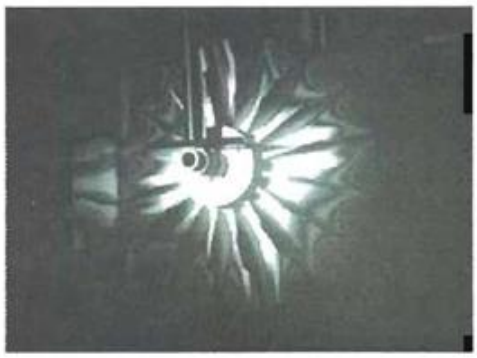


Fig 3.7.4. Holographic interferogram of a turbine compressor, front view.

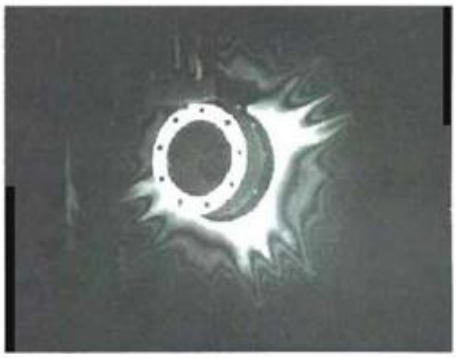


Fig 3.7.5. Holographic interferogram of a turbine compressor, back view.

Painting diagnostics

Over time, priceless paintings suffer damage due to humidity, temperature variations, and other unavoidable, everyday occurrences. More specifically, paintings are structurally composed of several layers of primers, made of mixtures of gesso and glue, and paint, all stacked upon a canvas or wood base. As a result of slowly-accruing damage to a painting, detachments between layers can occur, usually between the first layer of primer and the wood (or canvas) base. These detachments, of course, are harmful to the health of the painting, and need to be monitored to allow for identification of works which require restoration.

Double-exposure holographic interferometry is a proven method for measuring the location and severity of the detachments. Due to natural temperature and humidity variations, it is found (hat local displacements in detachments occur at a rate on the order of a few microns per minute. Therefore, by recording each of the exposures in the double-exposure holographic interferometry setup with a separation of several minutes, fringe patterns on the surface of the painting will be sufficiently dense to identify detachments. However, to heighten the sensitivity of the method, slightly warmed air is passed over the surface layers of the painting; since the detached regions of the painting will disperse heat into the canvas or wood support more slowly than «healthy» parts of the painting, these regions will stand out with higher contrast (more fringes) in a holographic interferogram recorded during the cooling phase. As shown in Fig 3.7.7, the detachments appear as dislocations in the fringe pattern of the cooling painting.

Fig 3.7.6. Holographic image of a panel painting (Santa Caterina, Pier Francesco Florentine, fifteenth century).

Fig 3.7.7. Holographic Interferogram made using thermal-drift method, revealing detachments.

3.7.4. Summary

Holographic interferometry is technique used widely for characterizing the mechanical response of structures to sources of deformation. Through double-exposure holographic interferometry, one can detect flaws in a structure while under constant stress. With real-time holographic interferometry, one can perform a rapid sweep of vibrational excitation and look for resonant frequencies. Finally, using time-average holographic interferometry, one can create high-visibilty interferograms of an object under vibrational stress for detailed study.

References

M. Franson, S. Slansky. Coherence in optics. M., Science
L.M. Soroko. Basics of coherent optics and holography. M. Science
M. Born, E. Wolf. Basics of optics. M., Science, 1970
4. R. Collier, K. Berkhard, L. Leane Optical holography. M., Mir; 1973
5.M. Miller. Holography. M., World; 1976
6. Optical holography. Under. Ed. J. Colefield, v.1,2. M.Mir; 1982
7. L. Perina. Light coherence. M. Mir, 1974
8. Yu.I. Ostrovsky. Holography and its application. M. Mir, 1973.
9. J. Strouk. Introduction to coherent optics and holography M., Science 1976
10. Kogelnik H. Coupled Wave Theory for Thick Hologram Gratings // The Bell System Technical Journal, 1969, Vol.48, No9, P.2909-2947.
11. Ryabuho, V.P. Rainbow Holograms // Physical Education in Universities. 2003. V. 9. Issue 4 Pp.88-99.
H. Coufal, D. Psaltis, G. Sincerbox. Holographic Data Storage. Springer, 2000.
P. Hariharan. Basics of Holography. Cambridge University Press, 2002.
J. W. Goodman. Introduction to Fourier Optics. McGraw Hill, 2005.
G.K. Ackermann, J. Eichler. Holography: A Practical Approach. Wiley, 2008.
V. Toal. Introduction to Holography. CRC Press, 2011.
J. Rosen. Holography. Research and Technologies. 2011.
A.L.Tolstik, Multiwave mixing in solutions of complex organic compounds. Minsk: Belarusian State University. 2002.
Vest C. M., 1979. Holographic Interferometry. New York: Wiley.

Fein H. 2003. An application of holographic interferometry for dynamic vibration analysis of a jet engine turbine compressor rotor. Practical Holography XVII and Holographic Materials IX. SPIE, 5005, 307-315.
Christnacher F., Smigielski P., Matwyschuk A., Bastide M., Fusco D., 1999. Mine detection by holography. SPIE, 3745, 361-365,
Amadesi S., Gori F., Grella R., Guartari G., 1974. Holographic methods for painting diagnostics. Appl Opt. 13, 2009-2013.
Ruiinemalm A., Molin N., 2000. On operating deflection shapes of the violin body including in-plane motions. J. Acoust. Soc. Am. 106, 3452-3459.
Coppola G., De Nicola S., Ferraro P., Finizio A., Giilli S., lodice M., Magro C., Pierattini G., 2003. Characterization of MEMS structures by microscopic digital holography. MEMS/MOEMS: Advances in Photonic Communications, Sensing. Metrology, Packaging and Assembly. SPIE 4945, 71-78.
Ferraro P., De Nicola S., Coppola G., Finizio A., lodice M., Magro C., Pierattini G., 2004. Testing silicon MEMS structures subjected to thermal loading by digital holography. SPIE, 5343, 235-243.
Schanrs U. Juptner W., 1994. Direct recording of holograms by a CCD target and numerical reconstruction. Appi Opt. 33, 179-181.
Schanrs U., 1994. Direct phase determination in holographic interferometry using digitally recorded holograms. J. Opt. Soc. Am. A 11, 2011-2015.
Ginzburg V., 1998. Some practical applications of holography in science and industry. Practical Holography XII. SPIE, 3293, 205-216.
Brown G., 1998. Thirty Odd Years of Industrial Hologram Interferometry. International Conference on Applied Optical Metrology. SPIE, 3407, 236247.
Tiziani H., Pedrini G., 1997. Digital holographic interferometry for shape and vibration analysis. International Conference on Experimental Mechanics: Advances and Applications. SPIE. 2921, 282-287.

Chapter 4. Optoelectronics

4.1. Solid state physics

4.1.1. Atomic structure of crystalline solids

Chemical bonds. All materials are compositions of one or more chemical elements (atoms) which differ by the electronic structure and hence chemical activity. Forces, which hold atoms and ions in a molecule and/or the crystal are called chemical or interatomic bonds. Interatomic interaction is caused mainly by electrostatic forces of attraction between oppositely charged particles (electrons and cations, cations and anions) and the forces of repulsion between same-charged particles (electrons and electrons, same – charged ions).

Depending on the ratio of the potential energy of chemical bonds Epot and the kinetic energy Ekin of motion of the atoms or ions at a temperature above absolute zero, all substances in nature are presented in three states (Figure 1.1.1): gaseous, when the kinetic energy of the atoms or ions is much greater than their potential interaction energy – $E_{kin} >> E_{pot}$; liquid – with $E_{kin} > E_{pot}$; solid – if $E_{kin} < E_{pot}$. This ratio of energies determines the physical state of a system of atoms or ions and its properties.

To ensure the transition of a substance from gaseous or liquid to solid state, the particles (molecules or ions) must approach to a specific (optimal) distance when the attractive and repulsive forces between them are in equilibrium. This distance corresponds to the minimum in the V(r) dependence in Fig. 4.1. The position of this minimum is exactly corresponds to the equilibrium interatomic distance r0, which is typically less than 1 nm.

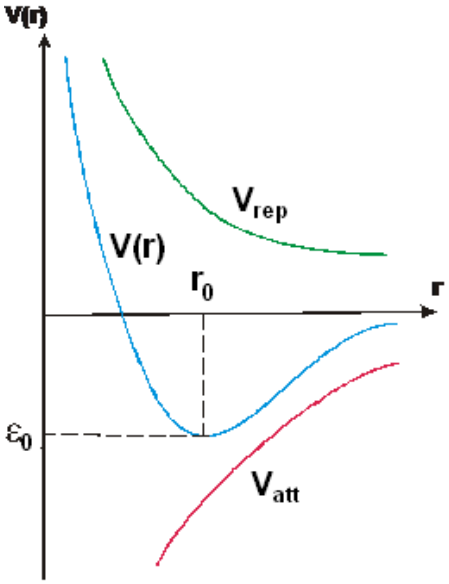


Fig. 4.1.1. The energy of the ions interaction vs interatomic distance. V(r) is total binding energy, Vrep – repulsion energy, Vatt – the energy of attraction, r0 - the equilibrium interatomic distance corresponding to the minimum binding energy.

The value of energy in the minimum $\varepsilon_0 = V(r_0)$ is just the energy of chemical bonds or the interatomic interaction energy. The value of ε0 per atom is defined as the difference between the total energy of the crystal E_{total} and energy and $N \cdot \varepsilon 0$ as in a crystal of N isolated atoms, divided by the number N of atoms:

$$\varepsilon_0 = \frac{\left(E_{total} - N \cdot \varepsilon_0\right)}{N} \qquad (4.1.1)$$

The value of this bond energy is dependent on the type of chemical bonding (see below) and for the crystalline materials is typically in the range from 0.01 to 5-7 eV.

As it was noted above, valence electrons, being on the outer electronic shells in atoms, play the main role in joining of ions in a molecules or solids. Interatomic bonding occurs because the atoms in the material are close enough to each other so that their outer electron shells begin to overlap. As a result of this overlapping, nature of the electrons motion changes dramatically and electrons located at a certain energy level of a single atom, are able to (a) pass corresponding to the energy level of the adjacent atoms without energy expenditure and thus become free to move along the solid, (b) move to another atom (forming a cation and/or anion) or (c) exchange pairs of electrons between neighboring atoms. Which of these processes is implemented, is primarily determined by the structure of the electron shells of the interacting atoms, i.e. their chemical nature. Depending on the structure of the atomic electron shells, chemical bonds between the atoms (ions) in solids are usually divided into 4 main (ultimate) forms - covalent, ionic, metallic, molecular.

The covalent bond in molecules or crystals is caused by interaction between atoms, when a pair of electrons is shared between two atoms. A major role in the covalent bonding plays a so-called exchange interaction, which has the quantum-mechanical nature. It is due to the Coulomb interaction of electrons with opposite spins, and the influence of the Pauli Exclusion Principle, which takes into account the correlation in the motion of the electrons due to the presence of spin. If diatomic molecule or crystal with a covalent chemical bonds consists of atoms of one element (e.g., H_2, N_2, diamond, Si,Ge), the distribution of electron density between atoms is symmetrical (a pair of electrons belongs to both bounded hydrogen atoms in the same degree). Molecules of this type are called by non-polar or neutral, since the centers of gravity for positive and negative charges (ions) coincide (see Fig. 4.1.2). If diatomic molecule or crystal consists of atoms of different elements, the center of the distribution of electron density in the electron pair can be displaced to one of the atoms. In this case, a covalent bond is called a polar. Molecules with a polar bond, in which the centers of positive charge do not coincide, are called polar or dipole.

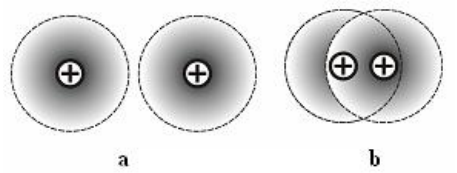


Fig. 4.1.2. Covalent interatomic bond in the hydrogen molecule: a – isolated atoms, b – a molecule with a non-polar covalent bond.

A classic example of covalent crystals are diamond, silicon and germanium. These atoms have two valence electrons in the s- and p-states. When approaching the atoms during formation of the crystal, atoms are rearranged to form 4 joint pairs of electrons, which are common to the nearest neighboring atoms. It can be seen from Fig. 4.1.2 covalent bond in such a structure.

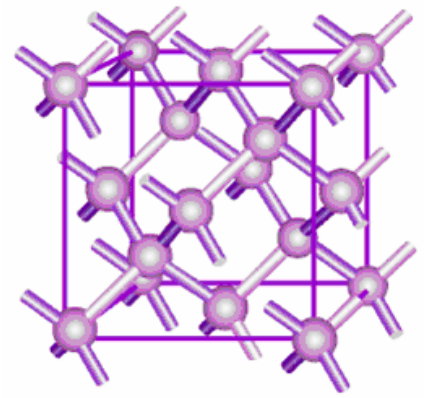


Fig. 4.1.3. Tetrahedral covalent bond in a crystalline diamond (silicon, germanium).

As can be seen from Fig. 4.1.3, electrons in such a diamond-like structure form four covalent bonds linking the neighboring atoms. Since such links are directed along axes of the regular tetrahedron, they are called tetrahedral. The distribution of electron density is found to be highly inhomogeneous, so that covalent bonds are directed along the highest densities of electrons combined. The angles between the bonds in silicon are 109o 29′. Covalent crystals usually have a very high binding energies, which can reach 5-7 eV. As a result, such crystals have the highest strength and refractoriness.

Ionic bonds are due to electrostatic (Coulomb) forces of attraction between positive and negative ions, formed by transfer of electrons from one atom to another (Figure 4.1.4).

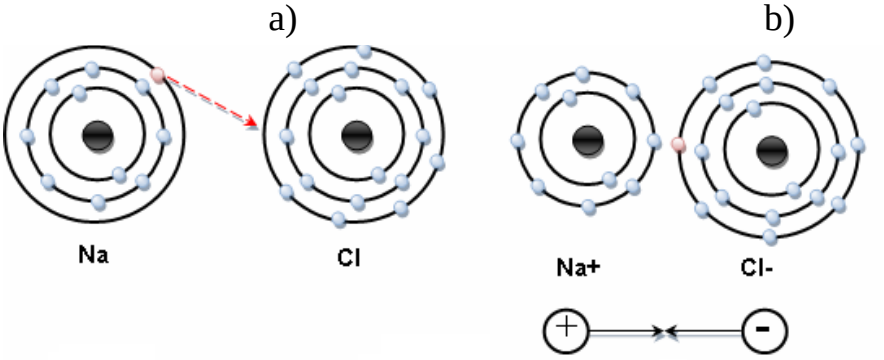


Fig. 4.1.4. Formation of ions (a) and ionic bonds (b).

The ionic bond is formed in chemical compounds in which one element is a metal and the other is close to the last group of the periodic table (for example, in

the crystals of alkali halides such as NaCl, KCl, KBr, LiF, etc., Fig. 4.1.4). In addition, such bonds are typical for many oxides and salts which are composed of ions of opposite signs (e.g., ZnO, CdO, NiO, CuO, etc.).

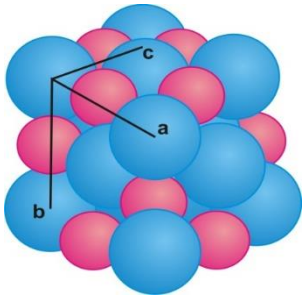


Fig. 4.1.5. Schematic representation of the anions and cations arrangement in an ionic NaCl crystal.

The ionic bond is less directed than covalent. Therefore number of nearest neighbors in the materials with ionic bonding is higher than for covalent crystals, and is usually 6 or 8. Ionic bond energy is in the range of 0.5-3 eV. Therefore, ionic crystals are less hard than covalent so that ion yield to covalent crystals in mechanical strength, fusibility and chemical durability.

Metal bonds are formed between atoms of the metal elements, which have the ability to donate valence electrons, becoming positive ions (cations). At the same time valence electrons that leave the atoms become free and called collective electrons (Figure 4.1.6). As a result, the metallic crystal can be thought of as a system consisting of positive ions, which are "immersed" in the gas of collective electrons. In this system, there is an electrostatic attraction between cations and the gas of free electrons. This structure causes omnidirectional nature of metallic bonds, leading to the possibility of forming crystals with a maximum number of nearest neighbors (up to 12).

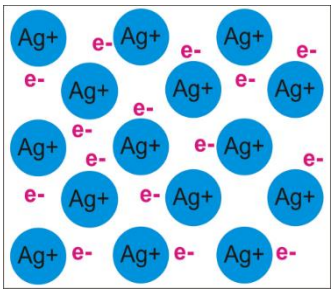


Fig. 4.1.6. A scheme of crystalline lattice with metallic bonding consisting of cations and free electron gas.

Energy of metallic bond is in the range of 1-5 eV, which determines a high mechanical strength and high melting points of metals. Due to the large concentration of free electrons metals have high electrical and thermal conductivity and also ductility.

Molecular or van der Waals bonds are formed between the individual molecules or atoms as a result of electrostatic attraction between the charges of opposite signs, which are formed during the formation of dipoles due to shifting of gravity centers of vcalence electrons and atomic nucleus. This electrostatic attraction forces called van der Waals forces. The presence of van der Waals forces is associated with the ability of neutral atoms or molecules to induce in each other instant small electric dipole moments due to fluctuations in the electron

density around the nucleus (polarization) due to collisions of atoms or other causes (Figure 4.1.7). On average, the interaction between the induced dipole moments in neighboring atoms will lead to their attraction, which is energetically favorable, as the energy of the system decreases.

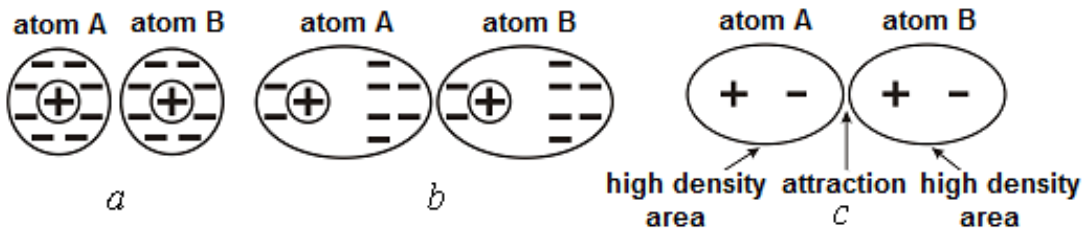


Fig. 4.1.7. The scheme of a van der Waals bonding: the initial atoms A and B (a), the atoms with polarized electron shells (b), attraction of the formed dipoles (c).

Under the influence of van der Waals forces the electrically neutral atoms of inert gases form crystals at low temperatures due to induced dipole-dipole interaction. With such forces the solid molecules of hydrogen H_2, nitrogen N_2, carbon dioxide CO_2 are also formed.

The binding energy values in crystals with the van der Waals interaction is for one or two orders of magnitude lower (0.01-0.2 eV) than in ionic ones. Therefore, such crystals have low melting and boiling points.

Crystalline and space lattice. Structural elements of solids (atoms, ions or molecules) are located either in an orderly fashion (Fig. 4.1.8a) or randomly (Fig. 4.1.8b). Substances with a regular arrangement of atoms in space, obtained under conditions of thermodynamic equilibrium, are called crystalline. Being synthesized in a non-equilibrium conditions, these substances become amorphous (highly disordered) and are characterized by a higher energy.

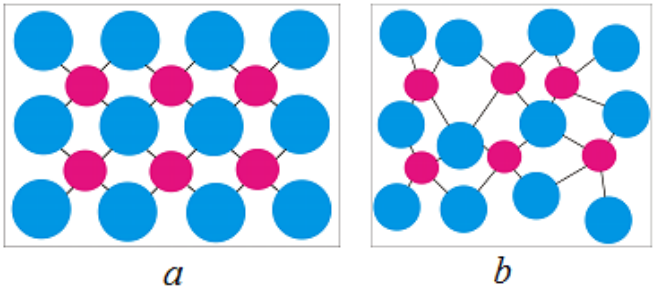


Fig. 4.1.8. Arrangement of atoms in a crystalline (a) and amorphous (b) solid material

Since the internal structure of crystalline substances is characterized by a regular periodic arrangement of their atoms, one say that the latter form a

geometrically regular periodic crystalline lattice (Fig. 4.1.8a). Due to the spherical symmetry of simple atoms, their crystalline lattice is often portrayed as a contiguous balls (Fig. 4.1.9a). However Bravais (to formalize and simplify the illustration of the atomic structure of crystals) introduced the concept of the space lattice in which the system of balls/atoms in space is presented by a scheme in which the centers of balls gravity are replaced by dots, called lattice sites (Figure 4.1.9b).

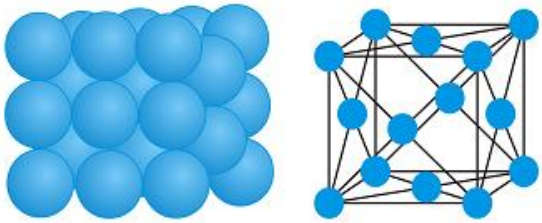


Fig. 4.1.9. Examples of crystalline (a) and space (b) lattices.

Thus, the space lattice (SL) is a system of equivalent points, or sites, put in line to crystalline lattice (CR), which reflects its basic symmetry properties. SL also can be defined as set of ordered points (sites) in the space when surrounding of every site is identical to the neighborhood of all the other sites. Since the lattice structure is the main feature of the crystals so the latter can be defined as a next. Crystal is a solid body in which the structural units (atoms, ions, molecules) are arranged regularly in the sites of SL, being grown under conditions of thermodynamic equilibrium, and have the shape of regular plane surface bodies.

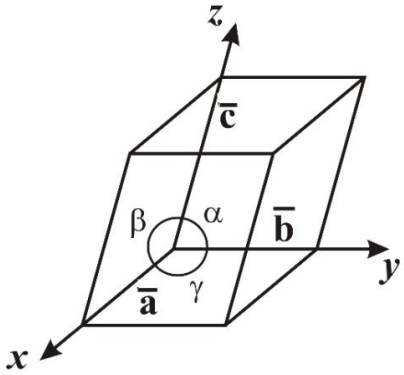


Fig. 4.1.10. An example of an oblique unit cell.

Fig. 4.1.10 shows the principle of the SL building with the help of three non-coplanar translation vectors $\vec{a}$, $\vec{b}$, $\vec{c}$, the directions of which respectively coincide with the directions of the axes x, y, z, forming a crystallographic coordinate system. The angles α, β and γ between the axes are called axial or coordinate angles. The described principle of SL construction allows to describe it mathematically. If one select any SL site as the origin, any other lattice point can be determined by the radius-vector

$$\vec{R} = m\vec{a} + n\vec{b} + t\vec{c}\,, \qquad (4.1.2)$$

called a translation vector. Here, m, n, t are random integers (from zero to infinity) which are called the site indexes. Lattice constructed using the translation operation (4.1.2) is called a simple Bravais lattice, and a parallelepiped, built into the crystallographic coordinates for the three basis-vectors, is called a unit cell of

the Bravais lattice). Modulo of the translation vectors, which determine the length of the cell edges in Fig. 4.1.10, are called lattice constants or identity periods. Lattice parameters. Lattice parameters can be determined by X-ray analysis (see below) and are measured in nanometers (1 nm = 10-9 m).

Primitive and nonprimitive cells. As can be seen from Figs. 4.1.10 and 4.1.11, structural lattice elements (ions, atoms, molecules) or sites, respectively, should be situated without fail.at the vertices of an unit cell of CD or SL. If the unit cell, based on the shortest translation vectors $\vec{a}$, $\vec{b}$, $\vec{c}$, has a lattice sites only at the vertices, it is called a primitive cell. Examples of such two-dimensional primitive cells are shown in Fig. 1.1.13a, b, c. Such cell has two important properties: it contains only a single site on one cell and has the least volume among all the elementary unit cells.

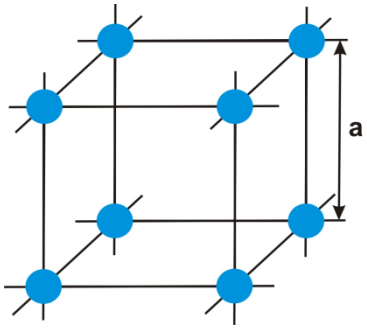


Fig. 4.1.11. Example Brava cell for a simple cubic lattice.

However, sometimes the description of the SL (and CR too) through non-rectangular elementary parallele-piped with the smallest unit cell volume (which may be oblique, as in Fig. 4.1.11b, c) is not appropriate. In this case, we can choose non-primitive cell wit the larger the volume, but rectangular, Fig. 4.1.11d,e. Such a cell contains not only the sites at the vertices, but inside itself, as well as on edges, and (or) faces. Therefore, when counting the number of atoms per unit cell,one should keep in mind that each site or atom belogs to many neighboring cells. For example, the cubic lattice shown in Fig. 4.1.11, every atom in the top of the cube, at the same time belongs to 8 cells.

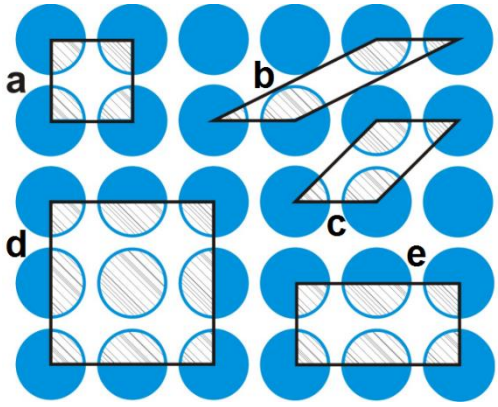


Fig. 4.1.12. Examples of primitive (a, b, c) and non-primitive (d, e) unit cells.

Closest ball packaging theory. As noted above, the spherical symmetry of the electron shells of some atoms (ions) makes it possible to represent them in the CL like a hard, non-compressible spheres of certain radius R, between which there are forces of attraction or repulsion. In this case, the atomic structure of a crystal may be concidered as an ordered spatial packing of hard spheres. The energy of such system will be minimal, if this is the closest packing.

Fig. 4.1.13 shows the case of close-packing when every ball in a flat bottom (first) layer of the balls is surrounded by six other similar R balls and six, respectively, triangular voids. At the construction of the second (upper) layer of close-packed spheres, we put balls in the voids of the same type. After packing the first two balls layers, two types of voids (pores) are generated - tetrahedral and octahedral. Tetrahedral pore is surrounded by 4 balls (numbers 1-4 in Fig. 4.1.13 and) of the two layers, the centers of which form a tetrahedron. Octahedral are through pores and each of them is surrounded by 6 balls (3-7 in Fig. 4.1.13) and by 3 balls in every layers, which are rotated by 60° relative to each other. The centers of the balls form an octahedron. The tetrahedral and octahedral pores have different volumes, which are typically characterized by a maximum size (radius r) of a ball which can be inscribed therein.

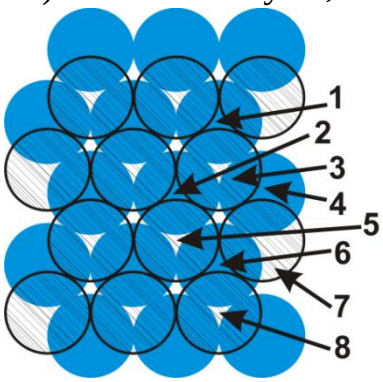
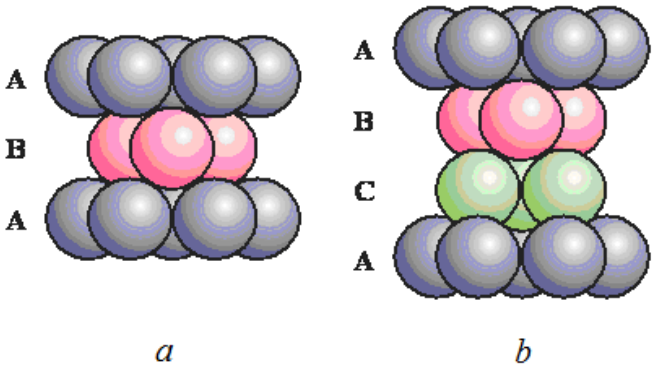


Fig. 4.1.13. Closest packing of two layers of balls. Balls of the second (upper) layer are shaded.

If the balls of the third layer are laid on the tetrahedral voids of the second layer, their centers will be located just above the centers of the balls of the first layer. Such close-packed structure is double-layered and is called the hexagonal close packed (hcp), with alternating layers of type ABABAB ... (Fig. 4.1.14a).



a *b*

Fig. 4.1.14. The hexagonal (a) and cubic (b) close packings.

If the balls of the third layer (layer C) are laid on the octahedral voids, and the balls of the 4th layer directly over the balls of the first layer, then this package is called a trilaminar or a cubic close packing with alternating layers of type ABCBCA... (Fig. 4.1.14b).

The main types of crystal structures. The nature of packaging (type of lattice) for each solid material depends on the electronic structure of its constituent atoms and type of chemical bond between them. Examples of metals have often found these types of crystal structures (types of unit cells) as a body-centered cubic (*bcc*) – Fig. 4.1.15a, face-centered cubic (*fcc*) - Fig. 4.1.15b and hexagonal close-

packed (*hcp*) - Fig. 4.1.15c. As shown, the elementary bcc cell is not primitive because contains 2 atoms for one cell, not only in the vertex 8 carbon atoms, but one atom in the center of the cube. In the fcc lattice every unit cell (also non-primitive) contains 8 carbon atoms at the vertices and 6 atoms in the centers of the faces of the cube (i.e., 4 atoms per unit cell). In the hcp lattice unit cell has the form of a hexagonal prism and contains 12 atoms at the vertices, 2 atoms at 2 faces and 3 atoms inside the prism (ie, 6 atoms per unit cell).

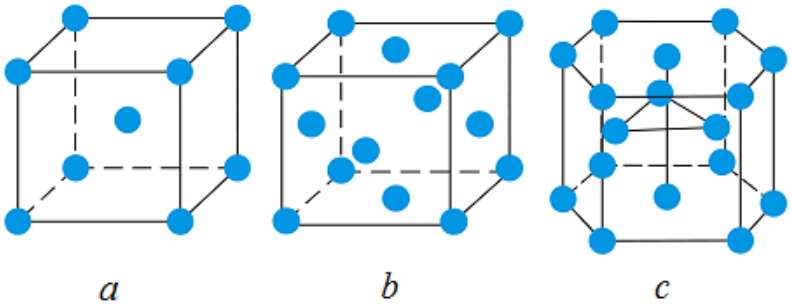


Fig. 4.1.15. The main types of crystal (spatial) lattices of metals: a – *bcc*, b –*fcc*, c – *hcp*.

In the case of cubic (bcc and fcc) structures all the edges of the cell (the length of translation vectors) are the same and form right angles to each other (Fig. 4.1.15a, b). Therefore, the parameters of cubic lattices are characterized by long edges of the cube and are denoted by the letter a. To characterize the hcp lattice we take two parameters – a hexagon side (along the x-axis) and height of the prism (along the axis z). When the ratio c/a = 1.633, the atoms are packed most tightly, forming the hcp structure. Some metals have a hexagonal lattice with a less dense packing of atoms: for example, for zinc c/a = 1.86, for cadmium c/a = 1.88.

The simplest structures of semiconductors are *diamond-like* (Fig. 4.1.16a), *sphalerite* (Fig. 4.1.16b) and *wurtzite* (Fig. 4.1.16c). The diamond-type structure is typical for diamond (consisting of carbon atoms), Ge and Si. The unit cell of a diamond refers to the *fcc* structure.

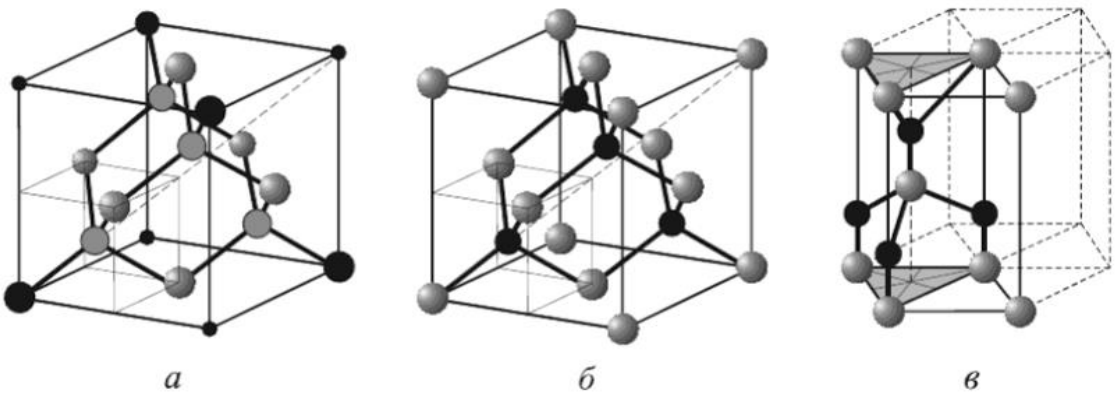


Fig. 4.1.16. Examples of unit cells for the diamond (a), sphalerite (b) and wurtzite (c) structures.

However, in addition to atoms at the 8 vertices and 6 centers of the cube faces, diamond crystal cell contains 4 extra atoms (gray balls are shown in Fig. 4.1.16a). These atoms are in 4 of 8 octants in which fcc unit cell can be divided in Fig. 4.1.16a. As a result, such cell contains 8 carbon atoms, and each of this atoms adjoins with 4 nearest neighbors (tetrahedral neighboring).

The sphalerite-type structure (ZnS, CdS, AlP, AlAs, GaAs, etc.) looks similar to diamond-like (Fig. 4.1.16b). However, 8 atoms belonging to the cell are divided into two classes: 4 atoms of one type belong the vertices and centers and 4 atom of another sort are in the centers of 4 (of 8) above mentioned octants.

Wurtzite structure (CdS, ZnS, ZnO, InSb, etc.) corresponds to the hcp lattice (Fig. 4.1.16c). However, in the unit cell of this structure there is an additional number of atoms of a different sort on the edges and inside the hexagonal prism. For this structure is also typical a tetrahedral environment.

Knowing the type of the lattice, we can determine a coordination number and atomic (or ionic) radius for the structure.

Coordination number of the lattice structure is the number of the nearest neighboring atoms surrounding every atom. As can be seen, in a bcc lattice coordination number is 8. For *fcc* and *hcp* lattices coordination number is 12. The atomic (or ionic) radius in the crystal lattice is defined as half the distance between the centers of the balls-atoms (ions) belonging to the edge on unit cell (Fig. 4.1.16).

The close packing model allows us to calculate the fill factor or the compactness Q of the crystal structure. It is defined as the fraction of the volume of the unit cell occupied by n atoms or ions of radius R:

$$Q = \frac{4\pi R^3 n}{3V} \cdot 100\% \quad . \qquad (4.1.3)$$

Here, V is volume of the unit cell, n – number of atoms per cell. According to (4.1.3), for a simple cubic lattice Q = 52 %, for the *bcc* structure – 68 %, and for the *fcc* and *hcp* lattices – 74 %. Thus, among these four structures the fcc and hcp lattices are the most closely packed. The changes in fill factor of lattices result in changes of interatomic pores dimensions, that is important for incorporation of foreign atoms into the crystal lattice (for example, the formation of complex compounds or alloys). As is seen, in the *fcc* and *hcp* lattice atoms occupy 74 % of

the total volume of the crystal lattice, and interatomic voids – 26 %. In *bcc* lattice atoms occupy 68 % of the total volume, and nly 32 % – pores.

From geometrical considerations, we can easily find that the radius r of the octahedral pores in the close-packed *fcc* and *hcp* lattices, having the same coefficient of compactness, is 0.41 of atom (ion) radius R, and the radius of the tetrahedral one is only 0,22R. Since the bcc lattice has a lower Q values, so octahedral pore radius therein is only 0.154R. In the same tetrahedral pores of bcc structure atoms can be held with a radius of 0,29R. In accordance with the theory of the close packings, spherical monoatomic crystal structure should be formed by hexagonal (hcp) or cubic (fcc) laws, as it is observed in many crystals of inert gases and the most metals. Deviations from the close packings (for example, the appearance of more loose bcc or diamond-like structures) may be associated with features of the interatomic interaction (the nature of the chemical bonds) in crystals. In non-monoatomic crystals, most large ions (usually anions) are arranged in the space under the laws of the close packing, whereas smaller ions (cations) are situated in the octahedral pores, such as in the structure of NaCl.

Methods for atomic structure description. As can be seen above, the unit cells of solid crystalline materials can fit the form of a rectangular and oblique parallelepipeds. Therefore, to descript the crystal structures with different kinds of unit cells, special crystallographic indexing method has been developed. It allows to describe the SL (CL) in the symbolic form. This method allows to describe uniformly positions of sites (atoms), and also specific directions and planes in lattices. This metod is irrespective to the crystallographic coordinate system (rectangular or oblique), and specific values of modulo for basic translation vectors (the lattice constants) in the coordinate axes.

In the crystallographic indexing method for describing the atomic (or site) coordinates in the lattice one can use the Eq. (4.1.2), where the radius vector is a linear combination of the basic translation vectors $\vec{a}$, $\vec{b}$, $\vec{c}$. In this case, the set of three indices in Eq. (4.1.2), written in double square brackets, gives the coordinates of every site. The [[mnt]] is called the symbol of the site. Examples of such symbols for the vertex atoms of the primitive unit cell are shown in Fig. 4.1.17a.

In crystallography all directions are passed (in the form of straight lines) through the sites (atoms) of the lattice. Therefore, every crystallografic direction in the crystal is always possible to characterize by the coordinates of two sites (atoms), through which it passes. If one of the sites is chosen as the origin, the coordinates of the second site will determine the preferred direction. Therefore,

the second site coordinates [mnt], recorded in single brackets, can be called a symbol of direction. The same symbol, enclosed in curly braces, describes the whole family of directions (lines) parallel to this. Some examples of the major crystallographic directions are shown in Fig. 4.1.17b.

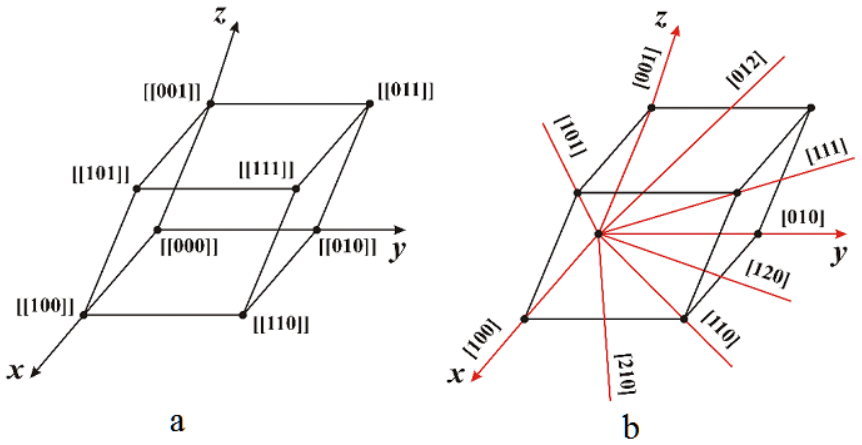


Fig. 4.1.17. Crystallographic simbols of sites (a) and directions (b) in the primitive unit cell.

Every crystallographyc plane also should pass through a system of the space lattice sites (see Fig. 4.1.18a). In this case, any plane that passes through the three lattice sites, that do not lie on a straight line, contains an entire grid of sites. In this case, to describe the position of every crystallographic plane (and system of parallel planes too) in the space, one should define a set of the intercepts, which are cutted off by the crystallographic plane on the coordinate axes x, y, z. According to Fig. 4.1.18a, these intercepts are equal ma, nb, tc, where the numbers m, n, t determine the lengths of the intercepts (in fractions of the modulo of unit translation vectors $\vec{a}$, $\vec{b}$, $\vec{c}$). Fig. 1.1.20a a system of parallel planes is given, where the first (the nearest to the origin) is characterized by m = n = 1/2 and t = 0. Fig. 4.1.18b shows two planes, labeled as (111) and (112), which cut the axes with intercepts having indices m = n = t = 1 (the non-shaded plane) and m = n = 1 and t = 1/2 (shaded plane).

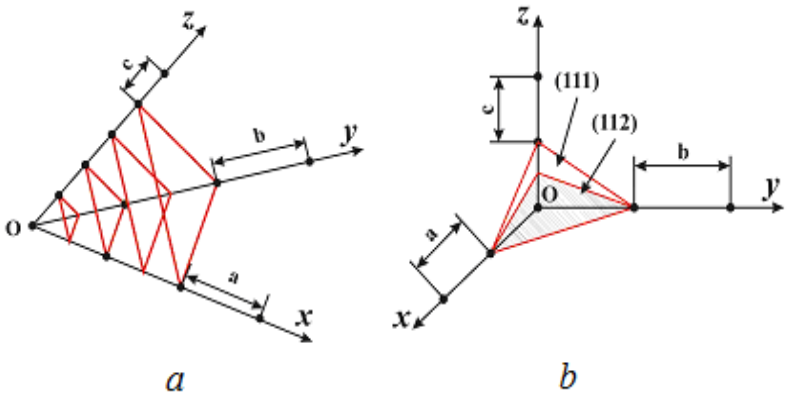


Fig. 4.1.18. Examples of the crystallographic planes construction in the space lattice: a – (221)-type plane intersects the axes x and y in the points with coordinates [[1/2,0,0]] and [[0,1/2,0]], respectively; b – planes (111) and (112) intersect the z-axis at the points [[0,0,1]] and

[[0,0,1/2]], respectively. The coordinates are given in fractions of the modulo of unit translation vectors $\vec{a}$, $\vec{b}$, $\vec{c}$.

However, in crystallography for convenience, the spatial arrangement of the planes is not characterized by *m, n, t* numbers or intercepts by the axes. They are usually characterized by the so-called crystallographic Miller indices. Miller indices (hkl) are enclosed in parentheses and indicate the reciprocal lengths of the intercepts, measured in units of the lattice parameter, namely - $h = 1/m$, $k = 1/n$, $l = 1/t$. Therefore for the planes in Fig. 4.1.18a, Miller indices are (221). In Fig. 4.1.18b the non-shaded plane is (111) and the shaded plane – (112).

If the plane cuts off negative intercepts by any axis, a minus sign is put above the corresponding index.

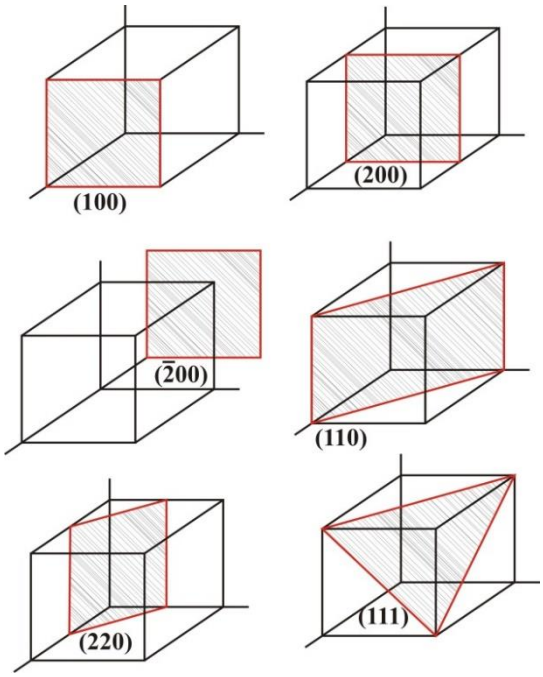
It follows that the geometric meaning of Miller indices is that the system of parallel planes wit the indices (*hkl*) dissects identity periods (or unit cell edges) *a, b, c* on *h, k, l* portions, respectively. This means that the symbols of the planes, shown in Fig. 4.1.19b, will (111) and (112). Fig. 4.1.19 shows images of major planes and their corresponding Miller indices for cubic unit cells.

The interplanar spacing d is the main characteristic of crystallographic planes family.

In particular, for the crystals with cubic symmetry d is defined by the relation

$$d^2 = \frac{a^2}{h^2 + k^2 + l^2}, \qquad (4.1.4)$$

where a is the lattice period (lattice constant).



To determine the atomic structure of solids the diffraction of X-rays, electrons and neutrons is used. All these methods are based on the general principles of the diffraction for waves on crystalline lattice if their wave lengths are close to the average interatomic distance in the material.

Fig. 4.1.19. (100), (200), ($\bar{2}$00), (110), (220) and (111) planes in the unit cell.

Defects in crystalline lattice. At thermodynamic equilibrium, the location of particles (atoms, ions, molecules) in a perfect crystal is characterized by a strict three-dimensional periodicity. However, real crystals, due to violation of the thermodynamic equilibrium conditions in the process of their production, are always contain distorted parts of the periodic lattice.

Any local deviation from the periodic structure of the crystal is called *the defect* of crystal lattice. These distortions are characterized by changes in the coordination of the atoms, a violation of the lengths and angles of interatomic bonds, the introduction of foreign atoms, the formation of foreign phases, etc. The presence of such defects always distorts the whole region of the crystal lattice around defect.

Low mobility and a large (almost infinite) lifetime of structural defects under normal conditions allow one to describe them using the visual geometric images. Classification of defects is usually carried out by the number of dimensions in which the distortions of the crystal structure are extending over distances exceeding the characteristic parameters of the lattice. Such approach allows to separate all defects on four classes: point (zero-dimensional), linear (one-dimensional), surface (two-dimensional) and bulk (three-dimensional).

Point defects. Under point defects we shall understand the isolated from each other damages of periodicity which are extended on a few interatomic distances in all three crystallographic directions. The so-called intrinsic point defects of the crystal lattice include *vacancies* (Fig. 4.1.20a) and own *interstitial atoms* (Figure 4.1.20b).

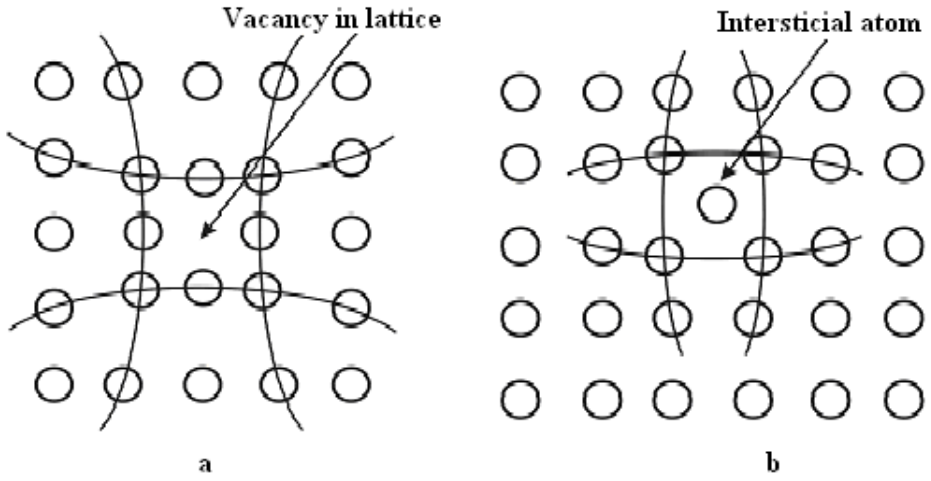


Fig. 4.1.20. Schematic representation of the vacancies (a) and interstitial atoms (b).

Impurity atoms in the crystal lattice are important types of point defects. They can substitute the lattice sites, defects 1 and 2 in Fig. 4.1.21a, or be incorporated into intersticials of the lattice, Fig. 4.1.21a-3. Simple point defects can interact creating their combinations, such as vacancy-interstitial atom (Frenkel defect), the vacancy-atom on the surface (Schottky defect), double and triple vacancies, vacancy-impurity, etc.

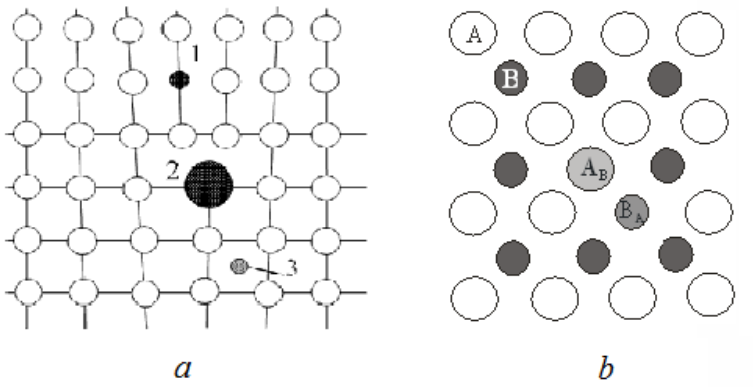


Fig. 4.1.21. Schematic representation of the simple point defects in a single-atom (a) and diatomic (b) crystal lattices. In a: 1 – impurity atom substitution of short-range, 2 – impurity atom substitution of long-range, 3 – introduction of an impurity atom. In b: A, B – atoms of different sorts; A_B, B_A – antisite defects.

There are the so-called antisite defects which occur in chemical compounds AB, whose lattice consists of two sublattices of type A and B (see Fig. 4.1.21b). Antisite defect A_B (B_A) is a point defect, which is obtained when atom A (B) occupies the site of sub-lattice of B (A) atoms.

Once again, we emphasize that since the introduction of a point defect is always a distortion of the crystal lattice in its vicinity (see Fig. 4.1.20), the concept of a point defect is the whole area of the distorted crystal lattice.

Point defects usually arise due to impact of heat or irradiation. The energies of the point defects formation are typically a few electron volts: ~ 2 eV for vacancies in Germanium, ~ 2.3 eV for the vacancy in silicon, ~ 1 eV for vacancies in copper, ~ 2.4 eV for the interstitial atoms in copper, ~ 2 eV for Schottky defects for NaCl and ~ 1.5 eV for Frenkel defects for NaCl.

Linear defects. Linear defects include damages of periodicity in the crystal lattice, which are extended on much interatomic distances along one crystallographic direction (axis), and along other directions – only on some of the interatomic distances (as for point defects). There are two main types of linear defects – edge and screw dislocations. Besides the chains of nano-and micro-cracks, and point defects are also related to linear defects. Linear defects can be both thermal and mechanical (generated as a result of plastic deformation) nature.

Edge dislocation presents linear range of lattice distortions, which occur formally along half-plane incorporated into the crystal lattice. Really edge dislocation is formed due to the incomplete shift of top part of the crystal elative the low one upon application of shear stress τ, perpendicular to the edge of this extra plane (dislocation line). The edge of this half-plane will coincide with the line of the edge dislocation. If the shift was not pass to the end of the crystal, dislocation can be represented as a certain distorted region of the crystal lattice around the edge of extra plane (shown by dotted line in Fig. 4.1.22). As follows from Fig. 4.1.22, the dislocation line (edge of extra plane) is perpendicular to the shear direction.

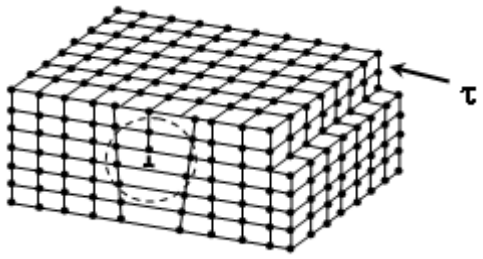


Fig. 4.1.22. Three-dimensional diagram of the crystal lattice with an edge dislocation arising from the action of shear stress τ, perpendicular to the dislocation axis (line).

If the extra plane is at the top part of the crystal, the dislocation is conventionally called positive if at the low part – then negative. Sign of the dislocation allows to evaluate the character of their interaction. Experiments show that the dislocations of the same sign repel each other, but the opposite – attract. The attraction of two dislocations of opposite signs accompanied by their annihilation (disappearance of defect region), as the two half-planes form a complete plane.

Screw dislocation is a distortion region in the crystal lattice, which occurs in the crystal due to subjection to shear stress τ, parallel to the line B-B` (see Fig. 4.1.23a). It is accompanied by a partial (on one interatomoc distance) shift of

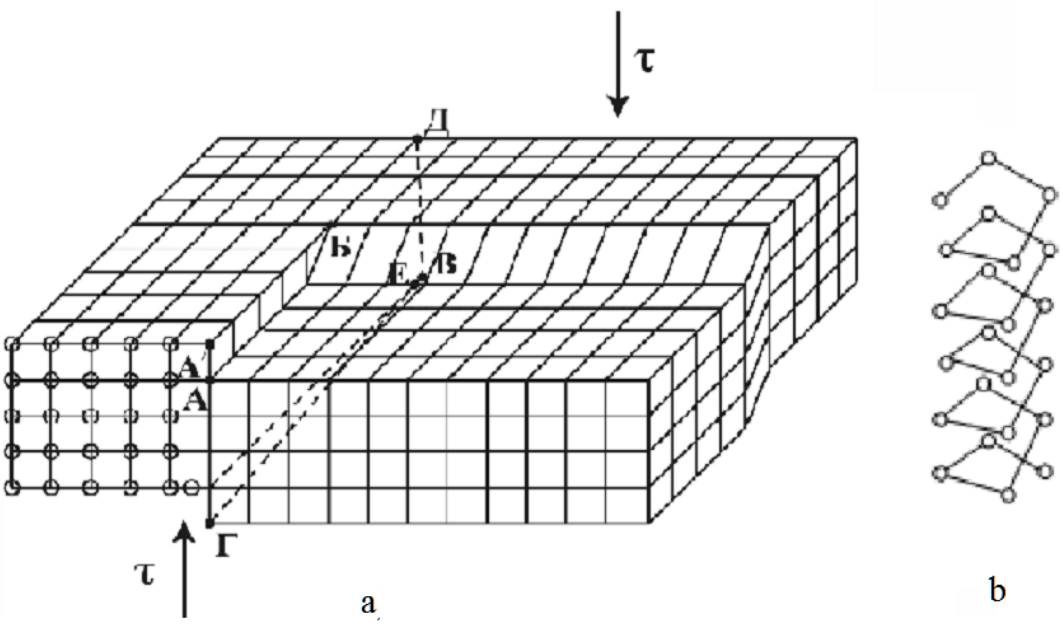


Fig. 4.1.23. Scheme of screw dislocations (a) and the arrangement of atoms along the dislocation line (b) under shear of two parts of the crystal relative to one another.

the atomic layers in the plane AA`BB` (Figure 4.1.23a). The atoms of the crystal lattice, which are shifted from their equilibrium positions along the axis B-B` (called the screw dislocation line), are arranged by twisted (spiral-like) curve along the dislocation line (Figure 4.1.23b).

The dislocation line separates shifted part of the crystal from the part of the crystal, where the shift has not yet occurred. Screw dislocation, formed by rotating clockwise, called the right and counterclockwise – left.

As follows from Fig. 4.1.23, crystal lattice is distorted around the dislocation line, causing the formation of the stress fields of compressing and stretching.

Screw dislocations are formed in the crystallization process, and also due to plastic deformation and phase transformations. By the density of dislocations we usually understand the total length of dislocations per unit volume.

The lattice distortions (stress fields) around dislocations results in attraction of imputity atoms. Accumulation of impurity atoms around the dislocations form the so-called Cottrell atmospheres or clouds that hinder the movement of dislocations during plastic deformation (see, below).

Two-dimensional defects. Two-dimensional or surface defects in the crystal are extended along two directions on distances which are many times greater than the typical values of the lattice parameter, while in the third direction – only a few interatomic distances. Surfaces of the crystals, stacking faults, and internal surfaces (interfaces), like grain boundaries and interphase boundaries are related to two-dimensional defects. Stacking faults occur when hexagonal and cubic crystalline layers alternate in the closed-packing structure (e.g., ABCABCABABC ...), Fig. 4.1.24a.

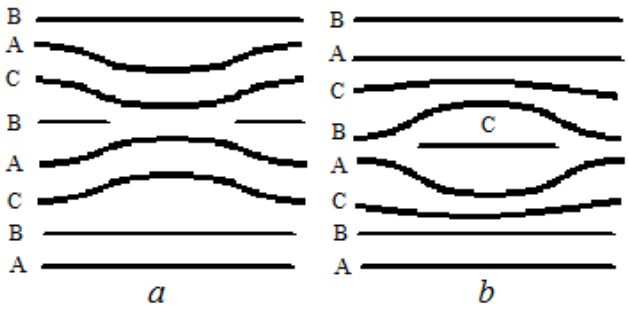


Fig. 4.1.24. Stacking faults.

Interfaces are typical places of crystal lattices matching neighboring grains in the polycrystalline material (Fig. 4.1.25b) or interphase boundaries in multiphase alloys (Fig. 4.1.25a).

The structure of the grain boundary depends on the angle of misorientation of the crystal lattices in the neighboring crystallites. Boundaries with the large misorientation angles represent crystallographicly strongly mismatched area S of two neighboring lattices (Figure 4.1.25b). In some cases, part of atoms at the grain boundary may belong to both lattices (for example, atom A), while the other atoms can be unmatched (B atom).

The structure of the grain boundary exerts a great influence on the properties of the crystals. Grain boundaries interact with dislocations being within the grains. Depending on the location of dislocation relative to the boundary and their sign, forces of attraction and repulsion can occur. Grain boundaries also attract point defects located at a few interatomic distances. Atmospheres of attracted impurity atoms (Cottrell atmospheres) inhibit the migration of boundaries.

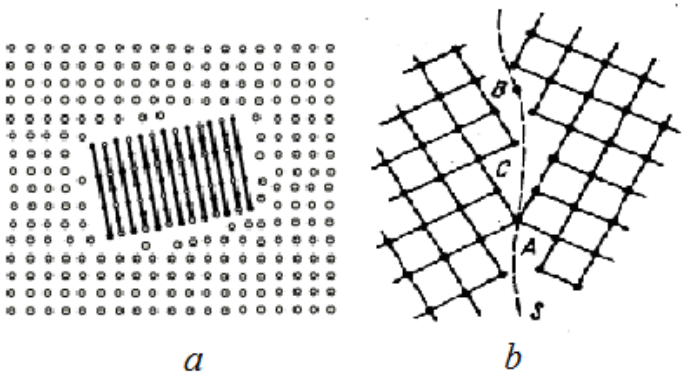


Fig. 4.1.25. Interphase (a) and intergranular (b) interfaces

Two-dimensional defects arise typically during the crystal growth due to violation of local thermodynamic equilibrium conditions.

Bulk defects. Bulk defects include micro- and nanopores, and also the inclusions of the other phases (Figure 4.1.25a), if their dimensions are much higher than the typical values of the lattice parameters in all three crystallographic directions. Bulk defects arise typically during the crystal growth due to violation of local thermodynamic equilibrium conditions or at intensive plastic deformation, heat treatment and other impacts.

4.1.2. Atomic dynamics

The oscillatory nature of the thermal motion of atoms. At finite temperatures, the atoms of the crystal lattice are always oscillate around their equilibrium positions – SL sites. As a result, in a crystal, being in thermal equilibrium with the environment, a steady state in the form of standing or running waves is established.

The character of these oscillations depends on the crystal symmetry (structure), the number of atoms in the unit cell, the type of chemical bonds, as well as the type and concentration of lattice defects. Displacements of atoms from the equilibrium positions during these oscillations is greater the higher the temperature, but they are much smaller than the lattice parameter up to the melting temperature when transforming of solids into liquid state.

The forces, that tend to hold the atoms in the equilibrium positions, as a first approximation can be considered proportional to relative displacements of atoms from equilibrioum positions as if they were connected to each other by elastic springs (Figure 4.1.26). Representation of the crystal in the form of a set of particles linked with perfectly elastic forces is called a *harmonic approximation*.

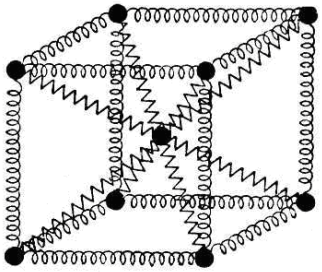


Fig. 4.1.26. Presentation of the bcc crystal in the form of a set of atoms bound to each other by spring bonds.

The harmonic approximation leads to the fact that a system of elastic waves, connected with displacements of atoms from their equilibrium positions, propagates in the lattice. In a crystal consisting of N identical atoms, there are 3N such waves that are called *normal* (or *intrincic*) *oscillations*, or *modes*.

Reciprocal lattice. In order to describe a system of waves (elastic, electron, electro-magnetic) in a crystal, the concept of the *reciprocal lattice* is used. This

concept is very useful in the analysis of many phenomena caused by the particle-wave nature of atoms, electrons, photons, etc. propagating in a periodic lattice.

The reciprocal lattice is closely related to the direct lattice. Relationship between the basic translation vectors $\vec{a}$, $\vec{b}$, $\vec{c}$ of normal (or direct) lattice and the basic translation vectors of the reciprocal lattice can be expressed as

$$\vec{a}^* = \frac{[\vec{b}\vec{c}]}{V},\ \vec{b}^* = \frac{[\vec{c}\vec{a}]}{V},\ \vec{c}^* = \frac{[\vec{a}\vec{b}]}{V}, \qquad (4.1.4)$$

where $V = \vec{a}[\vec{b}\vec{c}]$ is the volume of the cell unit in direct lattice.

There is a relationship between the crystallographic planes of the direct lattice, determined by Miller indices (*hkl*), and the vector $\vec{G}^*$ of the reciprocal lattice, which is a linear combination of the basic translation vectors $\vec{a}^*$, $\vec{b}^*$, $\vec{c}^*$. This relationship is defined by equation

$$\vec{G}^*_{hkl} = h\vec{a}^* + k\vec{b}^* + l\vec{c}^*. \qquad (4.1.5)$$

As follows from (4.1.4) and (4.1.5), the $\vec{G}^*$ vector is always perpendicular to the (*hkl*) plane and its modulo is proportional to the inverse interplanar spacing $1/d_{hkl}$ between the equivalent {*hkl*} planes of the direct lattice (see Fig. 4.1.27).

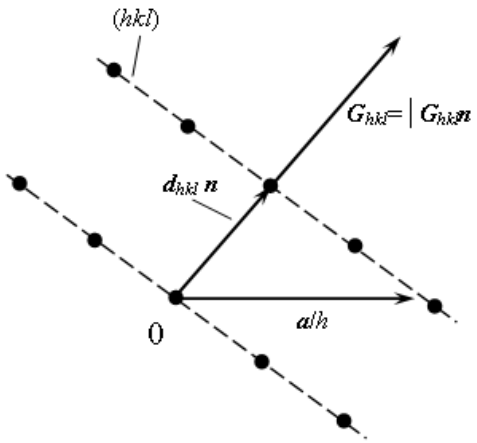


Fig. 4.1.27. To the reciprocal lattice definition.

The definition (4.1.5) of the basic translation vectors of the reciprocal lattice gives the next important rules:

The heteronymous main translation vectors of both lattices are always perpendicular to each other so that

$$(\vec{a}^*\vec{b}) = (\vec{a}^*\vec{c}) = (\vec{b}^*\vec{c}) = (\vec{c}^*\vec{a}) = (\vec{c}^*\vec{b}) = (\vec{b}^*\vec{a}) = 0. \qquad (4.1.6)$$

Having the same names basic translation vectors of both lattices are always parallel because

$$\left(\vec{a}^*\vec{a}\right)=\left(\vec{b}^*\vec{b}\right)=\left(\vec{c}^*\vec{c}\right)=1 \qquad (4.1.7)$$

The product of the primitive cells volumes of both lattices is equal to unity:

$$VV^*=1, \qquad (4.1.8)$$

where $V^*=\vec{a}^*[\vec{b}^*\vec{c}^*]$.

It can be shown that the *bcc* lattice has a reciprocal lattice type *hcc* (and vice versa) with a side of the cubic cell $2\pi/a$. For lattice primitive cell, both are the same.

Laue equations (interference condition). The reciprocal lattice concept makes it possible to establish a number of important relations for description of wave diffraction in crystals. In particular, Max von Laue received the equation, which makes it possible to determine the position of the interference maxima occurring at scattering of waves (X-ray, electron, elastic waves of atomic displacements, etc.) on the crystal lattice. In doing so, this equation does not use the known Wolfe-Bragg assumption concerning mirror-like character of X-ray waves reflection from atomic planes. Derivation of the Laue equation is based on the following assumptions:

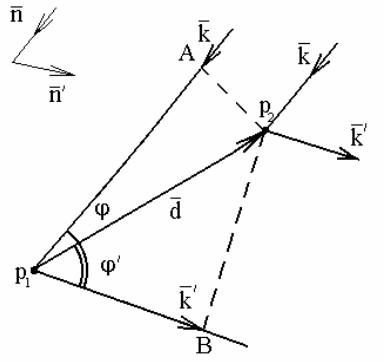
The crystal consists of a set of identical centers which are placed at the spatial lattice sites and scatter the radiation (X-Ray, electron and elastic waves).

Every center can scatter elastically the incident radiation (waves) in all directions.

The sharp peaks of intensity, which take place due to scattering of monochromatic radiation with wavelength λ by these centers, are observed only in those directions (for those angles θ), for which the rays, scattered by all lattice sites, are enhanced owing to the interference.

Laue equation can be derived from the condition that the path difference of the rays scattered from two different centers P_1 and P_2 (Figure 4.1.28) equals to a whole number $m\lambda$ of wavelengths. In accordance with Fig. 4.1.28, this condition can be written as equation

$$\Delta=\mathrm{P_1A}+\mathrm{P_1B}=\left(\vec{d}\left(\vec{n}-\vec{n}'\right)\right)=m\lambda\,, \qquad (4.1.9)$$



where $\vec{d}$ is the radius-vector, which characterizes the position of the centers P_1 and P_2 relative to each other, $\vec{n}$ and $\vec{n}'$ - the unit vectors showing the propagation direction (the direction of the wave vectors) of the incident and reflected waves, *m* - an integer (the order of reflection).

Fig. 4.1.28. For the derivation of the Laue equation.

Given that the wave vectors ($\vec{k}$, $\vec{k}'$) and the unit vectors ($\vec{n}$, $\vec{n}'$) for the incident and scattered waves respectively are related as $\vec{k} = \frac{2\pi\vec{n}}{\lambda}$ and $\vec{k}' = \frac{2\pi\vec{n}}{\lambda'}$, Eq. (4.1.9) can be easily converted to a relation

$$\left(\vec{d}\Delta\vec{k}\right) = 2\pi m, \qquad (4.1.10)$$

where $\Delta\vec{k} = \vec{k} - \vec{k}'$ is called *the scattering vector*.

Taking into account relations (4.1.9) and (4.1.10), for the reciprocal lattice vectors, equation (4.1.10) can be easily reduced to an equation

$$\Delta\vec{k} = \vec{G}^*_{hkl}, \qquad (4.1.11)$$

which is called *interference Laue equation*. Fig. 4.1.28 shows also that Eq. (4.1.9) can be easily reduced to the known Wolfe-Bragg condition $2d\sin\theta = n\lambda$.

Thus, according to Laue, diffraction peak, arising from the change of incident wave vector on the value of reciprocal lattice vector $\boldsymbol{G^*_{hkl}}$ at the wave scattering by a crystal lattice, is equivalent to Bragg reflection from a family of direct lattice atomic planes {*hkl*} normal to the scattering vector $\Delta\vec{k}$.

The dispersion laws. The nature of the atom oscillations in crystals is determined by the attraction and repulsion forces which links them. The energies of atomic attraction U_{att} and repulsion U_{rep} are differently dependent on the distances *r* between atoms (dot-dashed lines in Fig. 4.1.29) in accordance with relations

$$U_{\text{ï}\delta} = -\frac{C_1}{r^m}; \qquad U_{\text{î}\text{ò}} = \frac{C_2}{r^n} \qquad (4.1.12)$$

(C_1 and C_2, m and n are constants, characterizing the kind of interaction). So the total energy of interatomic interaction $U(r) = U_{att} + U_{rep}$ depends on the interatomic distance and can be represented by a curve with a minimum (solid curve in Fig. 4.1.29. The value of $U_0 = U(r_0)$ at the minimum point of the curve $U(r)$ is none other than *the bonding energy*.

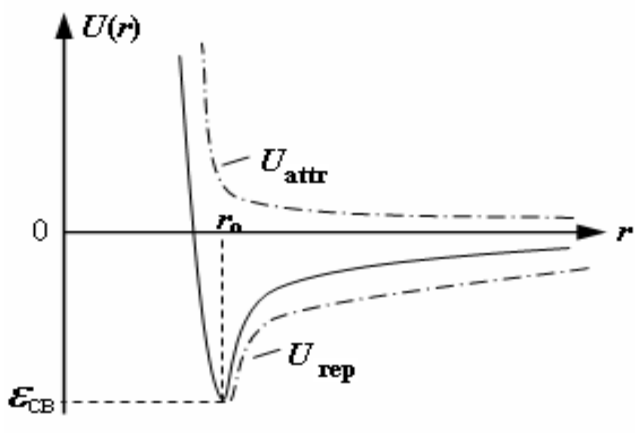


Fig. 4.1.29. The dependence interaction energies $U(r)$ of the atoms on the interatomic distance r: r_o is atomic position in the equilibrium state; $U_{att}(r)$ – the energy of attraction; U_{rep} (r) – the repulsion energy; $\varepsilon_{св}$ – chemical bond energy of the interacting atoms.

If this energy is high enough, the system of atoms forms a stable solid state structure when the atoms are at some distance r_0 from each other. Such bonded atoms have possibility only to fluctuate around their equilibrium positions. Formally, these oscillating atoms in the crystal can be presented as balls linked with elastic springs (Fig. 4.1.30). The presence of these spring-like bonds can transfer oscillations from one atom to another.

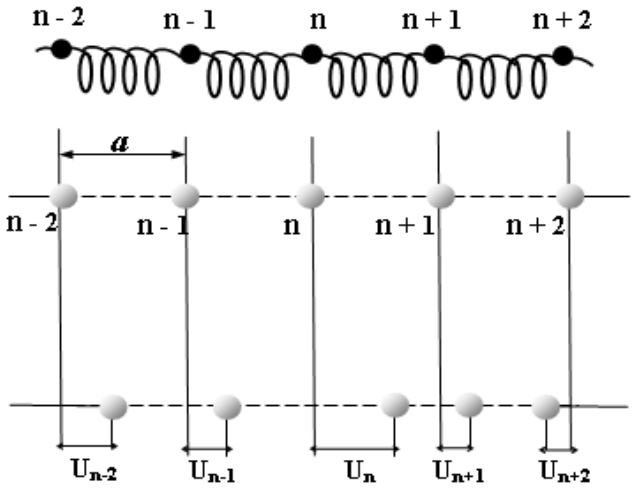


Fig. 4.1.30. The one-dimensional chain consisting of atoms of one kind.

As a result, at finite temperature in such a system interconnected, collective vibrational motion of the atoms is established, which is equivalent to propagating through the crystal of elastic waves due to atomic displacements. Such waves are called *normal oscillations* of the crystal lattice.

Consider the characteristics of these elastic waves for the one-dimensional chain of atoms (of the same kind) as the model of crystal.

The dispersion law for the elastic displacement waves in the one-atom chain. Let us consider the atomic vibrations in some frequency range on the example of a linear chain of N atoms with the length $L = Na$ (a is the lattice parameter), shown

in Fig. 4.1.30. We determine the force, which acts on every atom from neighboring ones, using two simplifying assumptions - the pairing interaction and elasticity (right of Hooke's law). Pairing interaction principle means that every atom interacts with only two the nearest neighboring atoms. In this case, the force acting on the *n*-th atom is determined by relation

$$F_n = F_{n+1} + F_{n-1} = -\mu(u_n - u_{n+1}) - \mu(u_n - u_{n-1}) = \mu(u_{n+1} + u_{n-1} - 2u_n), \qquad (4.1.13)$$

where μ is elastic coupling constant in Hooke's law, and u_n - displacement of the *n*-th atom from its equilibrium position during oscillations. Hence, the motion equation for the *n*-th atom has the form

$$m\frac{d^2u_n}{dt^2} = \mu(u_{n+1} + u_{n-1} - 2u_n), \qquad (4.1.14)$$

where *m* is atomic mass. The solution of equation (1.2.48) will seek as a flat elastic wave

$$u_n = A_n \exp\left(i\left(\vec{k}n\vec{a} - \omega t\right)\right), \qquad (4.1.15)$$

where $x = na$. Then, we obtain

$$F_n = m\frac{d^2u_n}{dt^2} = -m\omega^2 u_n. \qquad (4.1.16)$$

or

$$-m\omega^2 u_n = \mu(u_{n+1} + u_{n-1} - 2u_n). \qquad (4.1.17)$$

Substitution of (4.1.15) into (4.1.17) gives the expression

$$\omega^2 = \frac{\mu}{m}\left(2 - e^{+ika} - e^{-ika}\right) = 2\frac{\mu}{m}\left[1 - \cos(ka)\right] = 4\frac{\mu}{m}\sin^2\frac{ka}{2}, \qquad (4.1.18)$$

which can be converted to a so-called dispersion law for the one-atom chain

$$\omega = \pm 2\sqrt{\frac{\mu}{m}}\sin\frac{ka}{2}, \qquad (4.1.19)$$

which expresses the relationship between the oscillation frequency and the wave number $k = 2\pi/\lambda$ (or wavelength $\omega = 2\pi\upsilon/\lambda$, where υ – elasticc waves propagation velocity). Graphically, the dispersion law in the harmonic (elastic) approximation has the form shown in Fig. 4.1.31.

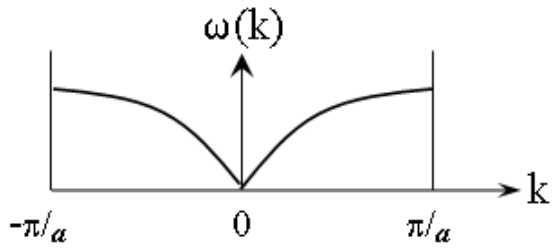


Fig. 4.1.31. The dispersion law for the one-dimensional atomic chain

Analysis of the dispersion relation (4.1.19) shows that it has the fundamental properties such as periodicity:

$$\omega\left(\vec{k}\right) = \omega\left(\vec{k}'\right), \tag{4.1.20}$$

where $\vec{k}' = \vec{k} \pm n\vec{G}^*$; and the lack of new oscillations outside the first Brillouin zone

$$-\pi/a \le k \le +\pi/a. \tag{4.1.21}$$

The relation (4.1.21) automatically means limiting of the frequencies spectrum

$$0 \le \omega \le \omega_{max} \tag{4.1.22a}$$

or wavelengths

$$2a \le \lambda \le L\ (\infty) \tag{4.1.22b}$$

of elastic waves in the crystal.

In the so-called *long-wavelength limit*, when $\lambda \to L\ (\infty)$ (or $k \to 0$, $\omega \to 0$), the dispersion law becomes linear:

$$\omega \approx \pm 2\sqrt{\frac{\mu}{m}}\frac{ka}{2} = \pm ka\sqrt{\frac{\mu}{m}} = v_0 k\,. \tag{4.1.23}$$

In this case, *the phase* and *group velocities* matches, and elastic waves dispersion, as a phenomenon, is missing. In other words, for small k or large λ values, the dispersion law $\omega(k)$ for elastic waves in a crystal, corresponding to the initial parts of the dispersion curve in Fig. 4.1.31, is a straight line. This means the independence of the phase velocity v_0 of elastic waves on the wavelength (the wave vector), as it should be carried out for a continuous media.

The sign ± after the equality sign in (4.1.23) means the presence of waves that propagate along the chain in two opposite directions.

The so-called *short-wavelength limit* ($\lambda \to 2a$, $k \to \pm\pi/a$, $\omega \to \omega_{max}$) in Eq. (4.1.19) gives the following relations for the phase and group velocities

$$\left.\begin{aligned} v_f &= \frac{\omega}{k} = \frac{2\sqrt{\frac{\mu}{m}}\sin\frac{ka}{2}}{k} \left| \frac{a}{a} = \frac{a\sqrt{\frac{\mu}{m}}\sin\frac{ka}{2}}{\frac{ka}{2}} = \frac{v_0\sin\frac{ka}{2}}{\frac{ka}{2}} \to \frac{2v_0}{\pi} \right. \\ v_{gr} &= \frac{d\omega}{dk} = 2\sqrt{\frac{\mu}{m}}\cdot\frac{a}{2}\cos\frac{ka}{2} = v_0\cos\frac{ka}{2} \to 0 \end{aligned}\right\}. \qquad (4.1.25)$$

This means that the waves with small λ (large k) values display dispersion phenomenon (depending the speed of elastic waves on λ or k): decrease of the phase velocity v_p with ω growth so that the dispersion law $\omega(k)$ becomes nonlinear - frequency growth is slowing down with k increase. As seen in Fig. 4.1.32c and from relation (4.1.23) for $k \to \pi/a$ (the edge of the Brillouin zone) the length of the shortest (standing) waves equals $\lambda \to \lambda_{min} = 2a$.

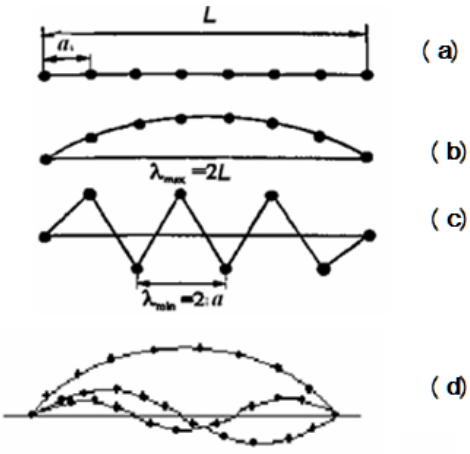


Fig. 4.1.32. Oscillations in monatomic chain (*a*) in one phase in the long-wave limit (*b*), in opposite phases in the short-wave limit (*c*) and at intermediate wavelengths (*d*).

Consider a stable standing wave pattern that occurs in the atomic chain of finite length $L = Na$, fixed at the ends of the chain (Figure 4.1.32a). In this case, oscillation with the smallest frequency ω_{min} or the maximum wavelength $\lambda_{max} = 2L$ (Fig. 4.1.32b)

$$\omega_{min} = \frac{2\pi v}{\lambda_{max}} \qquad (4.1.26)$$

occurs. At the same time, oscillations with the maximal possible frequency ω_{max} represent the standing wave with the minimal wavelength $\lambda_{min} = 2a$ (Fig. 4.1.32c). In this case, we obtain:

$$\omega_{max} = \frac{2\pi v}{\lambda_{min}} = \frac{\pi v}{a} = N\frac{\pi v}{l} = N\omega_{min}. \qquad (4.1.27)$$

For intermediate wavelengths the waves propagated are presented in Fig. 4.1.32d.

The above means that in the linear chain of N atoms N normal oscillations with frequencies ω_{min}, $2\omega_{min}$, ... , $N\omega_{min}$ or wavelengths $2a,4a,$ … , $2Na$ are possible. As the number of atoms N is large, so there are a lot of the normal vibrations, and the frequency difference between neighboring vibrations (adjacent frequencies) is very small. Therefore, the spectrum of frequencies of normal vibrations in crystals can be considered as quasi-continuous.

It follows from the above relations, the minimal elastic vibration frequency in a crystal is determined by its size L. Thus, for elastic waves in copper (v = $3{,}5\cdot10^5$ cm / s) with the sample size L = 100 cm minimal frequency $\omega_{min} = 1{,}1\cdot10^4$ s^{-1}. The maximal frequency is limited by the interatomic distance, so that for copper, with $a = 3{,}6\cdot10^{-8}$ cm, $\omega_{max} = 3\cdot10^{13}$ c^{-1}. Thus, the range of the normal vibrations in the crystal lattice includes both acoustic (in the range $\omega = 10^2 – 10^5$ s^{-1}) and ultrasonic (with $\omega \sim 10^{12}$ s^{-1}) waves.

The dispersion law of the elastic displacement waves in one-dimensional diatomic chain. Let us define the range of normal vibration frequencies through the linear chain with lattice parameter a which consists of two sorts of atoms with different masses, see Fig. 4.1.33.

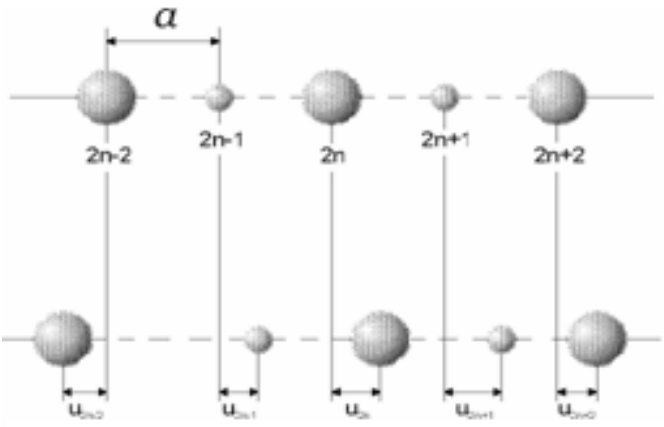


Fig. 4.1.33. The one-dimensional chain consisting of two sorts of atoms

Determining the force acting on every atom is also based on the principle of pairing interaction and the fairness of Hooke's law (elasticity). In this case, we have two equations of the type (1.2.48) and (1.2.49), separately for every sort of atoms. Therefore, the solutions (for two types of elastic waves) in such a crystal can be presented as

$$u_{2n} = Ae^{i[2nka-\omega t]}$$
$$u_{2n+1} = Be^{i[(2n+1)ka-\omega t]} \quad x_{2n} = 2na \quad x_{2n+1} = (2n+1)a \tag{4.1.28}$$

Substituting these solutions in the corresponding motion equations, we obtain

$$\left.\begin{aligned} -\omega^2 m u_{2n} &= m\frac{d^2u_{2n}}{dt^2} = \mu\left(u_{2n+1} + u_{2n-1} - 2u_{2n}\right) \\ -\omega^2 M u_{2n+1} &= M\frac{d^2u_{2n+1}}{dt^2} = \mu\left(u_{2n} + u_{2n+2} - 2u_{2n+1}\right) \end{aligned}\right\}, \tag{4.1.29}$$

which implies

$$\begin{cases} -\omega^2 mAe^{i[2nka-\omega t]} = \mu Be^{i[(2n+1)ka-\omega t]} + \mu Be^{i[(2n-1)ka-\omega t]} - 2\mu Ae^{i[2nka-\omega t]} \\ -\omega^2 MBe^{i[(2n+1)ka-\omega t]} = \mu Ae^{i[2nka-\omega t]} + \mu Ae^{i[(2n+2)ka-\omega t]} - 2\mu Be^{i[(2n+1)ka-\omega t]} \end{cases}. \tag{4.1.30}$$

Eliminating the unknown coefficients *A* and *B* from (4.1.30), we obtain the following relation for elastic displacement waves of two kinds:

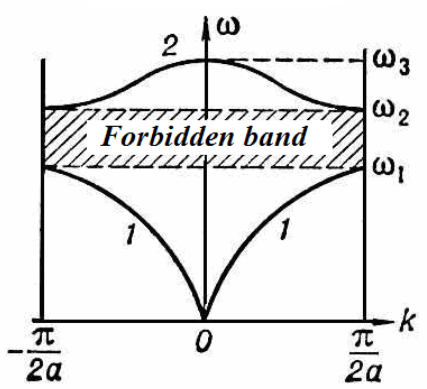
$$\left(2\mu - M\omega^2\right)\left(2\mu - m\omega^2\right) = 4\mu\cos^2 ka\,. \tag{4.1.31}$$

Solving the equation (4.1.31), we obtain the dispersion law for the diatomic linear chain in the form of a two-valued function

$$\omega = \pm\sqrt{\frac{\mu}{\overline{m}}\left(1 \pm \sqrt{1 - \frac{4\overline{m}\sin^2 ka}{M+m}}\right)}, \overline{m} = \frac{Mm}{M+m}\,. \tag{4.1.32}$$

Graphically, this law is shown in Fig. 4.1.33. The analysis of the dispersion law shows that it has the same fundamental properties as in singl-atomic linear chain: periodicity: $\omega\left(\vec{k}\right) = \omega\left(\vec{k}'\right)$, where $\vec{k}' = \vec{k} \pm n\vec{G}^*$, and the lack of new oscillations outside the first Brillouin zone

$$-\pi/2a \le k \le +\pi/2a. \tag{4.1.33}$$



Moreover, as is seen in Fig. 4.1.34, this dispersion law is characterised by the presence of two branches – acoustic 1 (low-frequency) and optical 2 (high frequency) 2.

Fig. 4.1.34. The dispersion law for the one-dimensional diatomic chain: *1* – acoustic mode, *2* – optical mode

From relation (1.2.66) it follows that for the acoustic branch 1 (bottom in Fig. 4.1.34), the dispersion is as follows:

$$\omega^2 = \frac{\mu}{m}\left[1-\sqrt{1-\frac{4\overline{m}\sin^2 ka}{M+m}}\right]. \qquad (4.1.34)$$

In *the long-wavelength limit* frequency range ($k \to 0$, $\omega \to 0$, $\lambda \to L(\infty)$) dispersion is linearized:

$$\omega^2 = \frac{\mu}{m}\left[1-\sqrt{1-\frac{4\overline{m}\sin^2 ka}{M+m}}\right] \approx \frac{\mu}{m}\left[1-\sqrt{1-\frac{4\overline{m}(ka)^2}{M+m}}\right] \to v_{02}k, \qquad (4.1.35)$$

where the group velocity is

$$v_{02} = \sqrt{\frac{2\mu a^2}{m+M}}. \qquad (4.1.36)$$

In the *short-wavelength limit* ($k \to \pm\pi/2a$, $\lambda \to 4a$, $\omega \to \omega_1$) limiting frequency ω_1 is achieved, and the phase and group velocities are expressed by the relations

$$v_f = \frac{\omega}{k} \to \frac{\sqrt{\frac{2\mu}{M}}}{\frac{\pi}{2a}} = \sqrt{\frac{8\mu a^2}{\pi^2 M}} << \frac{2v_0}{\pi}, \quad v_0 = \sqrt{\frac{\mu a^2}{m}} \qquad (4.1.37a)$$

$$v_{gr} = \frac{\partial\omega}{\partial k} \to 0. \qquad (4.1.37b)$$

For the *optical mode* 2 dispersion relation is as follows:

$$\omega^2 = \frac{\mu}{m}\left[1+\sqrt{1-\frac{4m\sin^2 ka}{M+m}}\right]. \qquad (4.1.38)$$

In the *long-wavelength limit* ($k \to 0$, $\omega \to \omega_3$), this branch of the dispersion law tends to saturate

$$\omega \to \omega_3 = \sqrt{\frac{2\mu}{m}}, \qquad (4.1.39)$$

so the phase and group velocities are of the form

$$v_{gr} = \frac{d\omega}{dk} \to 0 \qquad (4.1.40a)$$

$$v_f = \frac{\omega}{k} \to \frac{\omega_3}{0} \to \infty, \qquad (4.1.40b)$$

and the ratio of oscillation amplitudes for heavy and light atoms is

$$\frac{B}{A} = \frac{2\mu - m\omega_3^2}{2\mu\cos ka} \to \frac{2\mu - m\frac{2\mu}{m}}{2\mu} = 1-\frac{m}{m} = -\frac{m}{M}. \qquad (4.1.41)$$

In the *short-wavelength limit* ($k \to \pm \pi/2a$, $\omega \to \omega_2$) the limiting frequency is

$$\omega \to \omega_2 = \sqrt{\frac{2\mu}{m}}, \qquad (4.1.42)$$

so the phase and group velocities are of the form

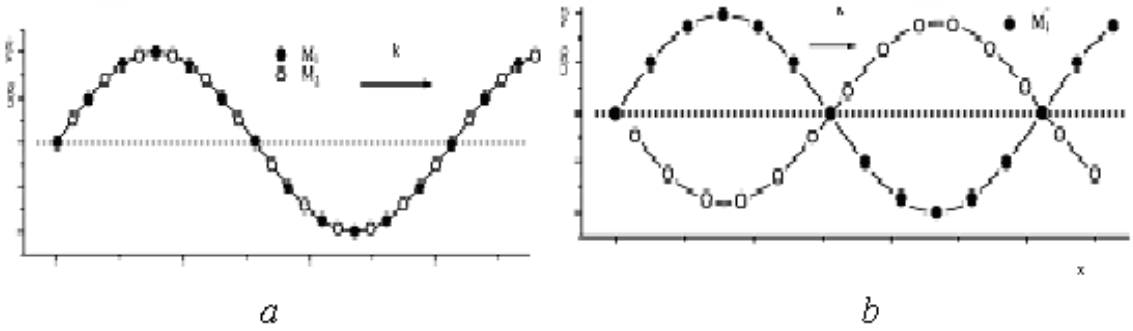
$$v_f = \frac{\omega}{k} \to \frac{\omega_2}{\pi/2a} = \sqrt{\frac{8a^2\mu}{\pi^2 m}} \qquad (4.1.43)$$

$$v_{gr} = \frac{d\omega}{dk} \to \frac{\omega_3}{k} \to 0, \qquad (4.1.44)$$

and the ratio of of oscillation amplitudes for heavy and light atoms is

$$\frac{B}{A} \to 0. \qquad (4.1.45)$$

Stable patterns of standing waves in a diatomic linear chain of finite length with their fixed ends for the acoustic and optical branches are shown in Fig. 4.1.35.



a *b*

Fig. 4.1.35. Oscillations of the atoms in one-dimensional diatomic chain. Oscillations of dissimilar atoms in-phase mode (*a*) for the acoustic branch and in antiphase (*b*) for the optical branch.

The dispersion law for elastic displacement waves in a three-dimensional lattice. If the three-dimensional crystal lattice consists of one type atoms, the total number of oscillating freedom degrees for every atom in a crystal is equal to three (because the oscillations can be carried along three mutually perpendicular axes) - two transversal and one longitudinal. Therefore, if the number of atoms in the crystal is N, it can be excited $3N$ normal oscillation modes with a particular frequency ω. This corresponds to the three branches of the dispersion law for acoustic waves (as, for examples, the 3 bottom curves for the two transversal and one longitudinal oscillation modes in Fig. 4.1.36). Note that in an anisotropic medium frequency depends not only on the magnitude of k, but also the wave propagation direction.

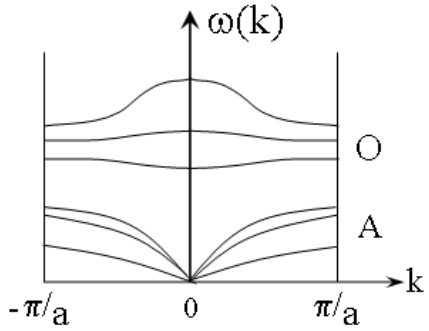


Fig. 4.1.36. Dispersion curves for the three-dimensional crystal

If the crystal consists of s sorts of atoms, the total number of vibrational modes is $3s$. Three of these modes form the acoustic branches of oscillations (three lower curves A of the dispersion law in Fig. 4.1.36) and the remaining ($3s$-3) will correspond to the so-called optical vibrations (the top three of the dispersion curves in Fig. 4.1.36 for the diatomic lattice with $s = 2$).

Concept of phonons. As shown in Section 1.2.1, the wave-particle duality of quantum particles (atoms) makes possible to write the relationship between the characteristics of the wave and particle properties as follows:

$$\varepsilon = \hbar\omega \qquad (4.1.46)$$
$$\vec{p} = \hbar\vec{k}\,.$$

In such a notation, left part of de Broglie relation presents corpuscular characteristics of wave-particle (its kinetic energy *E* and momentum *p*), and the right part – wave characteristics (its frequency ω and wave vector $\vec{k}$). With respect to the atomic oscillations in a crystal, such wave-particles (or, more correctly, *quasi-particles*) have the meaning of a sound quanta and are called *phonons*. The movement of these waves-particles in crystals is described by the Schrödinger equation and obeys the laws of quantum mechanics.

To describe the elastic displacement waves in a crystal as quasi-particles, we should present the crystal as a potential box filled with *phonon gas*. In this case, the collisions (interactions) of phonons should be conformed to the relevant laws of energy and momentum conservation. These laws can be easily obtained from the Laue equation (4.1.10) or (4.1.11), which holds not only for X-ray scattering, but for any scattering (including elastic waves) in the crystals. Laue equation can be written as the selection rules for the wave vectors giving the interference maxima in the scattering of rays on the direct lattice

$$\vec{k}' = \vec{k} + \vec{G}^*_{hkl}\,. \qquad (4.1.47)$$

Taking into account the de Broglie relation (2.1.80), equation (1.2.81) can be regarded as the form of momentum conservation law in the crystal

$$\vec{p}' = \vec{p} + \hbar\vec{G}^*_{hkl}\,, \qquad (4.1.48)$$

where $\vec{G}^*_{hkl}$ is the reciprocal lattice vector.

Fulfilment of relations (4.1.47) and (4.1.48) means that in the reciprocal space (space of wave vectors), there is a merely definite set of the wave vectors *k* (or momentums *p*) of the incident waves (X-ray, elastic, electronic) which can be scattered by the crystal in accordance with the Laue equation (the selection rules). This region of *k* and *p* values in the reciprocal space is called a *Brillouin zone*. In other words, the Brillouin zone is a region of the reciprocal lattice, which includes the whole set of wave vectors (momentums) of the falling

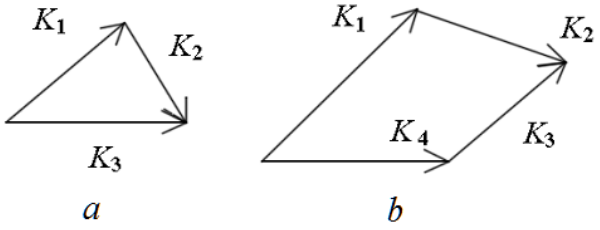


Fig. 4.1.37. Schemes of normal phonon collisions of types $\vec{k}_1 + \vec{k}_2 \to \vec{k}_3$ (*a*) and $\vec{k}_1 + \vec{k}_2 \to \vec{k}_3 + \vec{k}_4$ (*b*), when the conservation law for phonon momentum is performed.

on the crystal wave beam, for which there is a plane (*hkl*) of the direct lattice, giving a mirror reflection of waves with such wave vectors and leading to interference maxima.

The theory shows that there are the so-called *normal phonons collisions*, for which $\vec{G}^*_{hkl} = 0$. Normal collisions are of two types

$$\vec{k}_1 + \vec{k}_2 \to \vec{k}_3 \text{ and } \vec{k}_1 + \vec{k}_2 \to \vec{k}_3 + \vec{k}_4 \qquad (4.1.48)$$

(see, Fig. 4.1.37). Relations (4.1.48) show that the total momentum of the phonons at normal collisions $\vec{k}_1 + \vec{k}_2$ is conserved. As a result, the movement direction of the phonons, which is conditioned the heat transfer direction also remains the same.

In the crystals collision of phonons like $\vec{k}_1 + \vec{k}_2 \to \vec{G}^*_{hkl}$ are also possible, where $\vec{G}^*_{hkl} \neq 0$. In this case, the law of momentum conservation in the normal form is not satisfied. Such collisions are called *phonon collisions with a flip*. Peierls have shown that the conservation laws for momentum for such phonons, in contrast to the classic law

$$\vec{p}_1 + \vec{p}_2 = \vec{p}_3 + \vec{p}_4 \qquad (4.1.49)$$
$$\varepsilon_1 + \varepsilon_2 = \varepsilon_3 + \varepsilon_4$$

(in Fig. 4.1.37a), the phonon collisions with a flip will have a different form (Figure 4.1.37b):

$$\hbar\omega_1 + \hbar\omega_2 = \hbar\omega_3 \qquad (4.1.50)$$
$$\hbar\vec{k}_1 + \hbar\vec{k}_2 = \hbar\vec{k}_3 + \hbar\vec{G}^*_{hkl}.$$

The (4.1.50) coincides with the above selection rules (4.1.47) and (14.1.48). As can be seen from the expressions (4.1.50), at the interaction of phonons their number can be non-conserved.

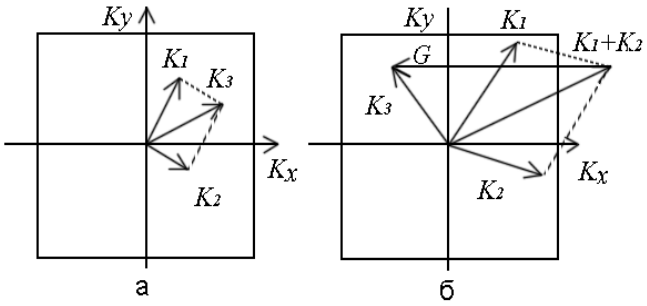


Fig. 4.1.38. The scheme of the normal collisions (*a*) and collisions with a flip, when the law of phonons' momentum conservation is not satisfied (*b*)

Furthermore, as seen from Fig. 4.1.38b, in such collisions the phonons' total momentum $\left(\vec{k}_1 + \vec{k}_2\right)$ is outside of the first Brillouin zone. This means that the momentum $\hbar\vec{G}^*_{hkl}$ is added to the momentum of the phonon system as a whole. Vector $\hbar\vec{G}^*_{hkl}$ is called by a *recoil momentum*.

As a result of a collision with a flip (Fig. 4.1.38b), the wave vector $\vec{k}_3$ of the "born" phonon determines the direction of phonon propagation, which differs from the direction of the total phonons' momentum. As will be shown below, such processes, when collisions strongly change the phonons motion direction (an therefore change the direction of thermal energy transfer), have a very strong influence on thermal conductivity of phonon gas (the crystal lattice). Note that, in the case of phonon collisions with a flip, phonons, generally speaking, does not have the mechanical momentum as the usual material particle. Therefore $\hbar\vec{k}$ is called *a quasi-momentum*.

All this means that every normal phonon mode (every elastic wave of vibrating atoms) with frequency ω can be associated with the quantum harmonic oscillator with a mass of oscillating atoms. The energy of such harmonic oscillator is $E_n = \left(n + \frac{1}{2}\right)\hbar\omega$, where n = 1,2,3 ... is oscillator energy level number. When quantum oscillator (normal mode) transits from one energy state to another, the energy can be changed merely by an amount $\hbar\omega$. Just as at the transition of an atom from one energy state in the lower state, the emission of a quantum electro-magnetic radiation – photons occurs, in the oscillating crystal we can speak about emission of quasi-particles – phonons with the energy $E_{\text{phon}} = \hbar\omega$ and quasi-momentum $\vec{p} = \hbar\vec{k}$.

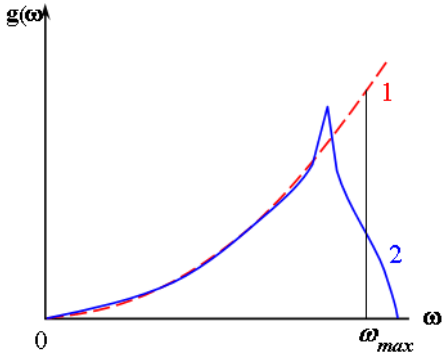
Note that phonons are bosons (particles with a quantum (zero) spin), so their energy distribution is described by the Bose-Einstein distribution. This function determines the average number of phonons n with frequency ω at temperature T. The higher temperature, the greater number of "born" phonons, and the higher phonon frequency, the less they number is generated at this temperature.

The number of possible oscillations (phonons) in a crystal. As indicated earlier, the number of possible wavelengths (and wave vectors), which can have the elastic displacement waves propagating through atomic chain of finite length, equals to the number of atoms in it:

$$\lambda = 2a; 4a; ...., 2Na = L; \qquad k = \frac{\pi}{Na}; \frac{2\pi}{Na}; \frac{3\pi}{Na}; \ldots \frac{N\pi}{Na} = \frac{\pi}{a} \Rightarrow k = (n\pi/Na). \qquad (4.1.51)$$

Distribution of these states by frequencies, in accordance with the Deby model, is called by Debye phonon spectrum

$$g(\omega) = g(k)\left(\frac{d\omega}{dk}\right)^{-1} = \frac{3}{8\pi^3} 4\pi k^2 v_0^{-1} = \frac{3}{8\pi^3} 4\pi \frac{\omega^2}{v_0^2} v_0^{-1} = \frac{3\omega^2}{2\pi^2 v_0^3}, \qquad (4.1.52)$$



with a quadratic dependence on frequency (curve 1 in Fig. 4.1.39).

Fig. 4.1.39. States density function for the Debye model (curve 1) and for a real crystal (curve 2).

Note also that for the Debye phonon spectrum (4.1.52), which lies in the frequency range

$$0 \le \omega \le \omega_{max}, \qquad (4.1.53)$$

the following normalising relation should be satisfied

$$\int_0^{\omega_{MAX}} g(\omega) d\omega = 3N. \qquad (1.2.95)$$

As can be seen from Fig. 4.1.39, the quadratic (by frequency) Debye phonon spectrum (curve 1) is significantly different from the real phonon spectrum of the crystal (curve 2) in the middle and high frequencies. The maximal frequency $\omega_{max} = \omega_D$ in the Debye phonon spectrum is called *the characteristic Debye frequency*.

4.1.3. Electric conductivity theory in metals

Crystalline solids greatly vary in their electrical properties. For example, metals are very good conduct electricity and is considered to be conductors. *Conductors* have a high electrical conductivity at a normal temperatures and a positive temperature coefficient of resistivity. At the same time, some crystals practically does not conduct current and are considered to be insulators (or dielectrics). *Insulators* have a very high electrical resistance: their main feature is the ability to polarization and the presentation of the internal electric field in them due to this reason. *Semiconductors* are intermediate conductors by their electrical conductivity between metals and insulators. Besides high dependence of conductivity on type and concentration of impurities, as well as the external energetic impacts (temperature, pressure, light, etc.).

Strong differences in the electrical conductivity between metal, semiconductor and dielectric materials are caused by the features of the distribution of the electrons by energy (energy spectrum) in crystals. This distribution is strongly influenced by the periodic arrangement of atoms in the crystal, forming, in particular, three-dimensional periodic potential in the field of which free electrons move. The nature of this motion of the electrons is also depends very strongly on their interaction with the crystal lattice, and between each other.

This section is devoted to the energy spectrum of electrons in crystalline materials and the motion of electrons in a periodic lattice potential, which allow us in the next sections to explain the reasons why all the crystals are divided on electric conductors (metals), poor conductors (insulators) and semiconductors and why they have different electrical properties.

Drude-Lorentz model for free electron gas. The first model describing the electronic properties of crystals (in particular, their electrical conductivity), has been established for metals and was called *the Drude model*. The Drude theory is based on the model of a metallic crystal as a system of N fixed positively charged ions, which form crystalline lattice and are embedded into gas of n_0 free electrons (Fig. 4.1.40). The last obey the laws of

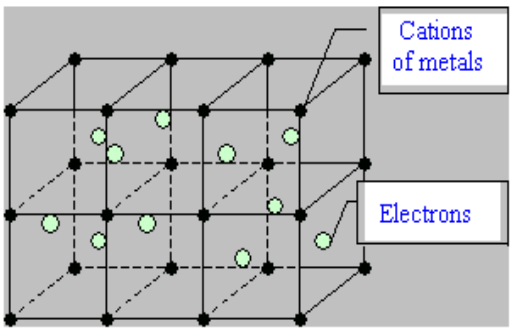


Fig. 4.1.40. Model of the free electron gas in a metallic crystal. classical Maxwell-Boltzmann statistics.

The chaotic and ordered motion of electrons in metal. The Drude model considers that the electrical resistance of the metallic crystals as a result of the scattering of electrons, moving under the influence of an applied electric field, by the fixed lattice ions (Fig. 4.1.41). Thus, the Drude theory is based on the following assumptions at the description of this electrons random motion in a crystal in the absence of an electric field:

1. In the interval between two successive collisions with ions, every electron moves by a straight line with the classical average velocity. This means that any type of Coulomb interactions with ions and other electrons is absent (this is called *the approximation of independent free electrons*).

2. Collisions between electrons and ions are treated as instant random events, suddenly changing the electron speed from some mean value to zero.

3. It is assumed that per unit of time, the electron experiences a collision with the probability $1/\tau$, where τ is the mean free time of an electron motion between two successive collisions.

4. It is also assumed that electrons come into equilibrium with the crystal only due to their collisions with the ions the lattice.

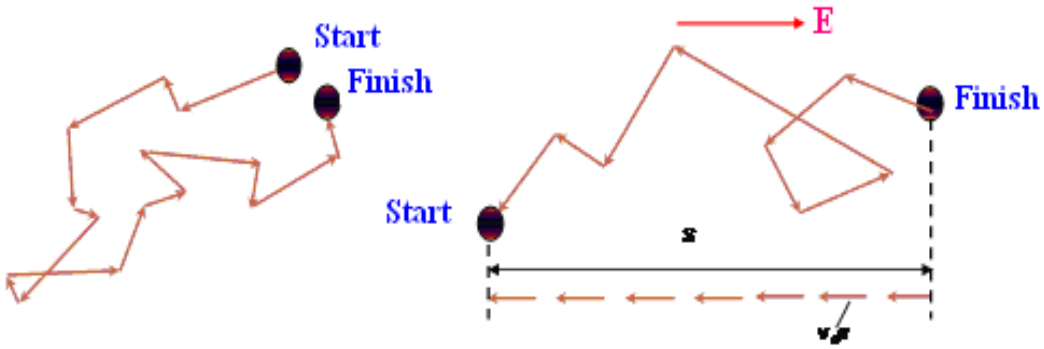


Fig. 4.1.41. Chaotic motion (*a*) and the ordered drift (*b*) of the of electrons in a crystal

These assumptions mean that in the absence of an external electric field, the electrons move randomly in space and do not have a preferred direction of movement. This is called *the chaotic (random) motion* (Fig. 4.1.41a).

When applying an external electric field to the crystal, the ordered movement of free electrons is superimposed on the chaotic motion. This second component is called *the drift of electrons* in an electric field (Fig. 4.1.41b). It is the electron drift provides electrical conductivity of crystals.

The electrical conductivity. Consider a group of electrons n_0, moving in an electric field E since the moment on $t = 0$ of its switching-off. As far as, according to our assumption 3, the probability of an electron collision with an ion is equal

to 1/τ, the number of electrons, which did not experience a collision for time *t*, is equal

$$n_t = n_0 e^{-\frac{t}{\tau}}. \tag{4.1.55}$$

Hence the rate of n_t decrease is equal to

$$\frac{dn}{dt} = -\frac{n_0}{\tau} e^{-\frac{t}{\tau}} = -\frac{n_t}{\tau}. \tag{4.1.56}$$

According to the classical Drude theory of conductivity, an electron moving in the electrictric field with strength *E*, is subjected to the effect of the force *F* = *eE*, which gives it the acceleration *a* = F/m = eE/m, where *m* – mass of a free electron. The electron, being accelerated, can not increase its velocity indefinitely due to collisions with ions and other electrons in the crystal lattice. The scattering of electrons by ions lead to the fact that the velocity to be gained in the direction of the electric field, practically falls to zero after each collision. As a result of such collisions, cycles "acceleration – scattering" for the electrons are repeated many times.

The average distance λ, which electrons pass from the collision to collision, is called *the mean free path*. Since the electron travels this distance for time *τ*, the latter is called *the mean free path time*. Thus the velocity vector, which electron gains in the electric field for the time t, is not more than $\vec{v}_t = -\frac{e\vec{E}t}{m} = \vec{a}t$.

This is an additional increase in the speed of electrons (due to the action of an electric field), which did not experience collisions with ions for the time t. It follows that the shift vector (drift) of the electron in the direction of the electric field is

$$\vec{x}_t = -\frac{e\vec{E}t^2}{2m} = \frac{\vec{a}t^2}{2} \tag{4.1.57}$$

(this drift is superimposed on the random thermal motion of electron, Fig. 1.3.2b).

Average total path (the total shift) in the direction of the electric field vector *E* for each of n_t electrons, which have not undergone collisions with ions for the time *t*, is equal to

$$< x_t >= \int_0^n x_t dn = \int_0^\infty \vec{x}_t \left(\frac{dn}{dt} \right) dt = -\frac{e\vec{E}n_0}{2m\tau} \int_0^\infty t^2 e^{-\frac{t}{\tau}} dt = \\ = -\frac{e\vec{E}n_0\tau^2}{m} \int_0^\infty \frac{1}{2} y^2 e^{-y} dy = -\frac{e\vec{E}n_0\tau^2}{m} = \vec{v}_D \tau n_0 \quad . \tag{4.1.58}$$

The average velocity v_D, which carriers gain under the influence of an electric field (it is called *the drift velocity*) for uniformly accelerated motion from the quiescent state is equal to half of the maximal velocity $v_D = 0{,}5eE\tau/m$. A more rigorous derivation, taking into account the distribution of free electrons in the crystal by the energies, leads to the expression:

$$\vec{v}_D = -\frac{e\vec{E}\tau}{m} . \tag{4.1.59}$$

According to the Drude model, for the loss of velocity, which electron gains in the electric field, only one collision is enough. So it gives

$$\tau = \frac{\lambda}{v}, \tag{4.1.60}$$

where $v = v_T + v_D$ is full velocity of the electron, which is a sum of the thermal v_T and drift v_D components, where usually $v_T >> v_D$. Note that in the classical Drude model the mean free path of electrons is independent on temperature and is given by

$$\lambda \approx \frac{1}{\pi R^2 N} , \tag{4.1.61}$$

where N is the number of centers of elastic electrons scattering (equal to the concentration of atoms in the crystal), and πR^2 is the cross section of electron scattering on a fixed atom with radius R.

The total density of the electron flow (current density) moving along the electric field can be expressed as the product of free electron charge on its drift velocity

$$\vec{j} = ne\vec{v}_D = \frac{ne^2\tau}{m}\vec{E} = \sigma\vec{E} . \tag{4.1.62}$$

This expression is called Ohm's law in differential form. Hence, the ratio front of the field vectors in (1.3.10), called the specific electric conductivity of the Drude electron gas will be equal

$$\sigma = \frac{ne^2\tau}{m}. \qquad (4.1.63)$$

The magnitude of the drift velocity in the electric field of unit strength

$$\mu = \frac{e\tau}{m} = \frac{\vec{v}_D}{\vec{E}} \qquad (4.1.64)$$

is called *the electron mobility*. From here one can get a second relation for conductivity of free electrons gas resulting from the Drude model:

$$\sigma = ne\mu. \qquad (4.1.65)$$

Substituting heat velocity to (4.1.63) or (4.1.65), we can obtain the temperature dependence of the conductivity for the Drude classical electron gas in the form

$$\sigma = \frac{ne^2}{m}\frac{\lambda}{\sqrt{\frac{3kT}{m}}} = \frac{ne^2\lambda}{\sqrt{3kTm}}. \qquad (4.1.66)$$

As will be shown later, this kind of dependence $\sigma(T)$ does not coincide with the experimental results.

The heat capacity and thermal conductivity of the classical electron gas in the Drude model. According to the classical Maxwell-Boltzmann statistics, which holds true for the gas of n free electrons in the Drude model, the energy (kT/2) should fall at every electron degree of freedom. This means that the heat capacity, as the derivative of the electron gas kinetic energy by temperature, must be equal

$$C_e = \frac{\delta U}{\delta T} = \frac{3}{2}k_Б n, \qquad (4.1.67)$$

where k is Boltzmann constant.

The most impressive success of the Drude model was the explanation of the empirical Wiedemann-Franz law. This law states that the ratio of thermal to the electrical conductivity κ_e/σ for the most metals is directly proportional to

temperature, in doing so the proportionality coefficient is the same, with sufficient accuracy, for all metals.

We can calculate the thermal conductivity of the electron gas, assuming that the major part of the heat flow in the metal is transferred by conduction electrons (therefore metals conduct heat much better, than insulators!) Taking into account that electron gas obey classical statistics, we can get thermal conductivity of an ideal electron gas in form $\kappa_e \approx \frac{1}{3} < v_T >^2 \tau C_e$, where C_e is specific heat of the classical electron gas in the metal. Substituting (4.1.60) and (4.1.67) to this relation, we obtain

$$\kappa_e = \frac{1}{2} nk\tau \frac{3kT}{m} = \frac{3nk^2}{2m} \tau T \,. \tag{4.1.68}$$

Hence it is easy to derive the relation

$$L = \frac{\kappa_e}{\sigma T} = \frac{\frac{3}{2}\frac{nk^2\tau T}{m}}{\frac{ne^2\tau}{m} T} = \frac{3}{2}\left(\frac{k}{e}\right)^2 = const \,, \tag{4.1.69}$$

which became known as the Wiedemann-Franz law. It is easy to calculate that the constant L in this equation, called the Lorentz number, is equal to $1{,}11 \cdot 10^{-8}$ $W/\Omega \cdot K^2$.

Boltzmann kinetic equation. The discrepancy between the calculated and experimental curves $\sigma(T)$ for metals required to transform the Drude model by taking into account the influence of the electric field on the distribution function f of the electrons by velocities (energies). According to Lorentz, the rate of change of f after the electric field switching on can be considered as the sum of two terms

$$\frac{\partial f}{\partial t} = \left(\frac{\partial f}{\partial t}\right)_{\text{ïîëí}} = \left(\frac{\partial f}{\partial t}\right)_{\text{ïîëå}} + \left(\frac{\partial f}{\partial t}\right)_{\text{ñòîëêí}} , \tag{4.1.70}$$

where contribution

$$\left(\frac{\partial f}{\partial t}\right)_{\text{ïîëå}} \approx -\frac{e\vec{E}}{m}\frac{\partial f}{\partial \vec{v}} \tag{4.1.71}$$

is called by *the drift term,* and contribution

$$\left(\frac{\partial f}{\partial t}\right)_{\text{ñòî ëêí}} \approx \frac{f - f_0}{\tau_r} \tag{4.1.72}$$

is *collisional term.* The parameter τ_r in this case is called *the relaxation* time *of electrons:* this is the time during which the electronic system comes into equilibrium with the lattice by collisions between electrons and ions. The value of f_0 is the unperturbed function of the Maxwell-Boltzmann in the absence of an electric field.

For the stationary case ($t > \tau_r$) relation (4.1.70) becomes the well-known Boltzmann kinetic equation

$$\frac{\partial f}{\partial t} = \frac{e\vec{E}}{m}\frac{\partial f}{\partial \vec{v}} + \frac{f_0 - f}{\tau_r} = 0, \tag{4.1.73}$$

whence

$$\frac{e\vec{E}}{m}\frac{\partial f_0}{\partial \vec{v}} = -\frac{f_0 - f}{\tau_r}. \tag{4.1.74}$$

From this equation we can get the expression for the Boltzmann distribution function for the stationary state, perturbed by the electric field,

$$f \cong f_0 + \frac{e\vec{E}}{m}\frac{\partial f_0}{\partial \vec{v}}\tau_r. \tag{4.1.75}$$

Here we consider that the shape of the distribution function in the electric field is not changed, i.e. it is assumed that

$$\frac{\partial f}{\partial \vec{v}} = \frac{\partial f_0}{\partial \vec{v}}. \tag{4.1.76}$$

In the improved Drude-Lorentz model electron relaxation time is assumed to be

$$\tau_r = A < v_T >^j. \tag{4.1.77}$$

Here A is the mean free path length λ, if j = -1, which corresponds to the elastic mechanism of electron scattering on neutral (uncharged) ions. For elastic scattering of electrons on the fixed ions of the crystal lattice the A coincides with the expression (4.1.61) and then the relaxation time depends on the velocity, but not on its direction:

$$\tau_r = \frac{\lambda}{< v_T >} = \tau. \tag{4.1.78}$$

Substitution of (4.1.75) in the definition of the average velocity of the electron (by the mean-value theorem) provides an additional contribution to the electron velocity in the direction of the electric field $\langle v_T \rangle = \langle v_T \rangle + \langle v_D \rangle$, which determines the electric current. Here

$$< \vec{v}_D >= \frac{e\tau_r \vec{E}}{m}. \tag{4.1.79}$$

In the direction of the electric field (along axis x) the current density is

$$J_x = -\iiint e v_x f dv_x dv_y dv_z , \tag{4.1.80}$$

where f - the Maxwell-Boltzman function.

Since $J_x = \sigma E_x$ and the integral (1.3.27) comprising f_0, is equal to zero, we obtain

$$J_x = \sigma E_x = -\iiint \left(\frac{E_x e^2}{m} \right) \tau_r v_x \left(\frac{\partial f}{\partial v} \right) dv_x dv_y dv_z . \tag{4.1.81}$$

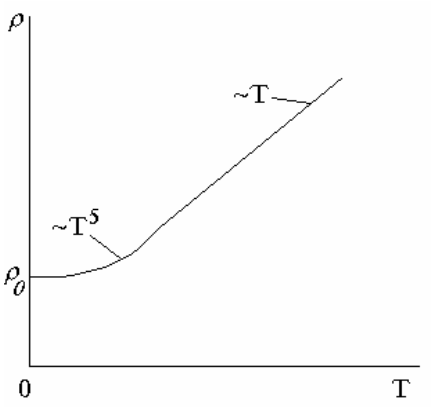
Using (4.1.78), we get relation,

$$\sigma = \frac{4ne^2\lambda}{3\left(2\pi mkT\right)^{1/2}} = \left(\frac{3\pi}{8} \right)^{1/2} \frac{ne^2\tau}{m}. \tag{4.1.82}$$

which is almost identical with the Drude formula (14.1.66).

The difficulties of the classical theory of electrical conductivity. The Drude-Lorentz model for free-electron gas, despite the apparent simplicity, allowed to explain Ohm's law (4.1.62) and to understand qualitatively some of the experimental data concerning the binding energy, electrical conductivity (for alkali metals), thermal conductivity, Wiedemann-Franz law, etc. However, when comparing with the experimental data, significant discrepancies between theory and experiment were revealed.

Drude theory allows to estimate the values of the relaxation times, mean free path and mobility of the electrons from measurements of conductivity and Hall effect. It is easy to estimate that $\tau \approx 10^{-14}$-10^{-15} s and the mean free path $\lambda \approx 0.1$-1 nm (for $v_T \sim 10^7$ cm/s). However, experiments have shown that for a very pure metals λ value can reach 1 cm when the temperature lowering, which is quite unclear in terms of the Drude theory. In addition, the real temperature



dependences of conductivity in metals were quite different (see Fig. 4.1.42), than the Drude (4.1.66) or the Drude-Lorentz (4.1.82) formulae give.

Fig. 4.1.42. The experimental temperature dependence of the conductivity in a metal

Experimental data on heat capacity C_e also differ substantially from the predictions of the Drude model: real $C_e \sim T$ whereas the Drude electronic capacity is constant, according to (4.1.67). Moreover C_e values for room temperature was by two orders of magnitude less than was predicted by Drude model for metals.

The measurements of the Hall effect for very pure substances at low temperatures showed that the experimental results for concentrations n of electrons in for alkaline (monovalent) metals are close to the Drude value of n_0, corresponding to one electron per atom. In the noble metals (as monovalent) value $n/n_0 = (1.3\text{-}1.5)$. At the same time, for the divalent Be and Mg R_H was found to be positive (!) and $n/n_0 = (0.2\text{-}0.4)$.

Thus, it has been shown that the Drude model can not explain some properties of metallic crystals, in particular, the difference between the measured values of the electron density in metals of different groups of the periodic table and a positive sign of R_H for some metals (in this model, R_H shoulf have only negative sign!). The Drude model also gives the wrong temperature dependence of the electrical conductivity and heat capacity of free electrons gas. Moreover, this model could not explain, in principle, the division of materials on metals and dielectrics.

Quantum theory of free electrons in metals (Sommerfeld model). The disadvantages of the classical Drude-Lorentz model were overcome by the quantum theory of free electrons in metals, developed by Sommerfeld. The Sommerfeld model uses the idea of quantization of the kinetic energy of free electrons, the Pauli exclusion principle and quantum Fermi-Dirac statistics.

The dispersion relation of the electrons in the crystal (the dependence of the electron energy on the wave vector). According to de Broglie's hypothesis (see above) and collectivized movement of free electrons in the internal electric field, produced by ions of the crystal lattice, in the weak-coupling approximation can be considered as movement of almost free particles. In this case, the expression for the kinetic energy of a free electron having a quasi-momentum p and mass m, can be written as

$$\varepsilon = \frac{p^2}{2m} = \frac{h^2 k^2}{2m}, \qquad (4.1.83)$$

where $k = 2\pi/\lambda$ is the wave number. In this case, according to the Sommerfeld model, the metallic crystal can be replaced by a kind of rectangular potential box (Fig. 4.1.43a), in which free electrons can be considered as noninteracting particles of an ideal gas having a parabolic dispersion law (Fig. 4.1.43b). In this case, the electrons have discrete energy spectrum and occupy energy levels (Fig. 4.1.43a) not more than two electrons to the level (due to the Pauli principle). The highest energy level that electrons in a metal can take at $T = 0$ K is called the Fermi level (see Fig. 4.1.43a).

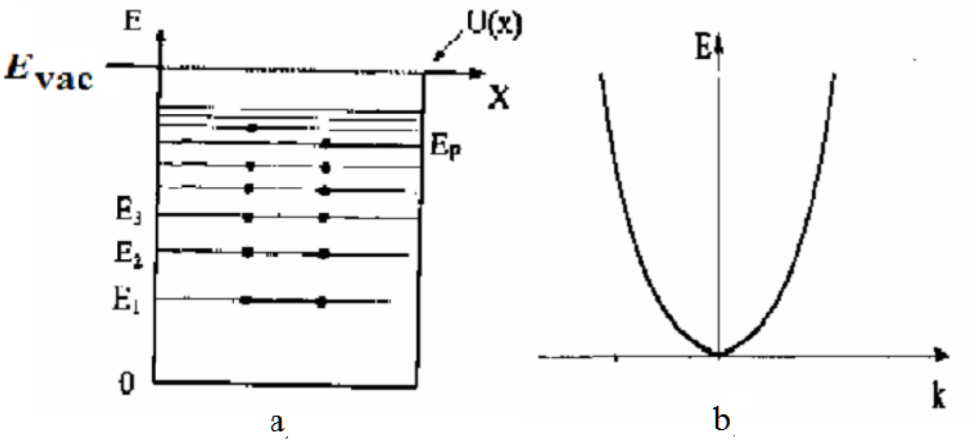


Fig. 4.1.43. The system of energy levels (*a*) and parabolic dispersion law (*b*) of the conduction electrons in the metal

In other words, the movement of the free electron can be interpret as propagation of plane wave with the velocity $v = \hbar k/m$ (see below). Plot of the $\varepsilon(k)$ function has the same form as the $\varepsilon(p)$ curve. At the same time, by virtue of parabolic dispersion law for free electrons in a metal (Fig. 4.1.43b), the distribution of electron states in *k*-space is characterized by a spherical symmetry (Fig. 4.1.44).

The Fermi level of electrons in metals. According to the Sommerfeld model, at $T = 0$ K, all the *n* electrons in a metal tend to occupy the states with the lowest values of the energy *E* subject to the Pauli exclusion principle (no more than two electrons with opposite spins per state). In such a case, due to squared-like electron dispersion law in a metallic crystal, in the absence of an external electric field, all occupied electron states in *k*-space will be inside a sphere of radius k_F (see Fig. 1.3.6). The surface of the ball is called *the Fermi surface,* and the highest energy of electrons, as noted above, is *the Fermi energy*.

When an external electric field E is applied along the k_x, distribution of electrons by the states, shown in Fig. 4.1.44, is shifted on some distance δk_x. Obviously, after a sufficiently long time, the velocity of the electrons and displacement of the electron distribution in Fig. 4.1.44 can become very large. However, according to the Drude-Lorentz model, for time $t > \tau_r$ collisions between electrons and lattice ions will result in the distribution to the steady state (the right part of Fig. 4.1.44).

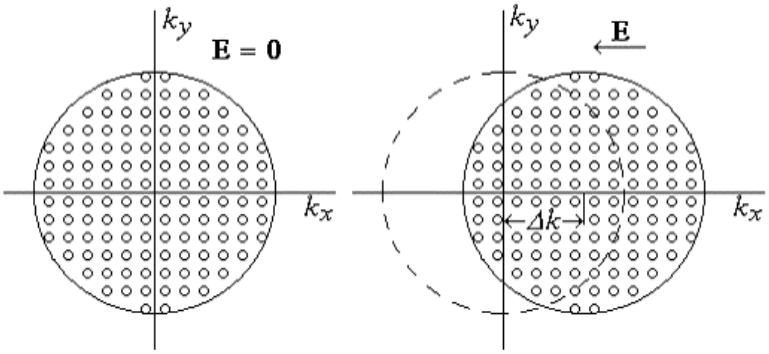


Fig. 4.1.44. Changing the distribution of electron states of a conductor in the Brillouin zone when exposed to an electric field *E*

It is easy to show that the Fermi energy is dependent on the concentration of free electrons *n* and at T = 0 is given by formula

$$\varepsilon_F = \frac{h^2\left(3\pi^2 n\right)^{2/3}}{2m}, \qquad (4.1.84)$$

which implies that $k_F \sim n^3$. As the temperature increases, the probability of electrons states (k values) occupying given by the Fermi-Dirac function is changed in accordance with Fig. 4.1.45. It should be noted that for all metals at all temperatures, including their melting temperature, the Fermi energy is 50-200 times greater than the value of thermal energy *kT*. Therefore, the electron gas in metals should be considered as highly degenerate electron Fermi gas.

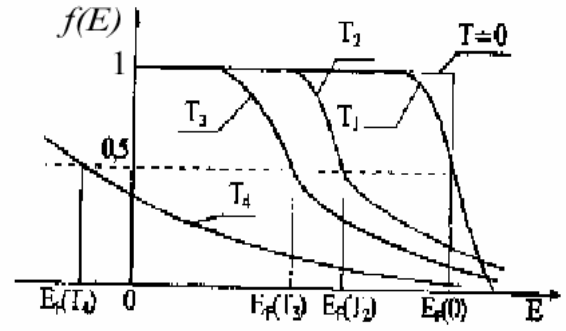


Fig. 4.1.45. The filling state function for the electron Fermi gas at different temperatures $T_4 > T_3 > T_2 > T_1 > 0$

The density of states and its dependence on energy. To determine the number of electrons with energies in a given interval (ε, $\varepsilon + d\varepsilon$), we need, except for the distribution function $f(\varepsilon)$, to know the density of states $g(\varepsilon)$. This function specifies the number of levels, which are attributable to unit energy (density of levels) in form

$$g(\varepsilon)=\frac{2\pi}{h^3}(2m)^{\frac{3}{2}}E^{\frac{1}{2}}. \qquad (4.1.85)$$

Extending the theory of the electron gas on the case of arbitrary temperatures and using the well-known normalization condition

$$n=\int_0^\infty f_{F-D}(\varepsilon)g(\varepsilon)d\varepsilon=(2s+1)\frac{2\pi}{h^3}(2m)^{3/2}\int_0^\infty \varepsilon^{1/2} f_{F-D}(\varepsilon)d\varepsilon, \qquad (4.1.86)$$

we can determine the temperature dependence of the Fermi level. The calculations show that the Fermi energy increases with temperature

$$\varepsilon_F(T)=\varepsilon_F(0)\left[1-\left(\frac{\pi^2}{12}\right)\left(\frac{kT}{\varepsilon_F(0)}\right)^2\right], \qquad (4.1.87)$$

where the Fermi energy $\varepsilon_F(0)$ at T = 0 K is given by (4.1.84). The weak dependence of the Fermi level on temperature indicates that the temperature increase leads only to a slight (~ kT) smearing of the Fermi-Dirac distribution function (see Fig. 4.1.45).

The main content of the Sommerfeld model. Thus the essence of the Sommerfeld model is that the free electrons in the metal can be considered noninteracting particles of an ideal gas in a squared-like potential well (Fig. 4.1.43a). Since electrons are fermions, they will consistently fill the energy levels up to the Fermi level, which will determine the maximal kinetic energy of the electrons in the metal at T = 0. Thus, the Fermi level $\varepsilon_F(0)$ in a metal defines the boundary between filled and unfilled (empty) states.`

As a result, at the absolute zero function has a step-like form (Fig. 4.1.45), while when increasing the temperature, $f_{\text{F-D}}$ is spraded by the width of 2kT (Figure 4.1.45). This means that the electrons, lying below the Fermi level, jump to higher, empty energy levels due to thermal excitation, freeing states below ε_F. Moreover, the main part of the electrons is completely insensitive to the very substantial changes in temperature. In doing so, the value of the Fermi energy is uniquely determined by the concentration of electrons. Dependence $\varepsilon_F(0)$ of n is non-linear, because at the growth of $g(\varepsilon)$ more and more electrons can be located at higher energy levels in the energy interval $d\varepsilon$.

Such an approach to the description of the electron gas in metals has allowed to eliminate a part of shortcomings for classical Drude and Drude-Lorentz models. In particular, they have described the reasons of lack of electron contribution to the heat capacity of the metallic crystal at temperatures of the order and above the Debye temperature.

Metallic conductivity in the Sommerfeld model. The aforementioned drawbacks of classical models in the description of the electrical conductivity of metals have been eliminated through the use of the Sommerfeld model based on the quantum Fermi-Dirac function for the electron energy distribution. According to the Sommerfeld model, with increasing temperature the Fermi distribution is "smeared" only slightly (Figure 4.1.45), so that the main part of electrons is not involved in the formation of the electrical conductivity (and heat capaity) of metallic crystal.

Using the Fermi-Dirac statistics instead of the Maxwell-Boltzmann in the kinetic Boltzmann equation, it is possible to obtain a quantum expression for the conductivity. As it turned out, in this case, we receive the same formula as was in the Drude-Lorentz model

$$\sigma = \frac{ne^2\tau_r}{m}. \qquad (4.1.88)$$

However, the characteristic time τ_r is called *the relaxation time*, and has a completely different meaning. It is given by relation

$$\tau_r = \frac{\lambda(\varepsilon_F)}{v(\varepsilon_F)}, \qquad (4.1.89)$$

Since only applies to electrons (with the concentration on the order $kTn/E_F(0)$), which have an energy around the Fermi energy at T = 0 K. Then (3.1.44) implies that

$$\sigma = \frac{ne^2\lambda(\varepsilon_F)}{mv(\varepsilon_F)} = \frac{ne^2\lambda(\varepsilon_F)}{\sqrt{2m\varepsilon_F}}, \qquad (4.1.90)$$

where the Fermi velocity $v(\varepsilon_F) = \sqrt{\frac{2\varepsilon_F}{m}} \approx 10^8\ cm/s$ is much higher than the thermal velocity of the classical electron gas. This approach allowed Gryunayzen theoretically obtain the temperature dependence of the conductivity $\sigma \sim T^{-n}$, where n = 3÷5, which is close to the experimental ones.

Advantages and disadvantages of the Sommerfeld model. Sommerfeld model has allowed to remove some inconsistencies and explain the shortcomings of the classical models of the electron gas. In particular, it has explained why not all free electrons may be involved in the energy transfer and contribute to the electric current. It also allowed to calculate correctly the laws of thermionic emission and electrical conductivity for metallic crystals. However, this model has not been able still to explain the reasons for the existence of insulators, semiconductors and even more semiconductors. In particular, it remained unclear fundamentally different temperature dependence of the electrical conductivity in metals (power-like) and dielectrics (exponential), a positive sign of the Hall effect in some crystalline materials, as well as the high sensitivity of semiconductors to external influences (temperature, radiation, light, magnetic and electric fields, etc.).

4.1.4. Zone theory of crystalline solids

Tight-binding approximation. Consider what happens to the energy of the electrons in the atoms, which at first are at large distances r, and then draw closer to a distance r_0, which corresponds to the equilibrium interatomic distances in the crystal (~10^{-8} cm). In the remote (non-interacting) atoms the total energy of the electrons takes only a number of distinct values of E_1 , E_2 , E_3, ... (Fig. 4.1.46a). When $r = r_0$, potential energy curves $U(r)$ in the spaces between the atoms are overlapped, so that potential barriers, separating the electrons in adjacent atoms, drop (Fig. 4.1.46b). As a result, the electrons, being on the highest level (E_3 in this figure), can migrate freely from atom to atom, and therefore belong the entire crystal that corresponds to Sommerfeld.

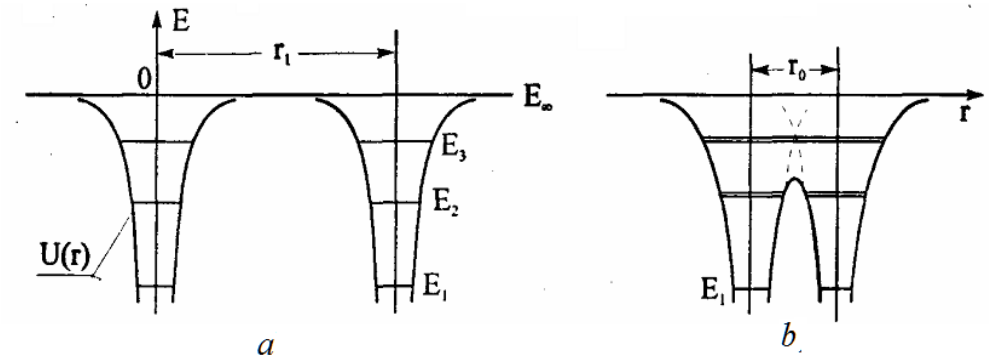


Fig. 4.1.46. The energy of the electrons in the isolated (non-interacting) atoms (*a*) and atoms in the crystal lattice (*b*)

If the crystal contains N atoms, then all electrons being on the outer levels in isolated atoms could be in the same state in the crystal. However, the Pauli principle forbids the presence of more than two electrons at one energy level. Therefore, at the formation of the crystal the energies of these electrons should be changed due to the electron interaction. In the simplest case, the interaction between the N electrons on these external levels (level E_3 in this case in Fig. 4.1.46*a*) should result in their splitting onto N close sub-levels, every of which may be occupied by only two electrons with opposite spins. The same applies to higher (excited) levels E_4, E_5, …, E_∞. Fig. 4.1.47 schematically shows how discrete levels of N isolated atoms (being at $r >> 10^{-8}$ cm) expand in the zone (band) with r decreasing. In the simplest case N sub-levels appears in every allowed band, having energy minima at $r = r_0$.

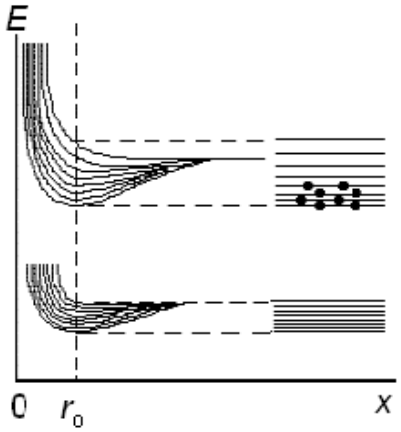


Fig. 4.1.47. Scheme of energy bands creation in the crystal from atomic energy levels when atoms approach equilibrium interatomic distance r_o

For the valence electrons the width of the allowed energy bands is of a few electron volts $\Delta E \sim h/\tau \sim 1$ eV (close to the energy of interatomic bonds). It follows that for typical value $N \approx 10^{22}$ cm^{-3} for concentration of atoms in crystals, the interval between levels constitutes ($\Delta E/N \sim 10^{-22}$ eV). In other words, the levels in the allowed band are so close that even at low temperatures thermal energy of the electrons kT is much greater than this value, so that this band can be considered as a zone of *quasi-continuous* values of the allowed energies.

As a result, to move through the crystal can not only electrons with energy E_3, but the electrons, which are located at the levels of E_2, which are separated in the crystal low potential barrier (see Fig. 4.1.46b). As a result, due to the tunneling effect or the lattice thermal energy kT, these electrons are able to overcome these barriers with the width of about 10^{-8} cm, and also become common throughout the crystal. Therefore, in this case, a range of allowed energies, although with a smaller width, is also appears.

Electrons, located at lower energy levels (like E_1 in Fig. 4.1.46b), are separated from those of the electrons in the neighbor atoms much more high and wide potential barriers, so they remain localized on their atoms. This means that the deep level E_1 is not splitted and does not create the allowed band for the electron states.

Therefore, if the levels E_2 of the electrons in the atom were filled with electrons, in the crystal these electrons form allowed band of free electrons. If a high-energy levels E_3 in isolated atoms were free of electrons under normal conditions, in the crystal such levels form the empty band (free of electrons).

The mentioned approach, when the energy spectrum of electrons in a crystal is based on the presence of energy levels in isolated atoms, is called *the strong-coupling approximation*. It illustrates well (at least qualitatively) the general regularities of the energy bands formation at coming together of isolated atoms when crystal lattice is formed, Fig. 4.1.47.

Thus, as follows from the strong-coupling approximation, the kinetic energy of the electrons in the crystal presents a set of wide empty and filled with electrons bands, which are separated by forbidden zones (the electrons with such kinetic energy are absent!). Such a form of energy spectrum of electrons is called *band (zone) energy spectrum*. The band spectrum (see Fig. 4.1.47) is common to all solids and determines the most of the properties of crystalline solids (electrical, magnetic, optical, etc.).

An electron in a periodic field of the crystal. Single-electron adiabatic approximation. As the crystal is composed of about 10^{23} cm^{-3} atoms and valence electrons, the most complete data on the electronic subsystem of the crystal can be obtained by solving the full Schrödinger equation with taking into accounts of all kinds of interactions: every electron interacts with every atom and every electron. Because it is difficult, we need to simplify the task. One of the simplifications in the calculation of the electron energy spectrum suggests that the electronic and lattice subsystems move independently of each other, that is without exchange by energy (*adiabatic approximation*). Another simplification assumes that behavior of all electrons is the same in different unit cells, and the action of the ions and the other electrons on a given electron can be replaced by an impact of some periodic potential field (*a single-electron approximation*).

A single-electron Schrödinger equation for electron moving in the field of chain of ions for the stationary case ($t > \tau_r$), will be the following:

$$-\frac{h^2}{2m}\frac{d^2\Psi(r)}{dr^2}+V(r)\Psi(r)=E\Psi(r), \qquad (4.1.91)$$

where the potential of the crystal lattice, in which free electron is moved, is periodic $V(x) = V(x + a)$. The solution of (4.1.91) for free electron will be in the form of a plane wave de Broglie

$$\Psi_k(x) = U_k(x)\cdot \exp(ik\cdot x), \qquad (4.1.92)$$

where

$$\Psi_k(x) = \Psi_k(x + Na),\ U_k(x) = U_k(x + Na). \qquad (4.1.93)$$

Note that factor $U_k(r)$ accounts for the effect of periodic crystal field on the electron

and reflects the fact that the probability to find an electron in a particular region of the crystal is repeated from one cell to another.

Solution of the stationary Schrödinger equation for the Kronig-Penney model. In addition to a strong coupling approximation another approach was proposed by Kronig and Penny. It shows that the presence of the band structure of the energy spectrum for electrons in a crystal is a fundamental consequence of the translational symmetry of the crystal lattice. For simplicity, in this approach, the case of one-dimensional lattice with periodic potential

$$V(x) = V(x+a) = V(x+2a) = ... \qquad (4.1.94)$$

which is a combination of an infinite number of potential wells of width b and potential barriers of height V_0 and width of $a = a - b$ (Fig. 4.1.48).

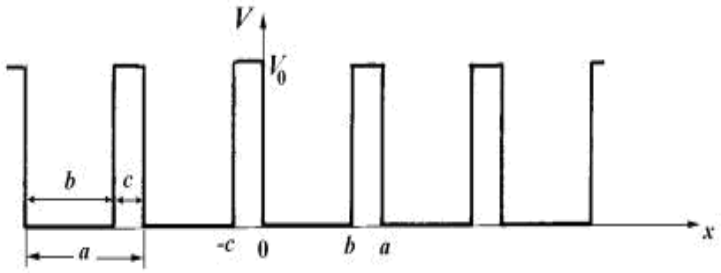


Fig. 4.1.48. The change of potential energy of an electron in a one-dimensional single-atom chain with recurrent rectangular potential wells in the model of the Kronig-Penney

Note here that the lattice period is $a = b + c$. In this case, the potential energy of the electron in the equation (4.1.91) is given by the relation

$$\begin{cases} V(x) = 0 & when \quad 0 < x < b \\ V(x) = V_0 & when \quad b < x < a = b + c \end{cases}, \qquad (4.1.95)$$

and the electron in the well is at the energy level $E < V_0$. The height of the potential barrier for electron is equal $\Delta V = V_0 - E$, whereas the thickness of the barrier is c.

If the velocity of an electron in the chain is equal to v, it realizes v/b approaches to the barrier per unit time at the motion in the well along the chain.

Then, the frequency of transitions of n electrons from one well to another (from one atom to another) is proportional to the probability of tunneling through a potential barrier. As follows from quantum mechanics, the probability of tunneling through a rectangular potential barrier (transition frequency) is exponentially dependent on the width of the potential barrier and its height

$$\nu = \frac{v}{b}\exp\left(-\frac{2h}{c}\sqrt{2m\Delta V}\right). \qquad (4.1.96)$$

Evaluation by formula (4.1.96) indicates that, when the potential barrier width $c \sim 10^{-8}$ cm, its height of about $V_0 \sim 10$ eV (close to the ionization potential of the isolated atom), velocities of electron in the atom $v \sim 10^8$ cm/s and radius Bohr orbit $b \sim 10^{-8}$ cm, the time during which an electron is in a particular lattice site equals merely only $\tau = 1/\nu \sim 10^{-15}$ seconds. In other words, the electrons of the outer atomic shells in the crystal are not localized near a specific site in the lattice, but move across the crystal at a velocity $v \sim 10^{-8}/10^{-15} \sim 10^7$ cm/s.

For the electrons of the inner atomic shells the potential barrier is wider and higher, and the probability of tunneling is much smaller than for the valence electrons. Consequently, electrons from substantially deeper atomic levels are localized in lattice sites.

For a selected Kronig-Penney potential, the general solution of Schrödinger equation will be in the form of a plane wave with an modulated amplitude like (4.1.92). Therefore the general solution of the Schrödinger equation for the electrons in one-dimensional Kronig-Penney model will look like:

$$\frac{\beta^2-\alpha^2}{2\alpha\beta}sh(\beta c)\sin(\alpha a)+ch(\beta c)\cos(\alpha a)=\cos\left(k(b+c)\right), \qquad (4.1.97)$$

where $\alpha = \sqrt{\frac{2m\varepsilon}{\hbar^2}}$ и $\beta = \sqrt{\frac{2m(V_0-\varepsilon)}{\hbar^2}}$. The expression (4.1.97) can be greatly simplified by assuming that the width of the barrier tends to zero $c \to 0$, and its height – to infinity $V_0 \to \infty$, but in a way that their product remains constant ($V_0 c \approx$ const). Under these conditions the expression (4.1.97) is transformed to

$$p\frac{\sin(\alpha a)}{\alpha a}+\cos(\alpha a)=\cos(ka), \qquad (4.1.98)$$

where the parameter p characterizes the "power" of the potential barriers that separate regions with zero potential

$$p = \frac{mabV_0}{\hbar^2}. \qquad (4.1.99)$$

Since parameter α defines kinetic energy E of the electron and k is its wave vector, the expression (3.1.58) is actually presents dispersion law $E(k)$ for an electron in a crystal lattice.

When $p \to 0$ (*weak-coupling approximation*), we can get from (3.1.58) that $\alpha a = ka$, whence

$$\varepsilon = \frac{\hbar^2 k^2}{2m}, \qquad (4.1.100)$$

which corresponds to the dispersion law for free electrons in the Sommerfeld model (see above).

On the contrary, if $p \to \infty$ (*approximation of absolutely bound electrons*), the energy of the electrons is independent of k. As seen from the equation (4.1.99), allowed energy values will have only those electrons for which $\sin(bc) = 0$, i.e. $\alpha a = n\pi$. Hence the energy of the electrons is given by relation

$$\varepsilon = \frac{\hbar^2 \pi^2}{2ma^2} n^2, \qquad (4.1.101)$$

where n is the integer. The expression (4.1.101) corresponds to the solution of the Schrödinger equation for a particle, which is in a one-dimensional potential well of width c and the walls with infinite height. The energy levels of a particle in such a well are discrete that corresponds to the case of isolated atoms.

The analysis shows that if the value of p is finite ($p >> 1$), the electron will overcome potential barriers easier and easier if its energy (parameter α) significantly increases compared with V_0 and there will come a time when the electron becomes like a free. In the intermediate case (*strong coupling approximation*), the energy of the electrons is characterized by the dispersion relation:

$$\varepsilon_n = \frac{\hbar^2 n^2 \pi^2}{2ma^2}\left[1 + \frac{2}{p}\left((-1)^n \cos(ka) - 1\right)\right] = \frac{\hbar^2 n^2 \pi^2}{2ma^2}\left(1 - \frac{2}{p}\right) + \frac{2\hbar^2 n^2 \pi^2}{2pma^2}(-1)^n \cos(ka) \qquad (4.1.102)$$

As follows from this relation, the dispersion law for the electrons is multi-valued, and consists of n allowed energy bands (zones). By choosing different values of n = 1,2,3..., we obtain the a whole set of dispersion laws for different energy bands:

$$\begin{cases} \varepsilon_0(k) = 0 \\ \varepsilon_1(k) = \dfrac{\hbar^2\pi^2}{2ma^2}\left(1-\dfrac{2}{p}\right) - \dfrac{\hbar^2\pi^2}{pma^2}\cos(ka) \\ \varepsilon_2(k) = \dfrac{4\hbar^2\pi^2}{2ma^2}\left(1-\dfrac{2}{p}\right) + \dfrac{4\hbar^2\pi^2}{pma^2}\cos(ka) \\ \varepsilon_3(k) = \dfrac{9\hbar^2\pi^2}{2ma^2}\left(1-\dfrac{2}{p}\right) - \dfrac{9\hbar^2\pi^2}{pma^2}\cos(ka) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases} \qquad (4.1.103)$$

Thus, using the approach of the Kronig-Penney, we have shown that the energy spectrum of electrons in a periodic potential field which is due to the translational symmetry of the lattice, really has a band structure.

Dependence of the electrons energies the crystal on the wave vector (quasi-momentum) for different types of allowed bands (4.1.103) is schematically shown in Fig. 1.3.124.49. Dot-dashed parabola in this figure corresponds to the limiting case of the energy of free electron with a quadratic dispersion law (4.1.100). For the electron in crystal this parabola is replaced by σ-shaped portions of sinusoid which are separated from each other by discontinuities in the energy spectrum at the boundaries of Brillouin zone at $k = \pm\pi n/a$, where n = 1, 2, 3, ... These discontinuities just are those gaps between allowed bands, which were mentioned above.

As is seen from Fig. 4.1.49, the parts of sine curves in different allowed energy bands I, II, III may belong to different Brillouin zones. The ranges of allowed values of the energy (allowed bands) I, II, III are separated by intervals of the forbidden values (energy gaps).

Since the electron states in all Brillouin zones are physically equivalent, we can use the procedure of shift (umklupp process) of the dispersion relation branches on the vector $G = 2\pi/a$ of reciprocal lattice. Using this procedure we can reduce the dispersion laws of different Brillouin zones only to the first Brillouin zone. Horizontal arrows in Fig. 4.1.49a indicate the directions of these shifts. The generalized dispersion law in Fig. 4.1.49*b* has a form of the “disrupted” parabola,

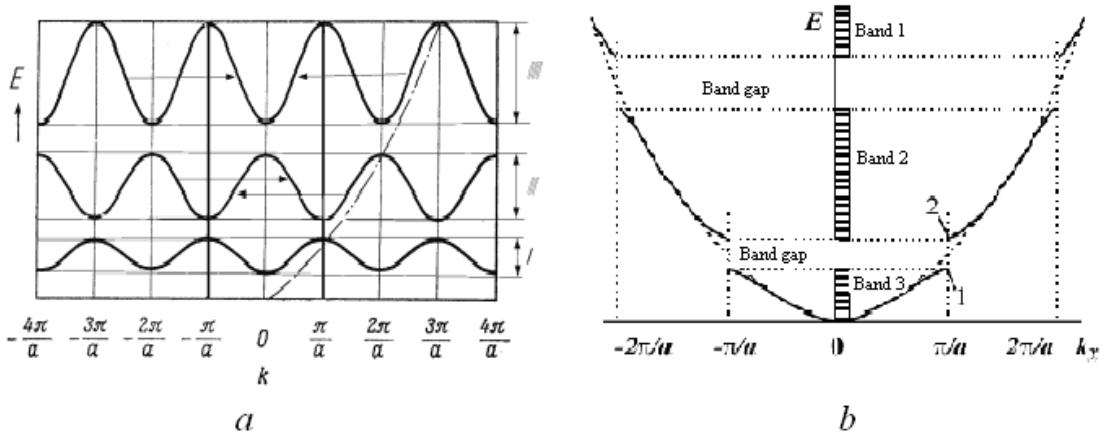
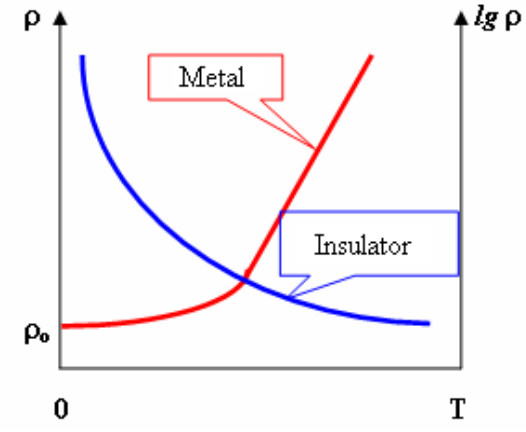


Fig. 4.1.49. The Kronig-Penney $E(k)$ dependences for different allowed energy bands for an electron in a one-dimensional lattice in the representation of extended zones (*a*) and the generalized dispersion law (*b*)

showing the dependence of $E(k)$ for a free electron in the first Brillouin zone $-\frac{\pi}{a} < k \le \frac{\pi}{a}$. Such an image is called *the representation of the reduced Brillouin zone*.

So, we have proved, using Kronig-Penney model, that the electron energy spectrum is divided into allowed zones (bands) and forbidden zones (gaps). The presence of gaps E_g in the energy spectrum means the lack of electrons with such kinetic energies.

Filling the energy bands by electrons. Dividing crystals on metals, dielectrics and semiconductors. As noted earlier in this chapter, all crystalline solids can be divided into metals, dielectrics and semiconductors primarily by the values of conductivity. For typical metals, at room temerature this value is of $10^8...10^6$ $(\Omega \cdot m)^{-1}$. For very good insulators electrical conductivity does not exceed 10^{-11} $(\Omega \cdot m)^{-1}$. Crystals with intermediate values of electrical conductivity are usually related to semiconductors.



However, as noted above, insulators, semiconductors and metals are different from each other not only by the resistivity values. In metals and dielectrics a fundamentally different types of the temperature dependence of the electrical conductivity are observed (see Fig. 4.1.50).

Fig. 4.1.50. The temperature dependences of the electrical resistance in metals and dielectrics.

Furthermore, semiconductors are also characterized by a very high sensitivity to various external impacts (temperature, light, magnetic and electric fields, etc.).

The division of crystalline solids on metals, semiconductors and insulators, as well as marked differences between their conductivities can be explained qualitatively on the basis of the foregoing model of electron energy spectrum. As follows from the model, such large differences in the electrical properties of solids are related to the structure and the degree of filling of energy bands by the electrons.

The number of electrons in a crystal, of course, depends both on number of atoms N and number of electrons per atom. The electrons in the crystal tend to occupy the lowest energy levels. However, the Pauli principle forbids to be more than two electrons at every energy level. Therefore, in the crystal the lowermost energy levels in the allowed bands are primarily filled. As a result, some (the lowest) bands becomes completely filled, whereas the uppermost are filled either partially or remain completely free of electrons (Fig. 4.1.51).

Band with the highest energies, which is completely filled with electrons, is called *the valence band*. Next, higher band, which can be either partially filled (Fig. 4.1.51*a*) or fully unfilled with electrons (see Fig. 4.1.51*b, c*), is called *the conduction band.*

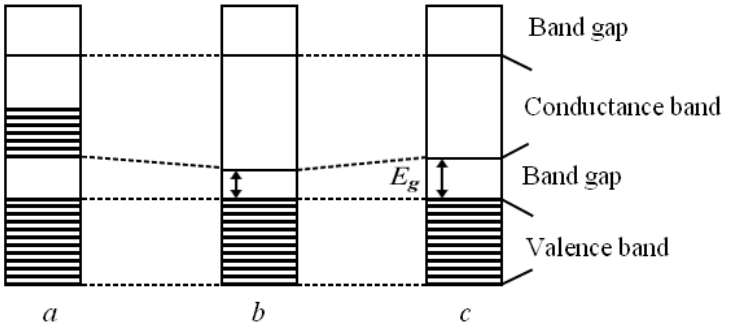


Fig. 4.1.51. Energy bands in the crystal and a scheme of their filling in the case of metal (*a*), dielectrics (*b*) and the semi-conductors (*c*)

So, it follows from the above, the structure of the energy bands (primarily, the value of band gap E_g) of the crystal and the nature of the filling of allowed zones by electrons render a decisive impact on the value of its electrical conductivity. If the magnitude of the electric field is not higher than 10^4 V/m, then at a distance equal to the mean free path (usually ~ 10^{-8} m), the electron in the crystal acquire energy of about 10^{-4} eV. It is clear that these values allow the electron to pass from one level to another but only within the same energy band. To move between the bands it needs the energy of the order of band gap E_g.

These considerations lead to the conclusion that the metallic crystal should have partially filled conduction band in the energy spectrum (Fig. 4.1.51a) to be highly conductive. If the valence band in the crystal is completely filled with electrons and the conduction band is empty (Figure 4.1.51b, c), weak external

electric field will not be able to throw electrons over the band gap from the valence to the conduction zone. Therefore, application of such field will not lead to an electric current in such a crystal and it behaves as dielectric or semiconductor. In this case, dielectrics crystals are characterized by a relatively wide band gap (Fig. 4.1.51c) with typical $E_g > 3$ eV. Thus, for diamond $E_g = 5.2$ eV, for boron nitride 4.6 eV and for alumina ~ 7 eV. In typical semiconductors bandgap is significantly less (Fig. 4.1.51b) and not more than 3 eV. For example, in germanium $E_g = 0.66$ eV, for silicon 1.12 eV, while the indium antimonide has $E_g = 0.17$ eV.

4.1.5. Electron dynamics in periodic lattice

Features of the electrons motion by the crystal are not only due to the external electric field, but also owing to their interaction with the crystal lattice. In general, the motion of the electron can be described by Newton's second law

$$\vec{F} + \vec{F}_{\text{int}} = m\frac{d\vec{v}}{dt}, \tag{4.1.104}$$

where electron is subjected to the impact of two forces – from the external electric field and from the inner periodic lattice field. However, as it turns out, the free motion of a single electron in a crystal can be described by the same equation of Newton, but which takes into account only the influence of the external force

$$\vec{F} = m^*\frac{d\vec{v}}{dt}. \tag{4.1.105}$$

This description is called *the effective mass approximation*. In this approximation, the electron in a crystal is considered as a quasi-particle described by a wave function whose energy and velocity depends on the wave vector through the dispersion law, but the mass is not equal to the mass of a free electron in a vacuum.

As noted above, the absolutely free electron is described by a monochromatic de Broglie wave and is not localized anywhere. Real electron in the crystal is necessary to compare the group of de Broglie waves with different frequencies ω and wave vectors $\vec{k}$. Center of this group is movingin space with the group velocity, which corresponds to the velocity of the electron energy transfer. For one dimensional case, it is expressed as

$$v_{\text{ãð}} = \frac{dx}{dt} = \frac{d\omega}{dk} = \frac{1}{\hbar}\frac{dE}{dk}. \qquad (4.1.106)$$

The increase in the electron energy *dE* under the impact of an external force *F* equals an elementary work *dA*, which accomplishes this force (such as an electric field) over the electron for infinitely small time interval *dt*

$$dE = dA = Fdx = Fv_{\text{ãð}}dt \qquad (4.1.107)$$

Substituting the expression (4.1.105) for the group velocity in (4.1.106), we obtain

$$dE = \frac{F}{\hbar}\frac{dE}{dk}dt\,. \qquad (4.1.108)$$

Hence

$$\frac{d}{dt}(\hbar k) = F\,. \qquad (4.1.109)$$

Extending this result to three dimensions case, we obtain the vector equality

$$\frac{d}{dt}\left(\hbar\vec{k}\right) = \vec{F}\,. \qquad (4.1.110)$$

As seen from this equation, the value $\hbar\vec{k}$ for electron in a crystal varies with time under the influence of external force in the same way as the particle momentum in classical mechanics (dP/dt) = F. Despite this, as noted in the case of phonons, $\hbar\vec{k}$ can not be identified with the momentum of an electron in a crystal, since the vector components are defined accurately within constant terms of the form $(2\pi/a)n_i$ (here a is lattice parameter, n_{i} = 1, 2 , 3, ...). However, within the first Brillouin zone, the quasi-momentum $\hbar\vec{k}$ has all the properties of the momentum.

We now calculate the acceleration *a*, acquired by an electron as a quasi-particle under the influence of an external force *F*. For the one-dimensional task

$$a = \frac{dv}{dt} = \frac{d}{dt}\left(\frac{1}{\hbar}\frac{dE}{dk}\right) = \frac{1}{\hbar}\frac{d^2E}{dk^2}\frac{dk}{dt}\,. \qquad (4.1.111)$$

At the calculation of the acceleration, we take into account that the electron energy is a function of time $E = E(k(t))$. Given that $\frac{dk}{dt} = \frac{F}{\hbar}$, we get

$$\frac{d\vec{v}}{dt} = \frac{1}{\hbar^2}\frac{d^2E}{d\vec{k}^2}\cdot\vec{F} \tag{4.1.112}$$

or

$$\vec{F} = \hbar^2 \frac{1}{\left(d^2E \big/ d\vec{k}^2\right)}\frac{d\vec{v}}{dt} = m^*\frac{\partial v}{\partial t}. \tag{4.1.113}$$

Comparing the expression (4.1.113) with Newton's second law, it is easy to see that the electron in a crystal moves under the influence of an external force such as a free electron in a vacuum would move under the impact of the same forces, if it has mass

$$m^* = \frac{\hbar^2}{\left(d^2E \big/ d\vec{k}^2\right)}. \tag{4.1.114}$$

The value of m* in (4.1.114) is called *the effective mass* of an electron in a crystal. Strictly speaking, the effective electron mass is irrelevant to real mass of the electron. It is *characteristic of the electrons in the crystal as a whole*.

By introducing the concept of effective mass, we describe real electron in a crystal as some new free quasiparticle having only two physical parameters of the real electron - its charge and spin. All other parameters - the quasi-momentum, effective mass, kinetic energy, etc. – are determined by interactiin of electron with the crystal lattice. Such quasi-particle should be called as *quasi-electron*, to emphasize its difference from the real electron in a crystal.

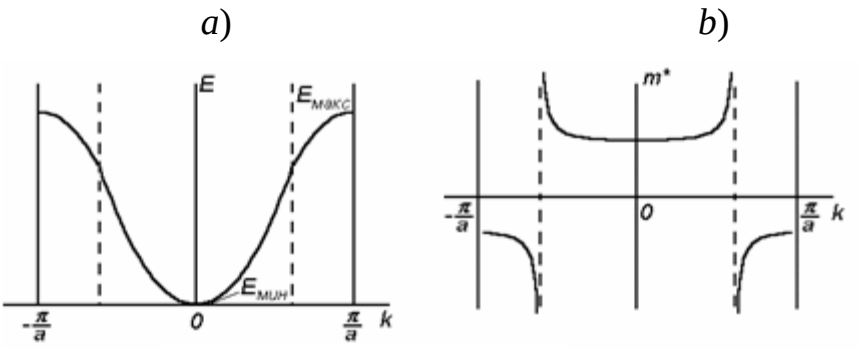


Fig. 4.1.52. The dispersion law (*a*) and dependence of the effective mass on the wave number (*b*) for an electron in a crystal

Features of the electron effective mass are connected with the type dispersion relation of the electron in the crystal (Fig. 4.1.52a). The dependence of the electron effective mass on its wave vector is shown in Fig. 4.1.525b. For electrons, which are located at the bottom of the energy band (in the center of the first Brillouin zone), the dispersion relation can be approximately described by the parabolic law

$$E = E_{\min} + Ck^2, \qquad C > 0. \tag{4.1.115}$$

Since the second derivative in this case equals $\frac{d^2E}{dk^2} = 2C > 0$, therefore, the effective mass is positive and practically independent on the energy of the electron (Fig. 4.1.52a). Behavior of such electrons in an external electric field is qualitatively similar to the free electrons in a vacuum: they are accelerated by an external electric field. The difference between these electrons and the free ones is that their effective mass may differ substantially from the free electron mass. For many metals, in which the concentration of electrons in the partially filled band is small, and they are located near its bottom, the conduction electrons behave similarly. Moreover, if these electrons are weakly bound to the crystal, their effective mass is merely slightly different from the real electron rest mass.

For electrons that are at the top of the energy band (near the edge of the first Brillouin zone), the dispersion relation can be loosely described by the inverted parabola

$$E = E_{\max} - D\left(\frac{\pi}{a} - k\right)^2, \qquad D > 0, \tag{4.1.116}$$

so that the effective mass becomes negative (Figure 4.1.52b). Such behavior of the electron effective mass can be explained as follows: during its movement by the crystal the electron has not only a kinetic energy E_{k} of its translation motion, but also the potential energy U due to its interaction with the crystal lattice. Therefore, part of the work A of external force can go into kinetic energy and change it to the value of ΔE_{k}, while another part to the potential energy ΔU:

$$A = \Delta E_{\text{K}} + \Delta U \tag{4.1.117}$$

If at the motion of an electron, not only the whole work of the external force turns into the potential energy, but also some part of its kinetic energy ($\Delta E_{\text{к}} < 0$),

then its velocity will decrease. In this case the electron behaves as a particle with a negative effective mass (or with a positive charge!). In the case where all of the work of an external force turns into potential energy ($\Delta E_к = 0$), the increment velocity and kinetic energy does not occur and electron behaves as a particle with an infinitely large effective mass (Figure 4.1.52b). Infinitely large effective mass of the electron corresponds to the inflection points of the dispersion curve (in Fig. 4.1.52a they are indicated by dashed lines).

Note that dispersion law for the free electron is

$$\varepsilon = \frac{\hbar^2 k^2}{2m_0}, \tag{4.1.118}$$

whereas for the electron in a crystal in the effective mass approximation, when taking into account (4.1.115), it will be of the form

$$\varepsilon = \frac{\hbar^2 k^2}{2m^*}. \tag{4.1.119}$$

4.1.6. Zone structure and statistics of semiconductors

As was shown in the part 4.1.5, band structure of defect-free dielectric and semiconducting crystalline materials are very similar having the less value of the forbidden band gap E_g in semiconductors. Consider the consequences to which this decrease of E_g leads.

Band structure in defect-free (intrinsic) semiconductors. Let us consider the crystal at $T = 0$ K with the valence band completely filled with electrons (Fig. 4.1.53a) and $E_g \approx 10kT$. In the case, when heating, part of electrons overcomes the band-gap Eg and can be transferred from the fully filled valence band (V-band) into the conduction band (C-band) as the result of interaction with the oscillating atoms of the crystal lattice (interaction with phonons), see Fig. 1.4.2b. As a result, the C-band states, occupied by electrons, appear, and the same number of empty states (holes) appears in the V band. This process is called *generation of electron-hole pairs* in the process of *band-to-band transitions*. Just such substances exhibit semiconducting properties, and defect-free semiconductors themselves are called *intrinsic*.

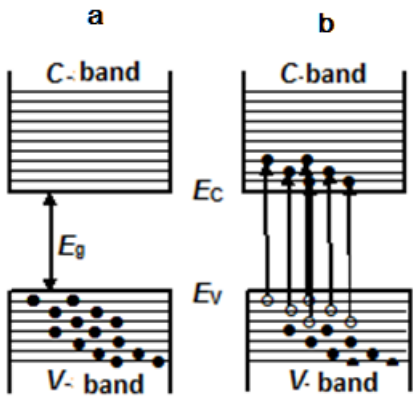


Fig. 4.1.53. Filling the C and V bands by electrons in the defect-free (ideal) semiconductor at $T = 0$ K (*a*) and at finite temperature (*b*). Black circles are electrons and white circles – holes.

In intrinsic semiconductors, when applying the external electric field, the electrons are rearranged by states (move up by the energy levels), as shown in Fig. 4.1.53b, in both conduction and valence bands. Moreover, if the electrons in *C*-band behave as a negatively charged particles with a positive effective mass, the holes in V-band act either as particles with a negative effective mass or as a positively charged particles with a positive effective mass. Furthermore, since there are exponentially few electrons in conduction band and holes in valence band, the defect-free semiconductor will conduct electric current worse than metal but better than dielectric.

As mentioned, the considered in Fig. 4.1.53 band structure can be realized only in the case of perfectly pure and defect-free (intrinsic) semiconductor crystals. Conductivity of intrinsic semiconductor is called the *intrinsic conductivity*.

The influence of defects on the band energy spectrum of electrons in semiconductors. In real crystalline semiconductors, impurities and other defects are always present, some of which have a substantial influence on their conductivity. For example, the addition to silicon of boron atoms in amount of one atom per 10^5 silicon atoms (10^{-3} at.%) results in the increase of its conductivity at room temperature in 1000 times. Semiconductors containing impurities are called *impurity (extrinsic) semiconductors*, and their electrical conductivity is called, respectively, the *impurity conductivity*.

Let us consider the changes that different types of impurities create in band structure and conductivity of semiconductor (for example, like silicon or germanium) at $T = 0$ K and when heating.

In pure silicon crystal, each atom uses four valence electrons for the formation of covalent bonds with four the nearest neighbors (Fig. 4.1.54a). Suppose that the pentavalent impurity atoms (e.g., phosphorus) are introduced into silicon crystal lattice. They are located in lattice sites (Fig. 4.1.54b) substituting the silicon atoms. Then, as in pure silicon, 4 electrons of each phosphorus atom

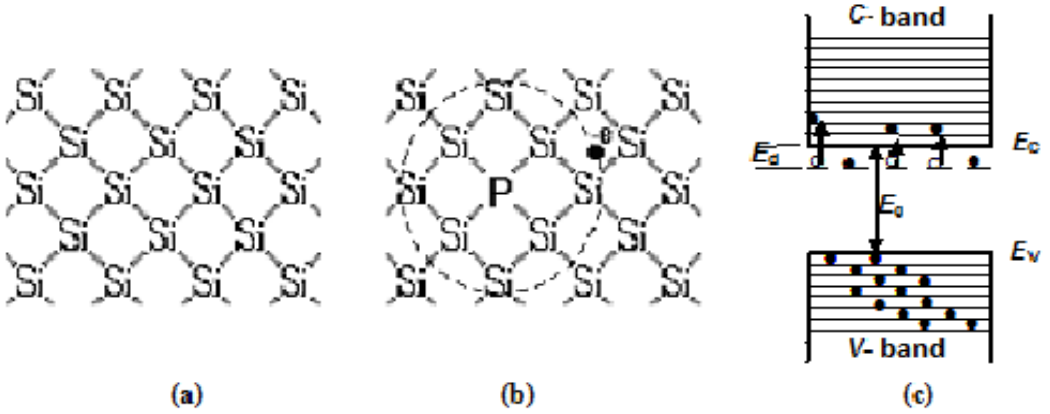


Fig. 4.1.54. Scheme of formation of electronic states in the semiconductor silicon: *a* – the covalent bonds in pure silicon; *b* – silicon with an impurity atom of phosphorus; *c* – band structure of silicon with donor impurities.

form 4 covalent bonds with 4 nearest neighboring silicon atoms but the fifth is not involved in the chemical bond. As a result, the extra electron is connected with the phosphorus atom much weaker than other 4 electrons bounded to silicon atoms. This means that at $T = 0$ K, this fifth electron is said to be localized on the impurity.

Such behavior of the phosphorus atom means that it will cause the appearance of energy levels in the band gap of silicon at the depth E_d below the bottom of C-band (Fig. 4.1.54c). Since these levels belong to electrons localized near the impurity atoms, they are represented in the band diagram by the dashed lines.

It will be easy to tear this extra electron away from the phosphorus atom (to make it free) in case of heating of the crystal, permitting its free motion through the crystal when an electric field is applied. In terms of band structure, this means that the electron can move up, getting energy from phonons, from the impurity level to the empty levels of C band (leaving localized hole at the impurity level in the gap). Such an electron can participate in the creation of an electric current when an external electric field is applied. Such an impurity which gives electrons to the C-band is called *donor impurity*, and semiconductors with such impurities are called *electronic semiconductors* or *n-type semiconductors*. E_d energy that must be expended to move the electron from the center of the donor type impurity into the conduction band is called the *ionization energy of donor impurity*. The donor center becomes positively charged after electron separation, since the hole is localized at it.

Atoms of the fifth group in the Mendeleev periodic table, such as phosphorus (P), arsenic (As) and antimony (Sb), are the most common donor impurities in silicon and germanium crystals.

Impurity atoms from the third group of the periodic table, such as boron (B), aluminum (Al), gallium (Ga) and indium (In) behave itself differently in silicon and germanium. For example, the substitution of one silicon atom in the silicon lattice by boron atom leads to the fact, that covalent bond of one of the 4 silicon atoms closest to the boron atom remains unfilled. This bond can be restored if one electron from the silicon-silicon chemical bond passes to the boron atom forming electron vacancy, or hole (see Fig. 1.4.4a). In the band diagram, this corresponds to the appearance of local boron impurity levels in the band gap of silicon near the top of the V band (Fig. 1.4.4b). At T = 0 K, this level is free, but electrons from the V-band can fill it if the crystal is heated. The holes formed in the valence band are the carriers of electric current in this type of impurity semiconductors.

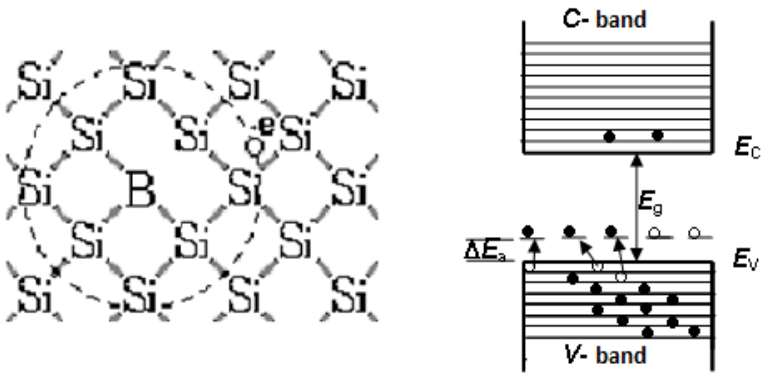


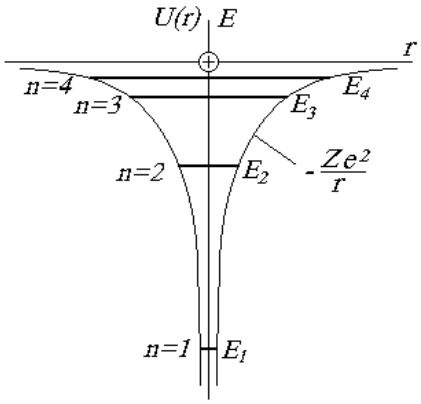
Fig. 4.1.55. Scheme of formation of the hole states and conductivity of the semiconductor silicon: *a* – silicon with boron impurity atom, *b* – band structure of silicon with acceptor impurities

Impurities capturing electrons from the valence band of semi-conductors are called the *acceptor impurities*, and the energy levels of these impurities are *acceptor levels*. Difference between the acceptor level energy and the energy of V-band top is called the *ionization energy of acceptor impurity* E_a. Semiconductors containing acceptor impurities are called *hole semiconductors* or *p-type semiconductor*s. After electron transfer from the silicon to acceptor center (formation of a free hole in the valence band), the latter becomes negatively charged.

Hydrogen-like impurities in the crystalline semiconductors. The binding energy of impurities in a semiconductor crystal (ionization energy of the donor or acceptor) can be estimated on the basis of a simple model, similar to the Bohr model of the electron in hydrogen atom. So these impurities are called *hydrogen-like impurities*.

The potential energy of an electron in a hydrogen atom is determined by a spherically symmetric field of its interaction with the kernel (Figure 4.1.56.):

$$U(r) = -\frac{Ze^2}{r}, \tag{4.1.120}$$



where Z - the atomic number of the atom, r - distance of the electron from the nucleus. For an electron bound to a hydrogen atom, complete its energy is negative (E < 0), whereas for the electron moving freely outside the atom E > 0.

Fig. 4.1.56. Energy diagram for a hydrogen-like atom

The Schrodinger equation for the valence electron in such an atom has the form

$$\Delta\Psi(x,y,z) + \frac{2m}{\hbar^2}\left(\varepsilon + \frac{Ze^2}{r}\right)\Psi(x,y,z) = 0\,. \tag{4.1.121}$$

For E <0 when electron is inside the atom, the equation (4.1.121) has finite and continuous solutions only for discrete energy values

$$\varepsilon_n = -\frac{m_e e^4 Z^2}{2\hbar^2 n^2}, \quad n = 1,2,3,... \tag{4.1.122}$$

The wave function $\Psi_1(r)$ for the state with the lowest energy (n = 1) in the spherically symmetric case has the form

$$\Psi_1(r) = \sqrt{\frac{1}{\pi r_B^3}\exp\left(-\frac{r}{r_B}\right)}, \tag{4.1.123}$$

where $r_B > 0$ is the so-called Bohr radius of the electron in a hydrogen atom.

The function $\Psi_1(r)$ determines the bulk density of the probability to find an electron in the space. More visual representation can be obtained using radial probability density. This value is introduced so that the product of the electron detection probability determined at a distance from nucleus between r and $r + dr$. The calculations show that the ionization energy of a hydrogen-like donor (acceptor) in eV can be expressed by the relation

$$\rho_r(r) = \frac{4}{r_B^3} r^2 \exp\left(-\frac{2r}{r_B}\right). \tag{4.1.124}$$

As can be seen from (4.1.124), the function $\rho_r(r)$ has a maximum at $r = r_B$. For the hydrogen atom numerical value значение $r_B \approx$ 0,05 nm coincides with the radius of the first Bohr orbit. This means that the first Bohr orbit corresponds

to a distance from the nucleus, when the probability of finding an electron is maximal.

For quantitative estimates of ionization energy for the impurity atom in a semiconductor we can use the Eq. (4.1.124), which, however, should be somewhat transformed. According to the hydrogen-like model of impurity atoms (for example, phosphorous in silicon), an electron is moving by a circular orbit in the Coulomb force field like the electron moving around the nuclei of ion impurities in of the hydrogen atom. The difference is in that the field of impurity ion is weakened by dielectric properties of semiconductor crystal (crystal polarization around ion impurity according the Bethe hypothesis). This effect is taken into account by the value of the dielectric constant of the crystal, which can be varied for semiconductors typically from 5 to 2000. It is necessary also to take into account the fact that the effective mass of an electron in a crystal is different from the free electron mass. In this case, considering the dielectric constant of the semiconductor ε and replacing the free electron mass m by its effective mass m_e* in crystals, we use instead of (4.1.122) for the energy of the electron in the hydrogen atom the following expression for the donor impurity ionization energy:

$$E_d = \frac{e^4 m^*}{2\left(4\pi\varepsilon\varepsilon_0\hbar\right)^2}. \qquad (4.1.125)$$

According to (4.1.122), the ionization energy of the free hydrogen atom is 13.6 eV. Then, in accordance with formula (4.1.125), we can obtain the value E_d in form

$$E_d = 13{,}53\frac{m_n^*}{m\varepsilon^2}, \text{ eV} \qquad (4.1.126)$$

Since the dielectric constant of silicon ε = 11,7, and m * / m ≈ 0,2, we get Ed ≈ 0,02 eV, which is very close to the experiment (see. Table. 1.4.1). Bohr radius of the electron in the hydrogen-like donor impurities will look like:

$$a_d = \frac{\varepsilon_0 h^2}{m_n^* e^2} = \frac{h^2}{me^2}\frac{\varepsilon_0 m}{m_n^*} = r_B\frac{m\varepsilon_0}{m_n^*}, \qquad (4.1.127)$$

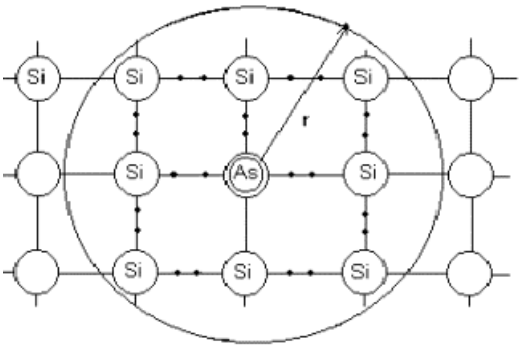


Fig. 4.1.57. To clarification of the physical meaning of the impurity Bohr radius in the semiconductor

that gives a few nanometers for a_d, which is substantially higher than electron radius r_B in hydrogen atom)). Since, as noted above, because of the high dielectric constant of the semiconductor Coulomb attraction of extra electron in donor impurity (As in Fig. 4.1.57) is largely reduced, so that the radius of the electron orbit is larger, involving several interatomic distances.

The ionization energy of the acceptor is described by the same equation:

$$E_a = 13{,}53\frac{m_p^*}{m\varepsilon_0^2}, \text{ eV}, \tag{4.1.128}$$

and the Bohr radius of the acceptor of hydrogen-like impurities in this case would be:

$$a_a = r_B\frac{m\varepsilon_0}{m_p^*}. \tag{4.1.129}$$

The experimental values of the ionization energies of some types of acceptors and donors in silicon and germanium are presented in Table. 4.1.1.

Table 4.1.1. The experimental values of the ionization energies for hydrogen-like impurities in silicon and germanium

Impurities	Ionization energies, eV	
	Gemanium	Silicon
Donors		
P	0,0120	0,044
As	0,0127	0,049
Sb	0,0096	0,039
Bi		0,069
Axceptors		
B	0,0104	0,045
Al	0,0102	0,057
Ga	0,0108	0,065
In	0,0112	0,16

Electrical conductivity of intrinsic and impurity semiconductors. According to the band structure of intrinsic semiconductor (see Fig. 4.1.56), its conductivity depends on the concentration of generated electrons and holes (moving in electric field in the opposite directions) and can be expressed by the following equation:

$$\sigma = \sigma_n + \sigma_p = ne\mu_n + pe\mu_p . \qquad (4.1.130a)$$

Here σ_n and σ_p are contributions to the conductivity of electrons and holes, n and p –their concentrations, μ_n and μ_p – mobility of electrons and holes, respectively. So, intrinsic conductivity in (4.1.65б) consist of two components – electron (first contribution) and hole (second contribution).

As can be seen, the value of the conductivity of the semiconductor and its temperature dependence are determined by concentration of carriers (electrons and holes) and their mobility, which in turn depend on the type of semiconductor. Electronic conduction component is determined by the first term in the formula (4.1.130a), and the second term is due to hole conductivity of the semiconductor.

Since in the intrinsic semiconductor electron and hole concentrations are the same (since $n = p = n_i = p_i$, where n_i and p_i are intrinsic concentrations of electrons and holes), the conductivity of intrinsic semiconductor will be equal to

$$\sigma = n_i e(\mu_n + \mu_p). \qquad (4.1.130b)$$

Conductivity of impurity semiconductors is determined by the concentration of charge carriers in corresponding allowed band (electrons in C-band or holes in V-band) and their mobility. In doing so, carrier concentration depends on the level of doping (impurity concentration) and temperature.

The main source of electrons in *n*-type semiconductor at low temperatures are donor impurities. When heating of impurity semiconductor (by thermal excitation), electrons pass from the donor levels in band-gap into the C band, so that the conductivity is defined by

$$\sigma_n = ne\mu_n . \qquad (4.1.131a)$$

At very high temperatures, when the electrons are excited into the C band due to their transition from the V band, the electronic semiconductor behaves like an intrinsic semiconductor (see equation (4.1.130).).

For the hole semiconductor, the following relation is true

$$\sigma_p = pe\mu_p . \qquad (4.1.131b)$$

Electrical neutrality of semiconductors. As noted in the preceding section, the impurity conductivity of semiconductors is caused by formation of some free electrons or holes in C- or V-band, respectively, due to the influence of temperature. Concentration of generated free charge carriers is determined by the balance between impurity-to-band and band-to-impurity transitions of electrons. First process is called *generation of carriers* and their return back - *carriers recombination*.

At steady process of generation and recombination of electrons in the *n*-type semiconductor their equilibrium concentration on donors (in this case, the donor atoms are electrically neutral) will be equal

$$N_d^0 = n_d = N_d - N_d^+ = N_d - p_d , \qquad (4.1.132)$$

where N_d is total concentration of donors (donor levels), n_d – electron density at the donors, N_d^+ – concentration of charged donors (which is equal to concentration of holes p_d localized on them).

In equilibrium state, the concentration of positive charges (localized holes and positively charged donors) and negative charges (free electrons and negatively charged acceptors) must be equal. Only in this case, the semiconductor is electrically neutral.

The electrical neutrality condition for in n-type semiconductor at low temperatures, when $kT << E_d$, and no band-to-band transitions with the formation of electron-hole pairs (as in the intrinsic semiconductor), will have the form

$$n = N_d^+ = p_d. \qquad (4.1.133)$$

The electrical neutrality condition of for electronic semiconductor at high temperatures ($kT >> E_d$) is equal

$$n = p_d + p , \qquad (4.1.134)$$

where

$$p_d = N_d^+ = N_d . \qquad (4.1.135)$$

Equilibrium concentration of holes localized on acceptors in *p*-type semiconductor in steady generation-recombination process equals

$$N_a^o = p_a = N_a - N_a^- = N_a - n_a, \tag{4.1.136}$$

where N_a is the concentration of acceptor levels, p_a – the holes concentration at acceptor levels, N_a^- – concentration of charged acceptors (which is equal to the concentration of electrons n_a localized on them).

Then the electrical neutrality condition for hole semiconductor at low temperatures
($kT << E_a$) is

$$p = N_a. \tag{4.1.137}$$

The *condition of electrical neutrality for hole semiconductor at high temperatures*
($kT >> E_a$) is

$$p = n_a + n. \tag{4.1.138}$$

Electroneutrality condition in a semiconductor doped with both types of impurities (such a semiconductor is called *compensated*):

$$p + p_d - n - n_a = 0. \tag{4.1.139}$$

The condition of electrical neutrality in an intrinsic semiconductor:

$$p_i = n_i \text{ or } p_i - n_i = 0. \tag{4.1.140}$$

For the equilibrium concentrations of electrons and holes in any semiconductor, the following relation is true

$$p_i^2 = n_i^2 = np, \tag{4.1.141}$$

which is called the *mass action law*. The assumption that degree of filling of the energy levels by charge carriers is much less than unity was used in the derivation of this law. This gas of carriers is called *non-degenerate gas*, and semiconductors are called *non-degenerate semiconductors*, respectively.

Equilibrium concentration of charge carriers in defect-free semiconductor crystals. Let us first analyze the temperature dependence of concentration of intrinsic carriers, which will allow to describe temperature dependence of electrical conductivity in pure (defect-free) semiconductors. Since, there is no localized levels in the band gap, electrons can occupy levels in C-band only due to band-band transitions (see Fig. 4.1.53b). So their concentration in C-band will be determined by the known relation

$$n = N_C \exp\left(-\frac{\Delta E_g}{2k_B T}\right), \tag{4.1.150}$$

i.e. near the bottom of C-band electrons behave almost as an ideal gas, for which the kinetic energy is $E = p^2/2m$. Thus, as was shown in Section 1.4 the same concentration of free holes

$$p = N_V \exp\left(-\frac{E_g - \varepsilon_F}{2k_B T}\right) \tag{4.1.152}$$

is generated, in V-band. Here N_C and N_V are effective densities of states in C-and V-bands, correspondingly.

Multiplying the equations (4.1.150) and (4.1.152), we get another form of the law of mass action:

$$np = n_i^2 = N_C N_V \exp\left(-\frac{E_g}{kT}\right) = 4\left(\frac{kT}{2\pi^2\hbar^2}\right)^3 \left(m_p^* m_n^*\right)^{3/2} \exp\left(-\frac{E_g}{kT}\right) \tag{4.1.154}$$

The Fermi level in an intrinsic semiconductor. Using the equality $n = p$ and relations (4.1.150) and (4.1.152), we obtain:

$$N_C \exp\left(\frac{\varepsilon_F}{kT}\right) = N_V \exp\left(-\frac{E_g + \varepsilon_F}{kT}\right). \tag{4.1.155}$$

Hence

$$\varepsilon_F = -\frac{E_g}{2} + \frac{kT}{2}\ln\frac{N_V}{N_C} = -\frac{E_g}{2} + \frac{3}{4}kT\ln\frac{m_p^*}{m_n^*}. \tag{4.1.156}$$

This relation means that at $T = 0$ Fermi level of an intrinsic semiconductor is located exactly in the middle of the band gap, if the effective masses of electrons and holes are the same (since for $m_p^* = m_n^*$, $\ln\left(\frac{m_p^*}{m_n^*}\right) = 0$): $\varepsilon_F(0) = -\frac{E_g}{2}$.

Substituting (4.1.156) to one of the relations (4.1.150) and (4.1.152), we obtain an expression for intrinsic carrier concentration in semiconductor

$$n_i = p_i = \sqrt{N_C N_V}\,\exp\left(-\frac{E_g}{2kT}\right). \qquad (4.1.157)$$

Temperature dependence of Fermi level in an intrinsic semiconductor is determined by the second term in equation (4.1.156). If the effective mass of holes in the V band is more than effective mass of electron in the C band, Fermi level is shifted with temperature closer to the bottom of the C band. In the opposite case, Fermi level moves toward the V band (Fig. 4.1.59).

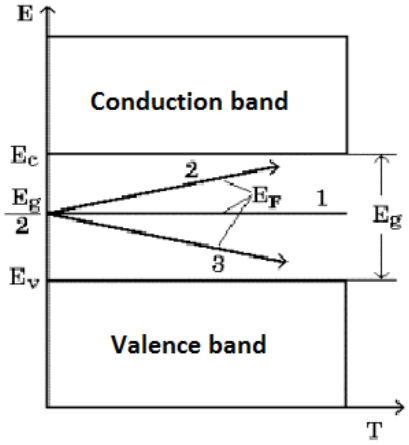


Fig. 4.1.59. Schematic view of temperature dependence of the Fermi level position in an intrinsic semiconductor: 1 - $m_p^* = m_n^*$; 2 - $m_p^* > m_n^*$; 3 - $m_p^* < m_n^*$.

For most semiconductors, hole effective mass is not much greater than the electron effective mass, and the shift of the Fermi level with temperature is insignificant. However, for example, InSb has the ratio $\frac{m_p^*}{m_n^*} \approx 10$, and band gap is small (E_g = 0.17 eV). Therefore its Fermi level is in the conduction band at T > 450 K. At this temperature, the semiconductor passes into degenerate state and behaves like a metal.

Electrons and holes concentrations in doped semiconductors. All the most interesting properties of semiconductor crystals which allow to use them in devices manifest themselves in the case of doped semiconductors. Impurities, even in relatively low concentrations, strongly modify electrical, optical and other properties of semiconductors.

To estimate the equilibrium concentrations of carriers in bands in the presence of donor and acceptor levels, we have to known positions of the band edges E_V and E_C, the ionization energies (positions of impurities levels) for acceptors E_a and donors E_d, donor N_d and acceptor N_a concentrations, effective masses of electrons m_n and holes m_p in corresponding bands and temperature T.

Calculations allow us to find the carrier concentrations n and p in doped semiconductors, as well as the position of Fermi level separately (see below).

However, the expressions for them, in general, are complex and sometimes have an unclear physical meaning. So, we consider only a few special cases.

If impurity levels (for definiteness by donor impurity, see Fig. 4.1.55b) correspond to E_d of about few hundredths of eV, the impurity is strongly ionized even at room temperature, and at higher temperatures it is ionized completely. In the last case, n value is almost equal to the impurity concentration N_d.

Formula (4.1.157), obtained above for intrinsic semiconductors, is true for doped semiconductor at high temperatures, when transitions band-to-band are take place.

At low temperatures, part of electrons (n_d) occupy donor levels and the electron concentration satisfies the relation

$$n = \sqrt{\frac{N_d N_C}{2}} \exp\left(-\frac{E_d}{2kT}\right). \qquad (4.1.158a)$$

Similarly, for the hole-type semiconductor at low temperatures

$$p = \sqrt{\frac{N_a N_V}{2}} \exp\left(-\frac{E_a}{2kT}\right). \qquad (4.1.158b)$$

Sometimes crystals contain both donor and acceptor impurities with concentrations N_d and N_a. In this case, electrons should pass ("fall down") from donor to acceptor levels. So, if $N_d > N_a$, only $(N_d - N_a)$ donors can participate in creation of the electronic conductivity. Such semiconductors are called *partially compensated*. In this case, concentration of dominating charge carriers (electrons) at $E_d/2kT >> 1$ and not very low concentration of compensating impurity N_a looks as follows:

$$n = \frac{N_d - N_a}{2N_a} N_C \exp\left(-\frac{E_d}{kT}\right). \qquad (4.1.159)$$

Fermi level in doped semiconductor. The Fermi level position in doped semiconductors can also be found from the condition of electrical neutrality of the crystal (1.4.10). For electronic semiconductor this condition can be re-written as

$$n = p_d + p \; = p + (N_d - n_d), \qquad (4.1.164)$$

where N_d is concentration of donor levels, n_d – electron concentration at the donor levels. Concentration of electrons in V band is equal to the sum of concentrations of holes in the valence band and concentration of positively charged donor ions (the latter, obviously, is equal to N_d - n_d). So, concentration of electrons at donor level can be calculated by multiplying the concentration of these levels N_d on the function of Fermi-Dirac:

$$n_d = \frac{N_d}{1+\exp\left(\frac{E_C-\varepsilon_F-E_d}{kT}\right)}, \tag{4.1.165}$$

where E_d is activation energy of donor levels.

Substitution of concentrations of electrons (4.1.150) and holes (4.1.152) as well as charged donor levels (4.1.165) into the electrical neutrality condition (4.1.164) leads to the following equation with respect to Fermi level ε_F for non-degenerate electron gas:

$$N_C \exp\left(\frac{\varepsilon_F-E_C}{kT}\right) - N_V \exp\left(\frac{E_C-E_g-\varepsilon_F}{kT}\right) + \frac{N_d}{\exp\left(\frac{E_C-\varepsilon_F-E_d}{kT}\right)} = N_d \tag{4.1.166}$$

Usually the equation (4.1.164) is not solved in general case because of its complexity. So below we are limited to reviewing special cases, for example, low or high temperatures:

$$\begin{cases} \varepsilon_F(T) = -\frac{E_d}{2} + \frac{kT}{2}\ln\frac{N_d}{2N_C} & for \quad kT << E_d \\ \varepsilon_F(T) = -\frac{kT}{2}\ln\frac{N_d}{2N_C} & for \quad kT >> E_d \end{cases} \tag{4.1.167}$$

From the equation (4.1.167) follows that, at absolute zero temperature, Fermi energy of electronic semiconductor is located exactly in the middle between the bottom of the conduction band and the position of donor levels. In so doing, temperature dependence of Fermi level is determined by the second term in the equation (4.1.167), which changes its sign with temperature. So, firstly, Fermi level moves toward the C band bottom with temperature increasing, and then (when impurity semiconductor is heated to high temperatures)– to the V band, as in intrinsic semiconductor (Fig. 4.1.53a). The free carriers, excited from impurity

levels, are called *the majority carriers*, while carriers with opposite sign generated due to band-band transitions are *the minority carriers*.

Similarly, we can obtain the expressions for temperature dependences of Fermi level in p-type semiconductors:

$$\begin{cases} \varepsilon_F(T) = -\dfrac{E_a}{2} + \dfrac{kT}{2}\ln\dfrac{N_a}{2N_V} & \text{for} \quad kT << E_a \\ \varepsilon_F(T) = -\dfrac{kT}{2}\ln\dfrac{N_a}{2N_V} & \text{for} \quad kT >> E_a \end{cases} \tag{4.1.169}$$

Schematic view of the temperature dependences of Fermi level in doped semiconductors is shown in Fig. 4.1.61.

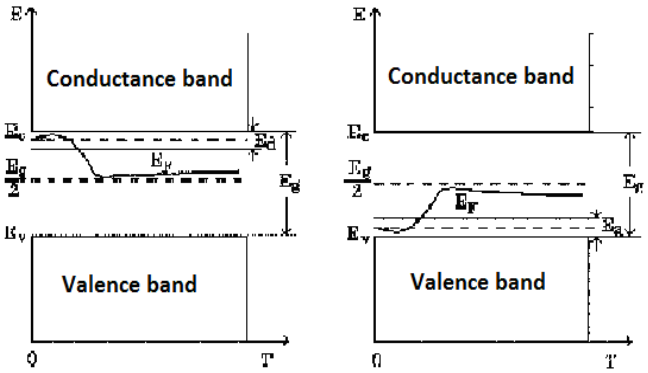


Fig. 4.1.61. Schematic view of temperature dependence of the Fermi level position in n- (*a*) and p-type (*b*) semiconductors

4.2. Semiconductor optical detectors

In order to become less dependent on the consumption of fossile fuels, in a lot of countries efforts are made to produce electrical power using renewable energy sources. When generating renewable electrical power, besides hydropower, tidal power, wind energy, geothermal energy or the use of bioenergy, also solar photovoltaics are gaining importance. Using photovoltaic cells, solar energy is converted into electrical energy.

In earlier days, photovoltaic cells were mainly used on remote places where no other electrical energy sources were available (e.g. satellites circulating around the Earth, light buoys at sea, private households at remote places where no public electrical grid is available,…). Nowadays, photovoltaic cells are also often installed on the roofs of private households (connected with the public low voltage grid) or on industrial sites (including entire solar parks). Due to the increased importance of photovoltaic energy, it is important to know and predict the power

production which is closely related with the solar irradiance (the solar radiation flux density expressed in W/m^2).

4.2.1. The voltage current characteristic of a photovoltaic panel

In general, a photovoltaic panel contains a number of photovoltaic cells connected in series. Figure 4.2.1 visualises typical voltage current characteristics of a photovoltaic panel (at a constant temperature of $25°C$) with different solar irradiances. For instance an incident solar power of $500 - 1000\ W/m^2$ corresponds with clear and sunny weather, an incident power of $120 - 500\ W/m^2$ corresponds with sunny partially clouded weather, an incident power of $50 - 120\ W/m^2$ corresponds with really clouded weather.

As Figure 4.2.1 visualises, the voltage current characteristic strongly depends on the solar irradiance. The higher the power available in the solar radiaton, the higher the generated voltage and (especially) the higher the generated current. Figure 4.2.2 visualises a single voltage current characteristic with a constant solar irradiance at a constant temperature. Notice the short circuit current I_{SC} when the photovoltaic panel has been short circuited. Notice the open circuit voltage U_{OC} when the photovoltaic panel has no electrical load. The

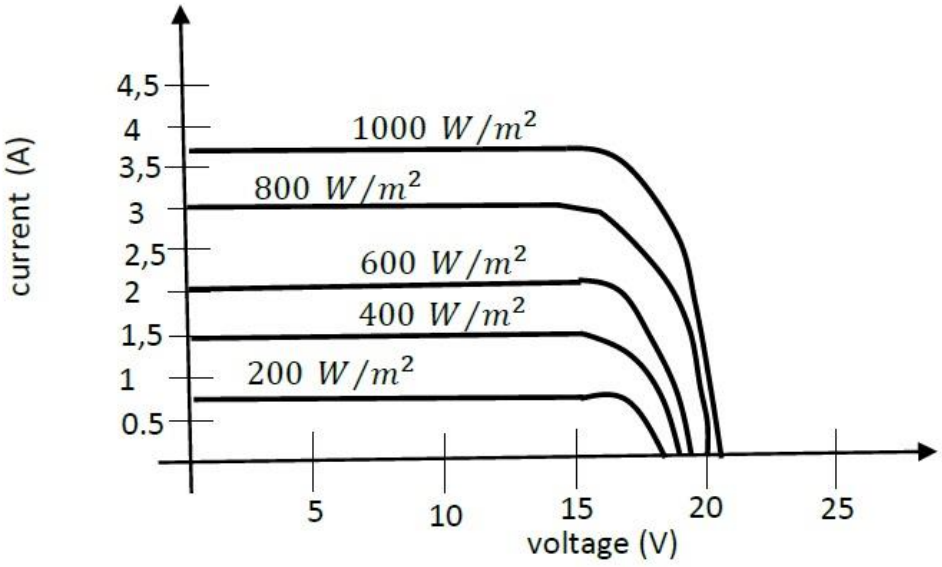


Fig. 4.2.1. Voltage current characteristics of a photovoltaic panel

electrical power is maximum in the "maximum power point" indicated in Figure 2. In this situation a current I_{MPP} and a voltage U_{MPP} account for a maximum power $P = U_{MPP}I_{MPP}$. When combining Figure 4.2.1 and Figure 4.2.2, it is clear the maximum power $P = U_{MPP}I_{MPP}$ strongly depends on the solar irradiance.

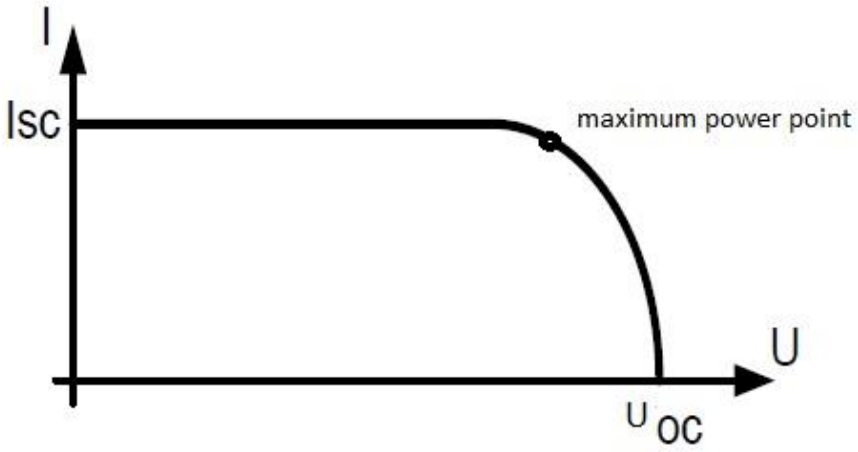


Fig. 4.2.2. The maximum power point

By measuring the solar irradiance on a site over a sufficiently long period of time, it is possible to predict the future irradiances and to predict more or less the power production of a photovoltaic installation. By combining a photovoltaic installation with solar irradiance measurements, it is possible to detect deficiencies in the installation. Based on the solar irradiance measurement, the normal power production is known and if the real power production is much smaller, something is wrong with the installation.

4.2.2. The use of a pyranometer and a pyrheliometer

A pyranometer measures the total solar irradiance in a hemisphere of the field of view. A pyranometer measures all solar irradiation i.e. diffuse and direct radiation will be measured together. In general, the pyranometer is mounted in the same plane as the photovoltaic panel. The pyranometer converts the solar irradiance (the power per unit area expressed in W/m^2) in an electrical voltage and this voltage is sampled by a data acquisition system. By converting the analog voltages to digital values, these digital values can be stored in a memory.

The spectral solar irradiance

Figure 4.2.3 visualises the distribution of the solar radiation. Figure 4.2.3 visualises the spectral irradiance (noted as "energy distribution") of a black body having a temperature of $6000\,K$ which corresponds with the temperature at the surface of the Sun. Figure 4.2.3 also visualises the extraterrestrial spectral irradiance (outside the atmospher of the Earth) due to the Sun (indicated as $AM0$ radiation). Since part of this radiation is absorbed by the atmosphere of the Earth (especially due to water vapour and CO_2), the spectral irradiance is reduced to the spectral irradiance indicated as $AM1.5$. The notation AM stands for "Air Mass". In case of $AM1.5$, the factor 1.5 indicates the solar rays travel throughout the atmosphere and need a distance to reach the surface of the Earth which equals 1.5 times the distance travelled by the solar rays when the Sun is directly overhead at sea level (i.e. the height of the atmosphere perpendicular to the Earth surface).

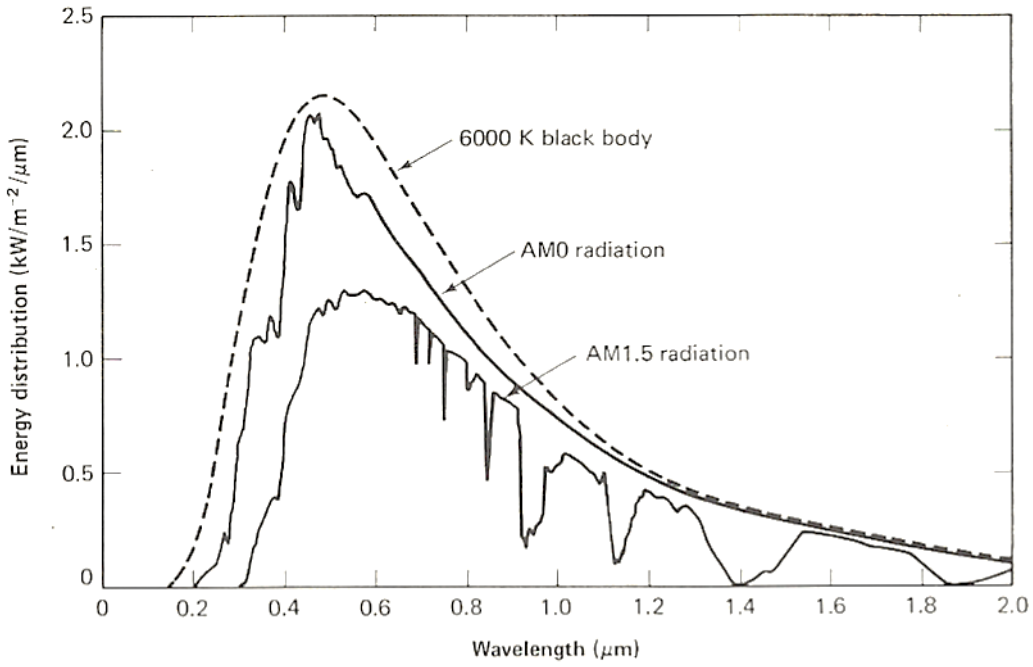


Fig. 4.2.3. The spectral distribution of the solar radiation[1]

Notice the spectral irradiance in Figure 4.2.3 is visualised as a function of the wavelength of the solar radiation (and expressed in $kW/m^2/\mu m$). By integrating the $AM0$ characteristic, a total solar irradiance of $1367\ W/m^2$ is obtained. By integrating the $AM1.5$ characteristic, a total solar irradiance of $1000\ W/m^2$ is obtained (notice this $1000\ W/m^2$ and an $AM1.5$ spectrum is

[1] Source:
http://solarwiki.ucdavis.edu/The_Science_of_Solar/Solar_Basics/B._Basics_of_the_Sun/III._Solar_Radiation_Outside_the_Earth's_Atmosphere

often taken as a reference when performing measurements on photovoltaic panels).

The electromagnetic radiation of the sun contains a whole range of frequencies corresponding with a whole range of wavelengths. When considering the $AM1.5$ spectrum, the wavelengths range from approximately $0.3\,\mu m = 300\,nm$ to $2\,\mu m$ or even $2.8\,\mu m = 2800\,nm$. This spectrum contains UV-light (having wavelengths below $400\,nm$), visible light and infrared radiation (having wavelengths larger than $700\,nm$).

Pyranometers and pyrheliometers

Figure 4.2.4 visualizes a pyranometer containing a glass dome to allow the solar radiation enter the pyranometer and allow the actual sensor performing the measurement. There exist several types of pyranometers. A distinction can be made between

- thermopile pyranometers,
- photodiode based pyranometers,
- photovoltaic pyranometers

and these different types will be discussed later on.

As already explained, a pyranometer measures diffuse and direct radiation together. By adding a shading device which blocks the direct radiation, only the diffuse radiation will be measured. By measuring the total radiation and also measuring only the diffuse radiation, the direct radiation can be calculated by subtracting both measurements.

Direct radiation can also be measured using a pyrheliometer. It is important the pyrheliometer is pointed in the direction of the Sun which requires a sun-tracking device.

Fig. 4.2.4. Pyranometer

4.2.3. The thermopile pyranometer

The use of a thermopile

A thermopile converts a temperature difference into an electrical voltage. More precisely, a thermopile consists of a number of thermocouples which are connected in series. A thermocouple consists of two different conductors (for instance copper and iron or copper and constantan) which are connected with each other. Such a thermocouple produces a temperature dependent voltage based on the Seebeck effect. By having two junctions between these two different conductors (one junction at a lower temperature and another junction at a higher temperature), the temperature difference is measured. Since the sensitivity of a thermocouple is quite limited, they are well suited to measure high temperature differences. If it is necessary to measure smaller temperature differences, a larger sensitivity is needed which is obtained by mounting several thermocouples in series (giving a thermopile). A thermopile typically contains 50 to 100 junctions.

The behaviour of a thermopile pyranometer

In case of a pyranometer, the hot side of the thermopile can be obtained using a black sector which absorbs the radiation of the Sun implying a temperature increase. The larger the irradiance, the larger the increase of the temperature. The cold side of the thermopile is connected with white sectors (or shadow area) which do not undergo a temperature rise due to the irradiance of the Sun. This way, a temperature difference is obtained which is (approximately) proportional with the solar irradiance to be measured.

The thermopile pyranometers are able to measure the irradiance while covering a broad spectrum ranging from $300\ nm$ to $2800\ nm$ with a mainly flat spectral sensitivity. This means a broadband irradiance measurement is obtained (approximately from $300\ nm$ to $2800\ nm$). Figure 4.2.5 not only visualises the solar radiation curve but mainly visualises the spectral response of a thermopile pyranometer (source: pdf-document: the Instruction Manual of the CM21 precision pyranometer of Kipp & Zonen: http://www.kippzonen.com/ProductGroup/3/Pyranometers).

Although by connecting several thermocouples in series the sensitivity increases, still relatively low voltages are obtained. In case a sensitivity of $10\ \mu V/W/m^2$ is obtained, an irradiance of $500\ W/m^2$ implies a voltage of $5\ mV$. For instance using opamp circuits, this small voltage can be amplified.

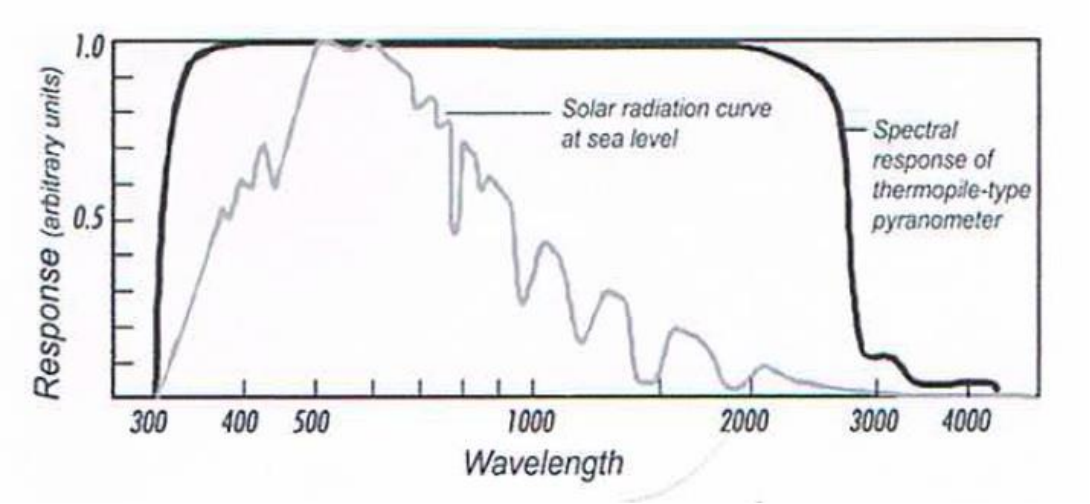


Fig. 4.2.5. Spectral response of a thermopile-type pyranometer

The use of appropriate materials

Thermopiles (and its thermocouples) have a better performance when a junction material is used which has a high thermo-electric coefficient (Seeback coefficient) relative to the other junction material. A higher Seeback coefficient implies a higher voltage is obtained due to the same temperature difference (inplying a higher sensitivity). Unfortunately, metals (e.g. gold, silver, copper) which have a low electrical resistivity (having a low electrical resistivity is also an advantage) in general also have a low Seeback coefficient. In a similar way, materials having a high Seeback coefficient (for instance antimony or bismuth) in general have a high electrical resistivity.

Nowadays, also silicon based thermopiles exist. These silicon based thermopiles are able to replace thermocouples based on e.g. antimony and bismuth. The Seeback coefficients of crystalline and polycrystalline silicon are really large and the electrical resistivity is relatively low. By changing the doping concentration, the Seeback coefficient and the electrical resistance can be adjusted.

4.2.4. The photodiode based pyranometer

Main working principle

A photodiode based pyranometer is a silicon pyranometer since the photodiode is made of silicon or germanium. Such a photodiode based pyranometer can cover a spectrum ranging from for instance $400\ nm$ to $900\ nm$ (although an appropriate choice of the semiconductor material allows to detect other parts of the spectrum).

Based on the photoelectric effect, a photodiode converts light into an electrical current. Figure 4.2.6 visualises the voltage current characteristics of a photodiode. A distinction can be made between the photodiode mode where the photodiode consumes power (left part in Figure 4.2.6) and the photovoltaic mode where the photodiode generates power (right part in Figure 4.2.6). As the solar irradiance increases, a higher voltage current characteristic, corresponding with higher currents, is obtained.

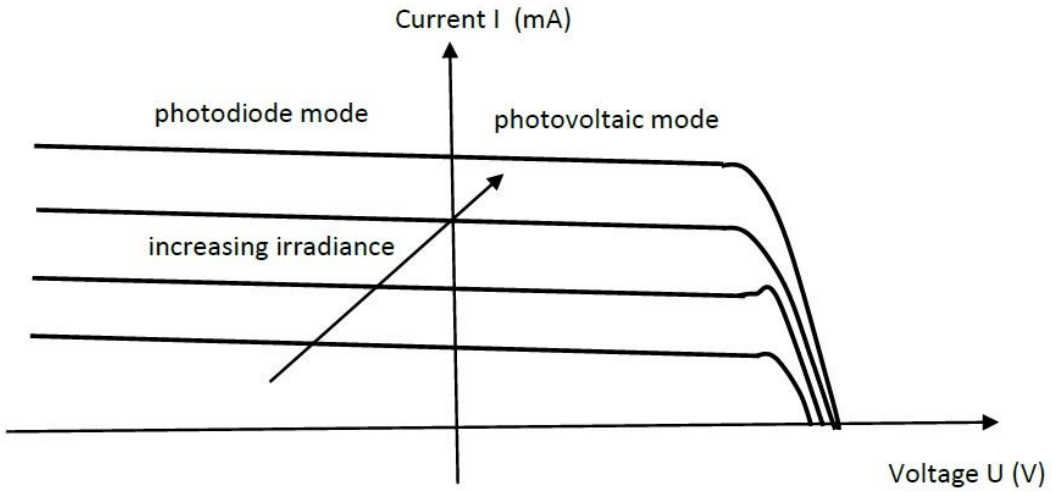


Fig. 4.2.6. Voltage current characteristics of a photodiode

When using the photodiode in photodiode mode, the circuit of Figure 4.2.7 is encountered. Notice the inverse polarization of the diode corresponding with a positive current I and a negative voltage U. Notice the photodiode consumes power as already mentioned. The behaviour of the circuit of Figure 4.2.7 is given by the equation (the load line)

$$U_{DC} = R.I - U\ .$$

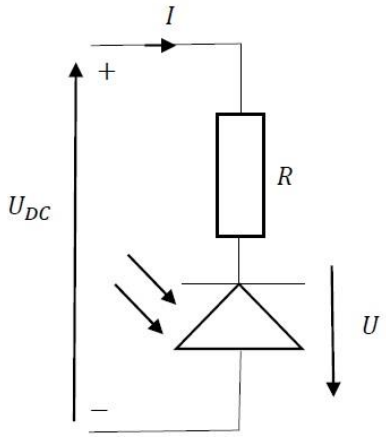


Fig. 4.2.7. The biased behaviour of a photodiode

The relationship $U_{DC} = R.I - U$ is visualised in Figure 4.2.8 corresponding with $U = -U_{DC}$ in case $I = 0$ and $I = U_{DC}/R$ in case $U = 0$. The actual voltage U and the actual current I depend on the irradiance of the sunlight and the associated voltage current characteristic. The operating point (giving U and I) is the intersection between the voltage current characteristic of the photodiode and the load line $U_{DC} = R.I - U$. As the irradiance increases, the current I increases (and $|U|$ decreases).

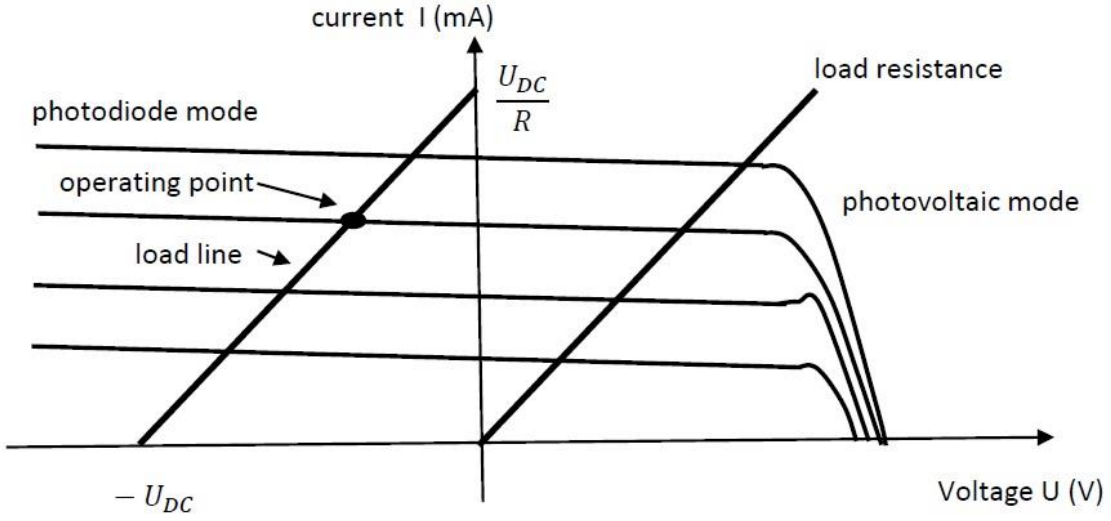


Fig. 4.2.8. Voltage current characteristics of a photodiode

In case the irradiance of the sunlicht increases, more photons are striking the photodiode which creates electron-hole pairs. To create an electron-hole pair, the photon needs sufficient energy i.e. its frequency must be sufficiently high. When using silicon photodiodes mainly wavelengths of $190\,nm$ to $1100\,nm$ are measured. When using germanium photodiodes mainly wavelengths of $400\,nm$ to $1700\,nm$ are measured. Germanium has a smaller bandgap than silicon implying the photon needs less energy which is related with lower frequenties and consequently higher wavelengths.

The holes are moving towards the anode and the electrons are moving towards the cathode implying a photocurrent which is a reverse current for the photodiode. This photocurrent increases as the irradiance increases.

Signal conditioning

Figure 4.2.9 considers the situation of the photodiode where a current I is flowing which is (approximately) proportional with the solar irradiance due to the

applied voltage U_{DC}. Using an operational amplifier, a current to voltage conversion is obtained. By an appropriate choice of resistor R_G, the sensitivity of the measurement (i.e. the amplitude of the output voltage) is determined.

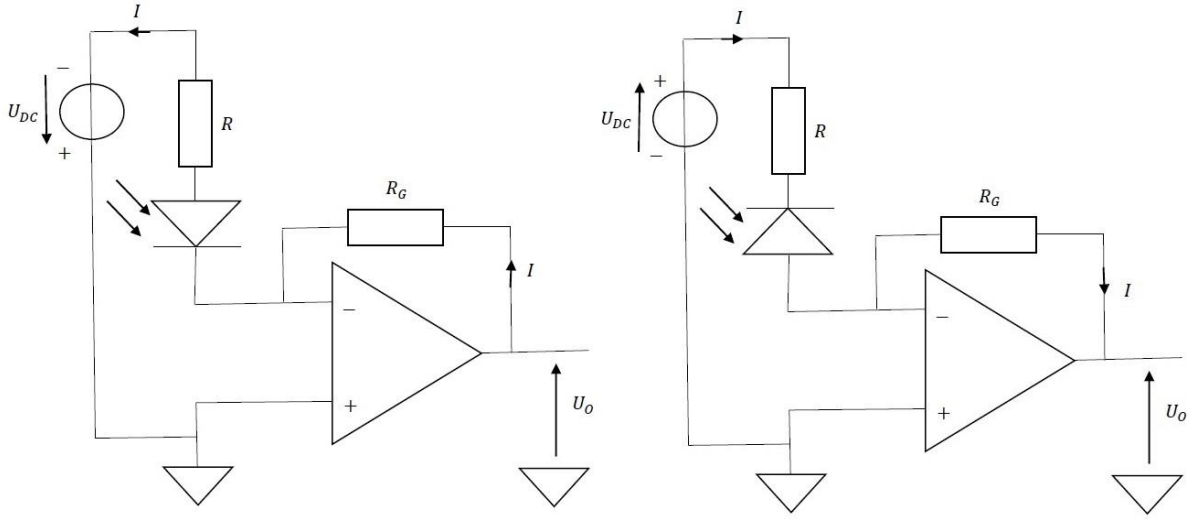


Fig. 4.2.9. Amplification of the photodiode information

Due to the negative feedback at the − input of the opamp, the voltage between the − input and the + input of the opamp is very small. Since the input impedance of the opamp is very large, almost no current is flowing at the − input of the opamp and the current I is flowing in the resistor R_G giving a voltage drop $R_G I$. Based on the voltage law of Kirchoff, this implies an output voltage $U_O = R_G I$ (left circuit) or $U_O = -R_G I$ (right circuit).

4.2.5. The photovoltaic pyranometer

When using a photodiode in a pyranometer in the photodiode mode, actually a bias is applied using the externally applied voltage U_{DC} as visualised in Figure 4.2.7. The photodiode is reverse biased since the cathode is connected with the positive node and the anode is connected with the negative node of the voltage source U_{DC}. Alternatively, a photodiode in a pyranometer can be used in the photovoltaic mode. In such a situation, no bias is aplied by using an external source.

A solar cell, which is actually a large area photodiode, is operating in a near short circuit condition i.e. it is loaded with a small resistor. The solar cell

approximately behaves as a current source and the generated current is (approximately) proportional with the solar irradiation (as visualised in Figure 4.2.1).

Signal conditioning

Figure 4.2.10 visualises the photodiode operating in photovoltaic mode implying a current I (approximately) proportional with the solar irradiance. Using an operational amplifier, a current to voltage conversion is obtained. By an appropriate choice of resistor R_G, the sensitivity of the measurement (i.e. the amplitude of the output voltage) is determined.

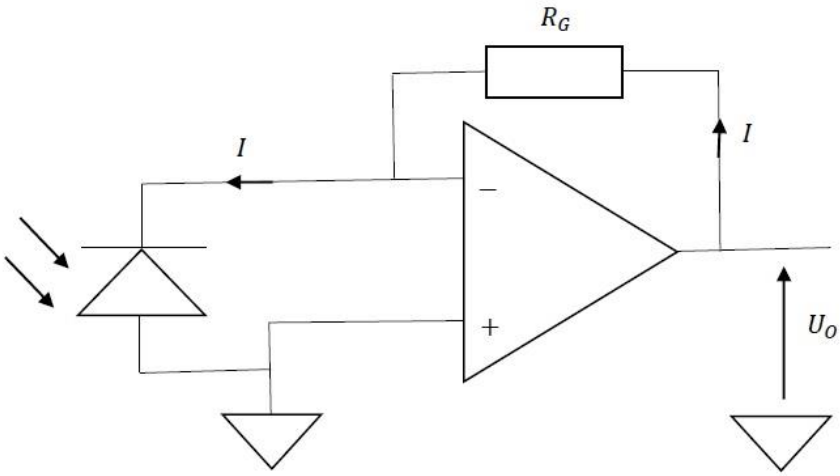


Fig. 4.2.10. Amplification of the solar cell information

Due to the negative feedback at the − input of the opamp, the voltage between the − input and the + input of the opamp is very small which short circuits the photodiode in photovoltaic mode (solar cell). Since the input impedance of the opamp is very large, almost no current is flowing at the − input of the opamp and the current I is flowing in the resistor R_G giving a voltage drop $R_G I$. Based on the voltage law of Kirchoff, this implies an output voltage $U_O = R_G I$.

4.3. Solar cells

4.3.1 Nature and spectral composition of solar light.

The energy of solar radiation originates from thermonuclear reactions of the proton-proton (at lower temperatures) and carbon-nitrogen (at higher temperatures) cycles responsible for the formation of a helium nucleus from four protons

$$4^1\mathrm{H} \rightarrow {}^4\mathrm{He} + 2\mathrm{e}^+ + 2\nu_\mathrm{e} + \Delta E,$$

where e^+ – positron, ν_e – electron neutrino. Every second about $6\cdot10^{11}$ kg ^{1}H are converted to ^{4}He. The mass defect $4\cdot1.008$ g (^{1}H) = 4.003 g (^{4}He) + 0.029 g comes to $4\cdot10^9$ kg resulting, in accordance with the Einstein relation, in the energy release ~ $3.8\cdot10^{26}$ J and we have

$$\Delta E = (4m\ 1_\mathrm{H} - m\ 4_\mathrm{He})\cdot c^2,$$

where c – speed of light.

The principal part of this energy is emitted in the form of electromagnetic radiation over the range from UV to IR. About 99 % of the solar radiation energy is associated with the wavelength interval 100 − 4000 nm (Fig. 4.3.1). At the present time the total mass of the Sun is approximately equal to $2\cdot10^{30}$ kg, that should be enough to provide its relatively stable existence, with a roughly constant energy release, for a period of about 10 billion years and more.

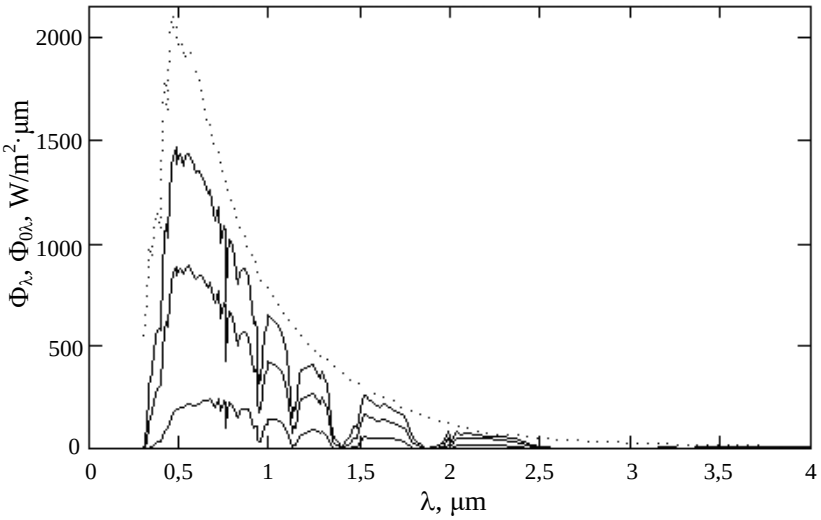


Fig. 4.3.1. Spectral distribution of the power of solar radiation Φ_λ incident on a horizontal area at astronomic noon in the central regions of Belarus for the winter solstice, vernal equinox, and summer solstice (solid curves from the bottom upwards, respectively) as compared with the initial solar spectrum $\Phi_{0\lambda}$ for the vernal equinox (dashed curve)

The total power Q of solar radiation over the whole wavelength range is $3.8 \cdot 10^{26}$ W. But in the environment the radiant energy of the Sun is dissipated inversely proportional to the squared object distance L

$$\Phi = \frac{Q}{4\pi L^2}.$$

The approximate distance between the Earth and the Sun is 149.5 mln. km. The mean radiant-energy density at the Earth orbit is 1370 W/m^2. This quantity is called the solar constant (Φ_0).

A solar spectrum is subdivided into three regions: UV (λ<390 nm) – 9 % of the total radiant energy; visible (390 nm <λ<760 nm) − 47 %; IR (λ>760 nm) − 44 %. Passing through the atmosphere, solar light is attenuated mainly due to absorption of IR radiation by water vapors, absorption of UV radiation by ozone, and scattering by the dust particles and aerosols contained in the air. The atmospheric effect on the intensity of solar radiation arriving to the Earth surface is determined by the atmospheric mass AM as follows:

$$AM = \frac{y}{y_0}\frac{1}{\sin\vartheta},$$

where y – atmospheric pressure, y_0 – normal atmospheric pressure (101,3 кПа), ϑ – angle of the Sun altitude over the horizon (Fig. 4.3.2).

A density of the luminous flux near the Earth surface is given as

$$\Phi_0 = \int_0^\infty \Phi_{0\lambda} e^{-\tau_\lambda m} d\lambda = \int_0^\infty \Phi_{0\lambda} e^{-\frac{\tau_\lambda h}{\sin\vartheta}} d\lambda = \int_0^\infty \Phi_{0\lambda} P^{\frac{1}{\sin\vartheta}} d\lambda,$$

where τ_λ – wavelength-dependent absorption factor in the atmosphere, m – distance covered by sunbeams in the atmosphere, h – height of the atmosphere,

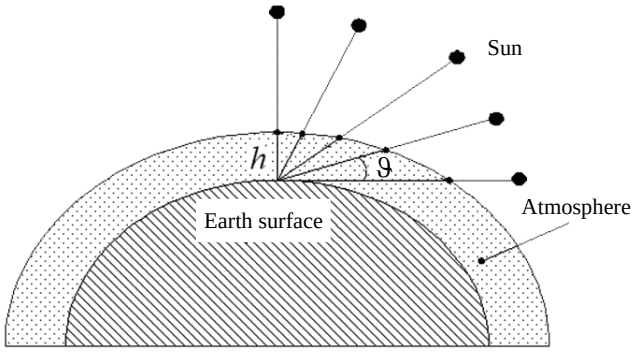


Fig. 4.3.2. Distance covered by solar beams in the atmosphere for different positions of the Sun over the horizon

$P = \Phi_{h\lambda}/\Phi_{0\lambda} = e^{-\tau_\lambda h}$ – transmission factor characterizing the atmospheric absorption.

For the mean latitudes, the flux of solar energy at the earth surface during the daytime is varied from sunrise (sunset) to the noon over the interval from 32.88 W/m^2 to 1233 W/m^2 in clear weather and from 19.2 μW/m^2 to 822 W/m^2 in cloudy weather.

As the spectral composition and density of the solar radiant flux is greatly dependent on the distance covered by sunbeams in the atmosphere, on the atmospheric density and composition (Fig. 4.3.1), there is a need for standardization of the measured parameters for solar cells (SC). According to the recommendation of the European Commission and of the International Electrotechnical Commission of the U.N. Organization, the atmospheric mass *AM* 1.5 with ϑ=41.81° (normal atmospheric pressure) has been adopted as a unified standard for measurements of SC parameters. The adopted solar radiation flux density was 835 W/m^2 that approximately corresponded to the mean radiant intensity on the Earth. Subsequently, it has been decided to conduct measurements of SC parameters taking a spectrum of the radiation associated with *AM* 1.5 and with the integrated radiation flux density 1000 W/m^2. *AM* 0 spectrum governs the operation of spacecraft SC. *AM* 1 spectrum is associated with solar radiation at the Earth surface when the Sun is at the highest (zenith) point, the total radiation power coming to ~ 925 W/m^2. *AM* 2 spectrum is realized at the angle ϑ=30° (normal atmospheric pressure). In this case the total radiation power is 691 W/m^2.

4.3.2 Light absorption in semiconductors.

When interacting with a semiconductor, optical radiation is partly absorbed, partly reflected from the surface, partly transmitted without absorption. Portions of the transmitted, reflected, absorbed energy are estimated by the corresponding factors. We distinguish the transmission factor $T = P_{\text{tr}}/P_{\text{inc}}$, reflection factor $R = P_{\text{ref}}/P_{\text{inc}}$, absorption factor $A = P_{\text{abs}}/P_{\text{inc}}$, where P_{tr} – power of transmitted radiation ; P_{ref} – power of the surface reflected radiation; P_{abs} – power of absorbed radiation; P_{inc} – power of incident radiation. The absorption factor α is equal to a value of the inverse scattering from the surface that is associated with attenuation of the initial power of incident radiation by a factor of *e*. At the depth *x* we have

$$P(x) = P_{\text{inc}} e^{-\alpha x}, \quad \alpha = -\frac{1}{x} \ln \frac{P(x)}{P_{\text{inc}}},$$

where $P(x)$ – power of radiation penetrating to the depth x. The absorption factor as a function of the incident radiation wavelength $\alpha(\lambda)$ is referred to as the absorption spectrum.

The operation of SC is based on the so-called intrinsic (or interband) radiation absorption in a semiconductor that corresponds to the photon energy spent on breaking of the valence band and on the electron transition from the valance to the conduction band (Fig. 4.3.3). This process results in the appearance of a free charge carrier (electron) in the conduction band and of a free charge carrier (hole) in the valence band – the intrinsic absorption of a photon by a semiconductor leads to the formation of an electron-hole pair of free carriers. For the transition of an electron to the conduction band, the energy of the absorbed photon should be higher than the forbidden bandwidth $E_{ph} = h\nu \geq E_g$, where E_{ph} – energy of the incident photon, E_g – forbidden bandwidth, h – Planck constant, ν – frequency of electromagnetic oscillations.

Because of this, a spectrum for intrinsic absorption in the long wavelength region has a marked threshold – the so-called photoemission (photoelectric) threshold, $\lambda_{\text{тр}} = ch/E_g$.

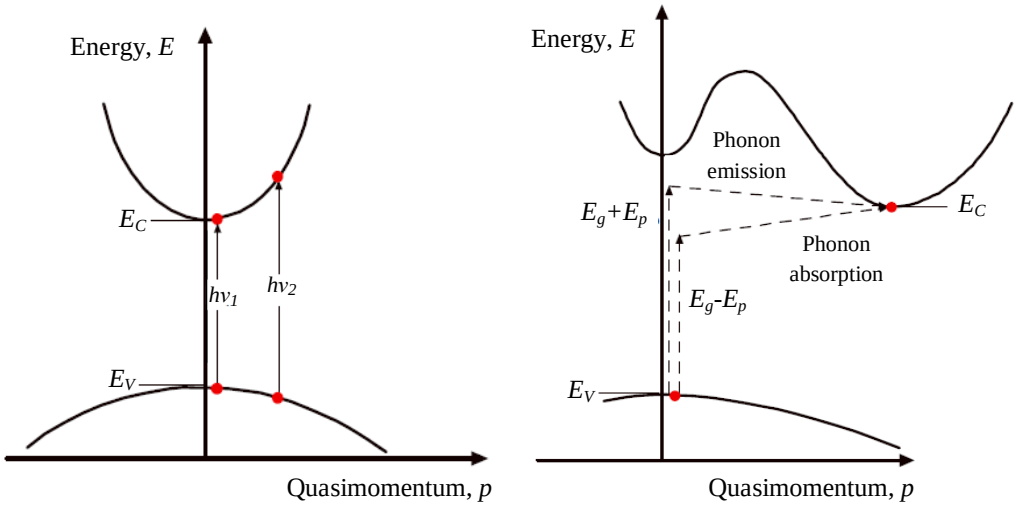


Fig. 4.3.3. Band structure of direct and indirect band semiconductors

In the case of intrinsic absorption the optical transitions may be both direct, without variations in the electron quasi-momentum (in direct band semiconductors), and indirect, with variations of the electron quasi-momentum

(in indirect band semiconductors). In direct band semiconductors the energy maximum of the valence band and the energy minimum of the conduction band are associated with one and the same value of the quasi-momentum. In indirect band semiconductors the energy maximum of the valence band and the energy minimum of the conduction band are associated with different values of the quasi-momentum (Fig. 4.3.3).

The probability of indirect optical transitions is significantly lower than the probability of the direct optical transitions. This is due to the fact that direct transitions are possible for the two-particle (electron–phonon) interaction, whereas indirect transitions necessitate involvement of the third quasi-particle, e.g., phonon. As a result, the optical radiation absorptivity of direct band semiconductors is, as a rule, higher than that of indirect band semiconductors by several orders of magnitude.

4.3.3 Photovoltaic effect in the p-n-junction.

SC makes it possible to convert the energy of optical radiation directly to the electric energy omitting the stages of thermal and mechanical energy. Operation of SC is based on the internal photoeffect in a semiconductor structure with the c *p*-*n*-junction (heterojunction, Schottky barrier).

Fig. 4.3.4 illustrates the principle of the *p*-*n*-junction formation. An electron (*n*-type) semiconductor contains some quantity of the donor-type impurity atoms,

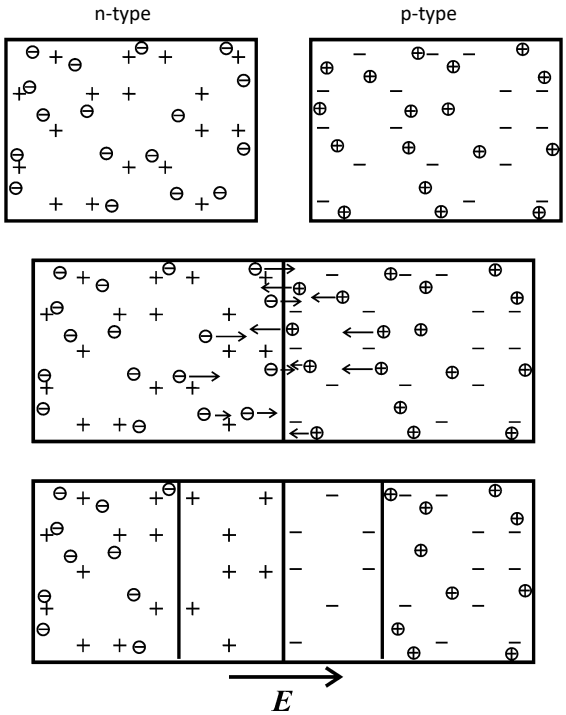


Fig 4.3.4. Formation of the p-n-junction

practically all of them being ionized at room temperature. That means the availability of n_0 free equilibrium electrons and the same number of immobile positively-charged ions. In a hole *p*-type semiconductor there are p_0 free equilibrium holes and the same quantity of immobile negatively-charged ions. In a *n*-type semiconductor electrons are the majority carriers and holes – minor carriers, whereas in a *p*-type semiconductor electrons are minor carriers and holes – major carriers.Contact of the *p* and *n* regions (Fig 4.3.4) leads, due to the density gradient of electrons and holes, to diffusion flux of the electrons from the *n*-type semiconductor to the *p*-type semiconductor and, *vice versa*, to the flux of the holes from the *p*-type semiconductor to the *n*-type semiconductor. The electrons, going from *n* region to *p* region, recombine with the holes close to the interface. Recombination of the holes from *p* region to *n* region is very similar. As a result, in the neighborhood of the *p*-*n* junction actually there are no free carriers: on both sides of *p*-*n* junction a *double charged layer* (*depletion layer* or *space-charge region* (SCR)) is formed by immobile ions. An electric field of SCR counteracts to diffusion of the majority carriers to the depletion region. Such a state is equilibrium and, in the absence of external disturbances, may be existent arbitrarily long.

Optical radiation absorbed in a semiconductor creates electron-hole pairs on condition that the quantum energy is in excess of the forbidden bandwidth. The process of separation involves the carriers generated in SCR of the *p*-*n*-junction and adjacent regions, with the sizes approximately equal to the diffusion length of minor carriers. Minor carriers generated in *p* and *n* regions at a greater distance from the junction than the diffusion length miss SCR due to recombination. Charge separation, in this case by the built-in electric field of the *p*-*n* junction, is defined as electromotive force (emf). In this way absorption of light by a semiconductor structure with the *p*-*n* junction results in origination of the photo-emf and, involving an external circuit, – to the current drawn in this circuit.

4.3.4. Equivalent circuit and current-voltage characteristic (CVC) of SC.

A typical SC is shown schematically in Fig 4.3.5. Most common type of SC is a semiconductor diode including two layers of different (electron - *n* and hole - *p*) conduction types and metal contacts to these layers. A front contact (to the illuminated surface) is made transparent or cellular, whereas a rear contact (to the

back surface) is solid. A special coating is applied to the front surface to protect a semiconductor layer against external effects, to lower reflection of incident radiation, and to decrease the surface recombination rate of the charge carriers.

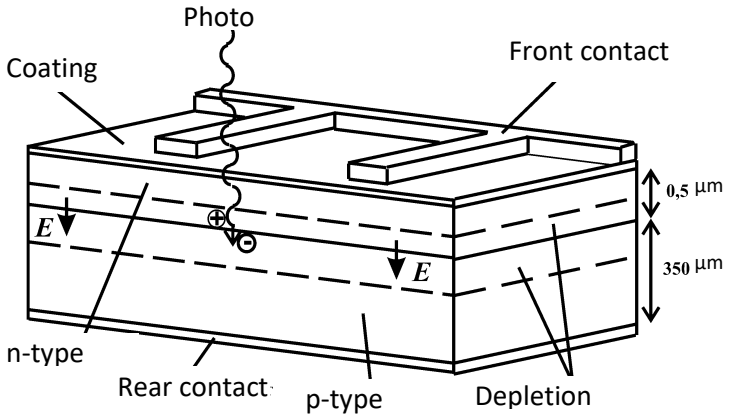


Fig 4.3.5. Schematic of SC made of polycrystalline silicon

Semiconductor material should be able to absorb the major part of solar radiation spectrum. To make their contribution to the photocurrent, minor carriers generated by radiation should arrive at SCR prior to their recombination (i.e. , during their lifetime). The junction depth from the surface should correlate with the radiation penetration depth of a particular semiconductor. Even such a weakly absorbing semiconductor as silicon has an optimum depth of the *p*-*n* junction about 0.5 μm. Because of this, the design and materials of SC should meet the specified requirements. To illustrate, a thin *n*-type layer has a relatively high resistance and hence, to lower the loss for electrical resistance, the photoactive material in this layer should have a high conductivity and the front contact network should be well engineered.

On illumination of SC that is based on the *p*-*n* junction, it carries the drift current of nonequilibrium minority carriers. In turn, the equilibrium majority carriers fail to get over the potential barrier and remain in the region of generation. Due to separation of the optically generated carriers, the density of holes in the *p*-region and of electrons in the *n*-region is increased to compensate for the volume charge of immobile impurity ions of SCR. The junction potential barrier is lowering by the value of photo-emf (open-circuit voltage). Because the potential barrier is lowered, the diffusion current of majority carriers is drawn through the barrier in a counter-direction to the photocurrent. In the stationary state the diffusion current density J_{dif} equals the drift current density including the

photocurrent density J_{ph} and the junction thermal- current density J_0, i.e., the following condition of dynamic equilibrium is met: $J_{dif} = J_{ph} + J_0$.

The difference $J_{dif} - J_0$ represents the diode current density and is denoted as J_d. In the case of an ideal *p-n* junction the diffusion current density and the thermal current density are related as

$$J_{dif} = J_0 e^{U_{OC}/V_T}.$$

Proceeding from the last equation, J_0 is referred to as the diode saturation current density. We have

$$J_{ph} = J_d = J_0(e^{U_{OC}/V_T} - 1),$$

where U_{OC} – open-circuit voltage; $V_T = kT/e$ – thermal potential. The open-circuit voltage may be expressed in terms of the photocurrent as

$$U_{OC} = V_T \ln(1 + J_{ph}/J_0).$$

However, U_{OC} (for any J_{ph}) is equal to or below the contact potential difference of the *p-n* junction because separation of the carriers stops at complete compensation of the electric field.

When the electrodes of SC are terminated in the external load, the voltage between them U is lower than U_{OC} and the diode current will not compensate for the photocurrent. In this case the current is drawn through the external load, with a density in the ideal diode approximation, and we have

$$J = J_{ph} - J_d = J_{ph} - J_0(\exp(eU/kT) - 1).$$

The last equation describes the current-voltage characteristic (CVC) of an ideal SC. For the ideal SC, the short-circuit current $J_{SC} = J_{ph}$.

An equivalent circuit of the ideal SC represents a current generator and an ideal diode connected in parallel (Fig 4.3.6).

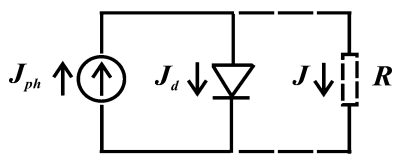


Fig 4.3.6. Equivalent circuit of the ideal SC: R – load resistance

In derivation of CVC for SC an equation of the ideal diode has been used and this is not always justified. The characteristic is often transformed, by introduction into the denominator of the exponent for the factor *A*, taking account of the recombination processes in SCR

$$J = J_{ph} - J_0(\exp(eU/AkT) - 1).$$

But even this expression conforms with the experiment insufficiently. As an area of SC is large, we should take into account the shunt resistance R_P (leakage resistance) too. High values of the current drawn through SC necessitate the inclusion of the series resistance R_S. Accordingly, the equivalent circuit of SC is modified as shown in Fig 4.3.7.

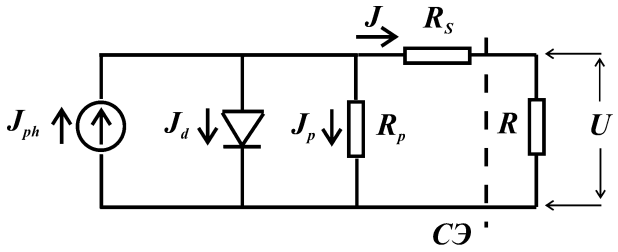


Fig.4.3.7. Equivalent circuit of SC

The current generator simulates the current J_{ph} initiated on illumination, the diode parallel with it allows for the injection (J_{dif} and J_0). The series resistance R_S includes the resistance of contact layers, resistance of each of the *p*- and *n*-regions in SC, metal-semiconductor junction resistance; the shunt resistance R_P reflects all possible channels of the current leakage parallel to the *p*-*n* junction.

In accordance with Fig 4.3.7, we derive an equation to describe, quite adequately, CVC of SC as follows:

$$J = J_{ph} - J_0\left(exp\left(\frac{e\left(U + JR_S\right)}{AkT}\right) - 1\right) - \frac{JR_S + U}{R_P}.$$

Fig 4.3.8 demonstrates the standard way to represent CVC of SC given by the last equation.

The open-circuit voltage (U_{OC}) – peak voltage arising at open terminals of SC when it is irradiated by solar light. The short-circuit current (J_{SC}) – peak current flowing through the terminals of SC in the case of short-circuit.

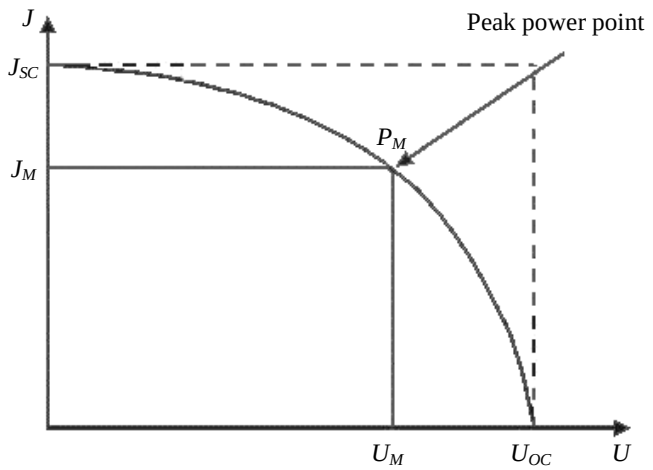


Fig 4.3.8. Current-voltage characteristic of SC

4.3.5. Spectral sensitivity of SC.

SC converts the energy of optical radiation with a particular spectral composition – spectral composition of solar radiation – to the electric energy. Because of this, spectral sensitivity of SC is a very important characteristic. Spectral sensitivity of SC is understood as a relationship between J_{SC} (J_{ph}, U_{OC}) and the wavelength of incident monochromatic radiation, normalized per unit energy of incident radiation with the same wavelength.

The reason for spatial selectivity of SC is that optical radiation with different wavelengths penetrates semiconductors to different depths (Fig 4.3.9) to create its distribution of the light-generated electron-hole pairs. The penetration depth is determined by the incident radiation wavelength and by the absorption factor of a semiconductor for the specific wavelength.

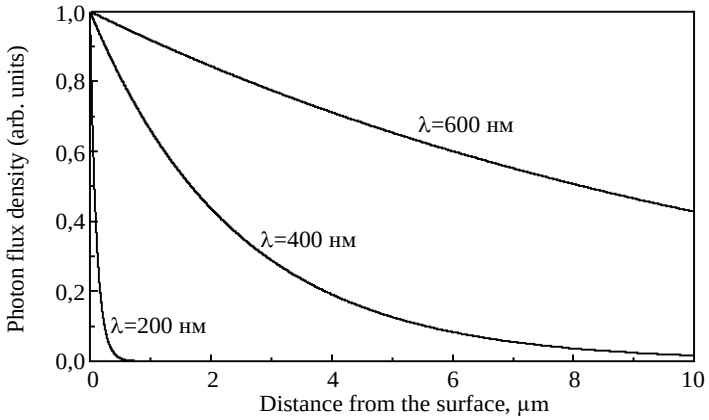


Fig 4.3.9. Penetration of incident radiation with different wavelengths into silicon

The ratio between the electron-hole pairs separated by the SCR field to the total number of the radiation-generated electron-hole pairs (quantum efficiency of SC) is determined by the SCR position (Fig 4.3.10) and by the radiation penetration depth at different wavelengths.

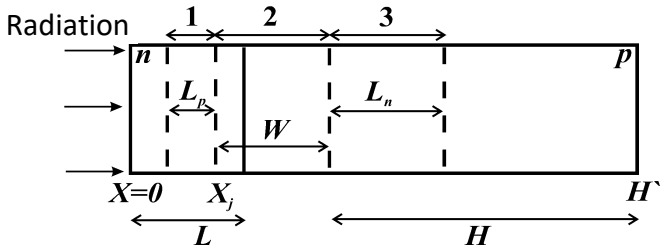


Fig 4.3.10. One-dimensional model of elementary SC: W – width of SCR; L_n – diffusion length of the electrons in p-region; L_p – diffusion length of the holes in n-region; x_j – boundary of SCR in n-region; L – junction depth; H – thickness of a quasi-neutral region of p-base; H' – total thickness:

We can distinguish in SC three regions associated with the photocurrent generation (Fig 4.3.10). Then we have

$$J_{ph} = J_p + J_n + J_{SCR},$$

where J_p – current density of the holes generated in region 1and arriving at SCR; J_n – current density of the electrons generated in region 3and arriving at SCR; J_{SCR} – current density for the carriers generated within region 2, i.e., in SCR.

Assuming that the doping profile corresponds to the abrupt *p-n* junction, we can derive expressions for J_p, J_n , and J_{SCR}. We get

$$J_{SCR} = eF(1-r)\exp(-\alpha x_j)\left[1-\exp(-\alpha W)\right],$$

where $F = F(\lambda)$ – incident photon flux density in a unit spectral interval; $r = r(\lambda)$ – portion of the photons reflected from the surface in a unit spectral interval; $\alpha = \alpha(\lambda)$ – absorption factor.

Let us derive an expression for J_n. At low excitation level, allowing for computation of the recombination rate of nonequilibrium carriers $\propto(n_p\text{-}n_{p0})$ in the *p*-type semiconductor, a one-dimensional steady-state continuity equation takes the following form: $g_n - \frac{D_n}{L_n^2}\left(n_p - n_{p0}\right) + \frac{1}{e}\frac{dJ_n}{dx} = 0$ for the electrons (*p*-type), where

n – density of free electrons; D_n – diffusion factor of the electrons; g_n – electron generation rate per unit area of the irradiated surface.

The electron current density

$$J_n = e\mu_n n_p E + eD_n \frac{dn_p}{dx},$$

where E – strength of an electrostatic field. We substitute the last expression into the continuity equation assuming that $E = 0$ (an abrupt p-n junction)

$$D_n \frac{d^2 n_p}{dx^2} - \frac{D_n\left(n_p - n_{p0}\right)}{L_n^2} + \alpha F\left(1-r\right)\exp\left(-\alpha x\right) = 0$$

$$J_n = eF\left(1-r\right)D_n\left(\frac{dn_p}{dx}\right)_{x_j+W}.$$

At the boundary conditions

$$n_p = n_{p0}\Big|_{x=x_j+W}, \qquad S_n\left(n_p - n_{p0}\right) = -D_n \frac{dn_p}{dx}\Big|_{x=H'},$$

where S_n – surface recombination rate of the electrons, we obtain

$$J_n = e\frac{F(1-r)\alpha L_n}{\alpha^2 L_n^2 - 1} e^{-\alpha(x_j+W)} \times \left[\alpha L_n - \frac{\left(\frac{S_n}{D_n}L_n\right)\left(\operatorname{ch}\frac{H'}{L_n} - e^{-\alpha H'}\right) + \operatorname{sh}\frac{H'}{L_n} + \alpha L_n e^{-\alpha H'}}{\left(\frac{S_n}{D_n}L_n\right)\operatorname{sh}\frac{H'}{L_n} + \operatorname{ch}\frac{H'}{L_n}}\right].$$

In a similar way, we have

$$J_p = \frac{eF(1-r)\alpha L_p}{1-\alpha^2 L_p^2}\left[\alpha L_p e^{-\alpha x_j} - \frac{\left(\frac{S_p}{D_p}L_p + \alpha L_p\right) - e^{-\alpha x_j}\left(\frac{S_p}{D_p}L_p \operatorname{ch}\frac{x_j}{L_p} + \operatorname{sh}\frac{x_j}{L_p}\right)}{\left(\frac{S_p}{D_p}L_p\right)\operatorname{sh}\frac{x_j}{L_p} + \operatorname{ch}\frac{x_j}{L_p}}\right],$$

where D_p и S_p – diffusion factor and the surface recombination rate of the holes.

The combined expressions derived for J_{SCR}, J_n, and J_p describe the spectral characteristic of SC sufficiently well. As follows from these expressions, the photocurrent is determined by the spectral composition of incident radiation, absorbance of a semiconductor, spatial position of SCR, diffusion and recombination parameters.

Efficiency and its limiting value.

The efficiency (coefficient of performance) – most important characteristic of SC – gives the efficiency of the solar energy conversion to the electric energy as

$$\eta = \frac{P_M}{P} = \frac{ff \cdot J_{SC} U_{OC}}{P},$$

where P – power of radiation incident on SC per unit surface area; P_M – peak output power of SC with respect to its surface area; ff – fill factor or CVC-shape factor. We get

$$ff = \frac{J_M U_M}{J_{SC} U_{OC}},$$

where J_M and U_M – current density and voltage associated with the peak power point P_M (Fig 4.3.8). The efficiency of SC indicates which part of the incident light energy it is able to convert to electricity.

Inability to use photons with the energy below the forbidden-band width of a semiconductor and the excessive energy of short-wavelength photons is greatly responsible for the energy loss in SC. The efficiency of SC is dependent on its reflectivity, spectral composition and intensity of incident radiation, on temperature, and some other parameters of SC. Shockley and Quesser were the first to apply the principle of detailed equilibrium between light absorption and radiation recombination for description of the photogeneration processes of SC. This sets the limit for a theoretically achievable efficiency of the ideal semiconductor SC.

Shockley and Quesser have developed the formalism for calculations of the limit efficiency of SC based on the *p-n* junction. They have supposed that generation of the carriers in SC is determined by the detailed equilibrium between light absorption and radiative recombination.

Mobility of the carriers in the ideal case is thought to be equal to infinity, whereas the Fermi level position is quasi-invariant. Besides, the surface recombination rate is zero. In the equilibrium state the Fermi energy quasi-levels (E_{FV} for the p-type and E_{FC} for the n-type semiconductors) are identical. Also, it is assumed that the electric contact between semiconductors of the two types is ideal. The voltage V initiated between the two electrodes is determined by the difference in Fermi quasi-levels of the majority carriers and may be derived from the following equation:

$$eV = E_{FC} - E_{FV}\,.$$

The total power of solar radiation Q is found from the Stefan-Boltzmann law as follows: $Q = \sigma T_s^{\,4}$, where σ – Stefan-Boltzmann constant, T_s=*5800 K* – surface temperature of the Sun.

When light is absorbed in a semiconductor, the electrons are activated to go from the valence band to the conduction band and the generation of electron-hole pairs takes place. In accordance with the detailed equilibrium principle, the inverse process also occurs and the electrons go from the conduction band to the valence band with photon emission. Such a process is known as radiative recombination. In the ideal case it is assumed that there is no radiative recombination.

The difference between the absorbed and emitted photons determines the photocurrent density limit J_{ph} for SC

$$J_{ph} = e(\Phi_s - \Phi_r)\,,$$

where Φ_s and Φ_r – densities of photon fluxes incident on the surface and emitted by the surface of SC, respectively. The photon flux is found from the Planck equation as

$$\Phi_s = \xi \frac{2\pi \cdot \sin^2 \vartheta_s}{h^3 \cdot c^2} \int_{Eg}^{\infty} \frac{\varepsilon^2 \mathrm{d}\varepsilon}{\exp\frac{\varepsilon}{k \cdot T_s} - 1}, \qquad \Phi_r = \xi \frac{2\pi}{h^3 \cdot c^2} \int_{Eg}^{\infty} \frac{\varepsilon^2 \mathrm{d}\varepsilon}{\exp\frac{\varepsilon - e \cdot V}{k \cdot T_a} - 1},$$

where ξ – absorptivity or emissivity (according to the Kirchhoff law, we assume that the absorptivity equals the emissivity); ϑ_s - solid angle at which the Sun disk is visible at the horizon; T_a – temperature of SC; k – Boltzmann constant.

Shockley and Quesser have supposed that all the incident photons with the energy in excess of the forbidden-band width in a semiconductor are absorbed, i.e., for them $\xi = 1$.

If the CVC fill factor of SC tends to unity, the efficiency of SC is given by $\eta = \frac{eV(\Phi_s - \Phi_r)}{\sigma T_s^4}$.

As distinct form Shockley and Quesser, we take into account that $\xi < 1$. The absorptivity ξ may be found with the help of the following formula according to the Bouguer-Lambert-Beer law:

$$\xi = \frac{I_0 - I_0 \exp(-\alpha d)}{I_0} = 1 - \exp(-\alpha d),$$

where I_0 – incident radiation intensity; α – absorption factor; d – thickness of an absorptive layer.

The absorption factor for the allowed direct dipole transitions is given by the following equation:

$$\alpha(h\nu) = \frac{\pi e^2 (2m_r)^{3/2} E_g}{n\varepsilon_0 m_e h^3 c\nu} \left(h\nu - E_g\right)^{1/2},$$

where n – refractive index of a semiconductor material; m_e – effective electron mass; m_r – reduced effective mass; ε_0 – electric constant.

For the majority of direct band semiconductors, the absorption factor may be calculated from the following expression:

$$\alpha(h\nu) = A \cdot E_g \frac{\left(h\nu - E_g\right)^{1/2}}{h\nu}.$$

The constant A is determined by the properties of a specific material, its value varying over the range about $[10^4; 10^5]$ $cm^{-1}(eV)^{-1/2}$.

With the use of the last equation, the limit efficiency may be calculated more accurately.

Let us consider radiation absorption in SC for two cases (total concentration and normal concentration of solar radiation) with due regard for a solid angle at which the solar disk is visible.

For the total concentration of solar radiation, we have $\vartheta_s = 4\pi\ sr$; for the case of normal concentration of solar radiation, we have $\vartheta_s = 6.65 \cdot 10^{-5}\ sr$.

Fig 4.3.11 demonstrates the calculation results for the limit efficiency of SC in the case ξ = 1, corresponding to the Shockley-Quesser model.

At T_a=0 K (curve 1 in Fig. 4.3.11) we obtain the well-known dependence of the limit efficiency of SC on the forbidden-band width of a photoactive semiconductor material that is in line with the Shockley-Quesser model. The curves for the cases of the total and normal concentrations of incident solar radiation are coincident at T_a=0 K because there is no term describing the radiative loss. As seen in Fig 4.3.11, the efficiency of SC at room temperature is decreased due to the nonzero term describing the radiative loss.

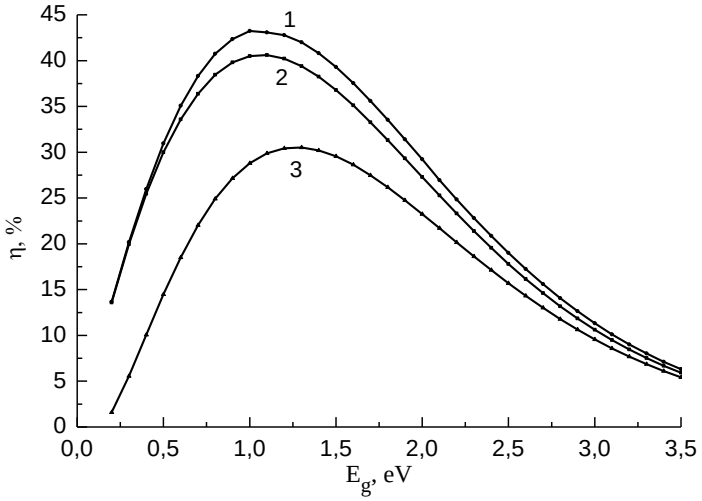


Fig 4.3.11. The limit efficiency of SC as a function of the forbidden-band width of a photoactive semiconductor material at ξ = 1: 1 – $T_a = 0$; 2 – $T_a = 300\ K$, $\vartheta_s = 4\pi\ sr$; 3 – $T_a = 300\ K, \vartheta_s = 6.65 \cdot 10^{-5}\ sr$.

In Fig 4.3.11 it is seen that at 0 K a maximum efficiency of SC is about 43.5% over the range of the semiconductor forbidden-band widths 1.0-1.1 eV, being independent of the SC illumination conditions. At a temperature of about 300 K, in the conditions of the total solar radiation concentration, a maximum efficiency corresponds to the range of the semiconductor forbidden-band widths 1.0-1.1 eV but its value is lower, approximately coming to 41%. For normal temperature, we observe a significant influence of the radiation concentration level on the dependence of the limit efficiency of SC on the forbidden band of semiconductor materials: in the case of the normal radiation concentration a maximum of the efficiency is shifted to the range 1.2 – 1.3 eV and equals 31% or so.

When the absorptivity $\xi < 1$, the limit efficiency of SC is dependent on the photoactive layer thickness and on the constant *A* from the expression for the absorption factor.

Figs. 4.3.12 and 4.3.13 demonstrate the calculation results for the limit efficiency of SC at different values of the photoactive layer thickness in the cases of the total and normal concentrations of solar radiation, respectively. It was assumed that a value of the constant *A* is equal to $10^5 cm^{-1} (eV)^{-1/2}$. It is obvious that, if for some semiconductor materials the constant *A* is by several factors lower than $10^5 cm^{-1} (eV)^{-1/2}$, the derived functions are associated with the photoactive layer thickness that is by several factors higher.

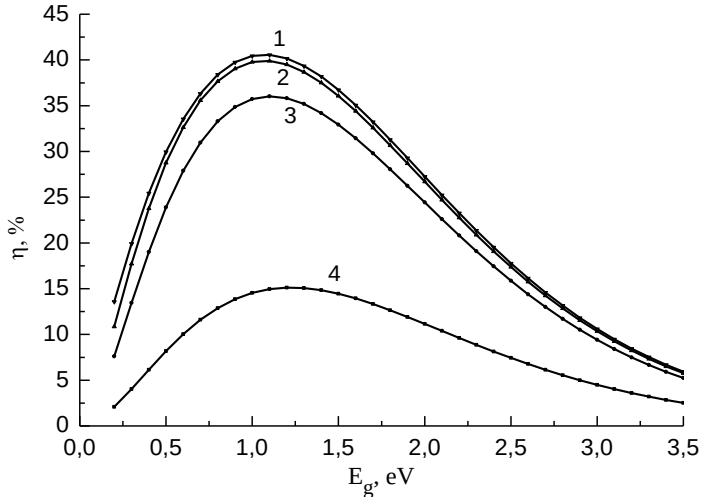


Fig 4.3.12. The limit efficiency of SC as a function of the forbidden-band width of a photoactive semiconductor material in the case of the total solar radiation concentration for different thicknesses of the photoactive semiconductor layer:
1 – d = 100 μm; 2 – d = 1 μm; 3 – d = 0.5 μm; 4 – d = 0.1 μm.

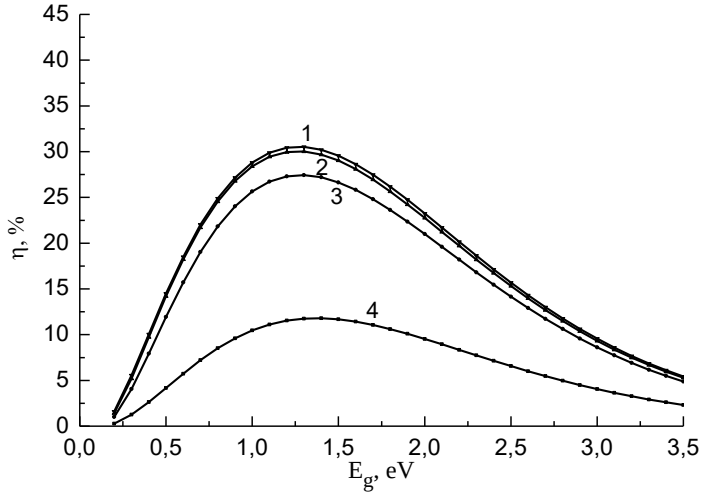


Fig 4.3.13. The limit efficiency of SC as a function of the forbidden-band width of a semiconductor material in the case of normal solar radiation concentration for different

thicknesses of the photoactive semiconductor layer: 1 – d = 100 μm; 2 – d = 1 μm; 3 – d = 0.5 μm; 4 – d = 0.1 μm.

Analysis of Figs. 4.3.12 and 4.3.13 indicates that the limit efficiency of SC is growing with the photoactive semiconductor-layer thickness. When the constant *A* from the equation for the absorption factor equals $10^5 \text{см}^{-1}(\text{eV})^{-1/2}$ and a thickness of the photoactive semiconductor layer is over 1 μm, the derived functions tend to those associated with the Shockley-Quesser model (Fig 4.3.11).

Obviously, it is inexpedient to increase a thickness of the photoactive layer of SC beyond 1 μm. Indeed, in the case of both total and normal concentrations of solar radiation the photoactive layer thickness of SC greatly influences the dependence of the limit efficiency of SC on the forbidden-band width that is below 1μm. As the photoactive layer thickness is lowering, a maximum of the SC efficiency is shifted to the region of high E_g. At the same time, an increase of the photoactive layer thickness beyond 1 μm brings no significant changes. Note that these conclusions are valid for materials with the constant *A* that is equal to $10^5 \text{см}^{-1}(\text{eV})^{-1/2}$, e.g., for GaN. But, for InSb with $A = 2 \cdot 10^4 \text{см}^{-1}(\text{eV})^{-1/2}$, the appropriate thickness of the photoactive layer comes to about 5 μm.

4.3.6. Requirements to photoactive materials for production of SC.

Fig 4.3.14 presents the curves for the theoretically achievable efficiency of a homogeneous SC as a function of the forbidden-band width of a photoactive material and the data about the maximum practical efficiency.

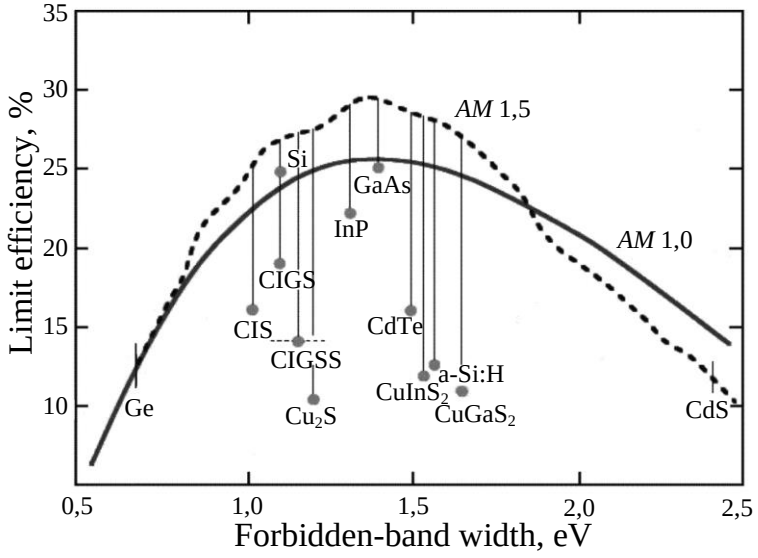


Fig 4.3.14. Maximal efficiency of SC as a function of the forbidden-band width of a semiconductor material (25 ºC): CIS – $CuInSe_2$, CIGS – $Cu(In,Ga)Se_2$, CIGSS – $Cu(In,Ga)(S,Se)_2$

On the one hand, a low value of the forbidden-band width E_g makes it possible to use effectively the greatest part of the solar spectrum and hence to increase J_{ph} improving the efficiency. On the other hand, lowering of E_g directly leads to lowering of U_{OC} that also affects the efficiency. A great value of E_g allows for the creation of SC with a high open-circuit voltage, while such a cell will have a low value of the photocurrent as photons with the energy below E_g are not absorbed in a semiconductor. Because of this, the curve for the theoretically achievable efficiency as a function of E_g has a maximum.

As seen, the efficiency of silicon SC (nearly 90 % of all SC produced) is below the theoretical limit, though it is approaching the limit, the value of E_g for silicon being close to the optimum one. High efficiency of silicon SCs, which are relatively inexpensive, has been achieved due to the advances in silicon microelectronic industry. However, we can hardly expect that the characteristics of silicon SC would be improved considerably or their prices would be reduced. This is associated not with E_g of silicon but with another important factor: absorptivity of silicon is not high because silicon is an indirect-band semiconductor. The fact that the absorptivity of indirect-band semiconductors is considerably lower than the absorptivity of direct-band semiconductors is of particular importance for manufacturers of SC from the viewpoint of the materials-output ratio. To illustrate, 90 % of optical radiation is absorbed in 1 μm of GaAs (direct-band semiconductor), whereas in silicon the same portion of optical radiation is absorbed in 100 μm. In this case, to activate the photogenerated carriers to SCR of the p-n junction, a diffusion length of the minority carriers should be rather great. This means that the material must be very pure and its crystalline structure must be perfect.

Proceeding from the above-mentioned physical limitations, it may be inferred that, despite a dominant role of silicon, direct-band semiconductors offer much promise for photovoltaics in the future.

Among direct-band semiconductors, of particular interest are the so-called solid solutions, e.g., $Cu(In,Ga)(S,Se)_2$, whose electric and optical parameters (E_g including) depend on a relative concentration of the components in a sold solution. Varying a relative concentration of In/Ga (ratio of the interchangeable elements in the cationic sublattice) and of S/Se (ratio of the interchangeable elements in the anionic sublattice), we can vary E_g of the solid solutions $Cu(In,Ga)(S,Se)_2$ over the range 1.0 to 2.4 eV, correlating it with the optimum value 1.2 – 1.6 eV.

A search for new materials needed to produce SC is still under way. The principal criteria for ideal photovoltaic materials may be summarized as follows:

- direct-band semiconductor;
- forbidden-band width over the range 1.2 – 1.6 eV;
- nontoxic and available for large-scale production;
- simplicity and reproducibility of the material deposition technique on large areas;
- high efficiency of photoelectroconversion;
- long-term stability of physical properties.

Unfortunately, no materials capable to meet all these requirements have been found by now. As a high light absorption factor is most important, numerous studies have been devoted recently to the so-called thin-film materials: to produce SC, it is sufficient to use a layer of the photoactive material with a thickness on the order of 1 μm. The ratio materials-output may be lowered critically as compared to the use of crystalline silicon.

Most likely, the three following scenarios in the development of photovoltaic materials are possible in the nearest future:

- further domination of mono- or polycrystalline silicon;
- development of the technologies for the formation of medium-thickness silicon films such as ribbon silicon or silicon on substrates of other materials;
- industrial production of thin-film materials such as α-Si, CIGSS or CdTe.

The long-term prospects are also associated with new design conceptions or with new innovative materials for SC, e.g., tandem SC or organic SC. At the present time we can hardly predict the most probable scenario. Most likely, several types of SC would be available for a long period, each of them finding its own niche at the market.

In general, one of the advantages of photovoltaics is diversity of the approaches liable to result in the development of highly-efficient, stable, and inexpensive solar cells.

4.4. Application of photovoltaic systems

When considering photovoltaic installations used to convert solar power into electrical power, a distinction can be made between smaller installations mounted on the roofs of private households and large scale solar parks. Figure 4.4.1 visualises a private household having photovoltaic panels mounted on the roof (very often oriented to the South side). Figure 4.4.2 visualizes a large solar park having a maximum generated power of 1.3 MW.

First, we will discuss a typical small scale installation containing photovoltaic panels as it is often owned by private households in Western Europe.

Fig.4.4.1 Photovoltaic panels mounted on the roof of a private household

Fig. 4.4.2. Large scale solar park

4.4.1 Photovoltaic cells, panels and strings

A photovoltaic panel (or solar panel or briefly panel) consists of a number of photovoltaic cells (or solar cells or briefly cells) connected in series. The number of photovoltaic cells depends on the desired DC voltage. When having sufficient solar irradiance, it is realistic to assume a generated voltage level of approximately $0.45\,V$ for each photovoltaic cell. In such a situation, a photovoltaic panel containing 36 photovoltaic cells generates a voltage of $16.2\,V$ which is sufficiently high to load a $12\,V$ battery without any problem. A situation where only one photovoltaic panel is used to load a battery can, for instance, be used to supply street lights or parking meters.

When considering private households, the generated power is generally not used to load a battery. Using a power electronic converter, the generated power is injected into the public AC grid. The power electronic converter converts the DC voltage of the photovoltaic panels into a $50\,Hz$ AC voltage as visualised in Fig. 4.4.3.

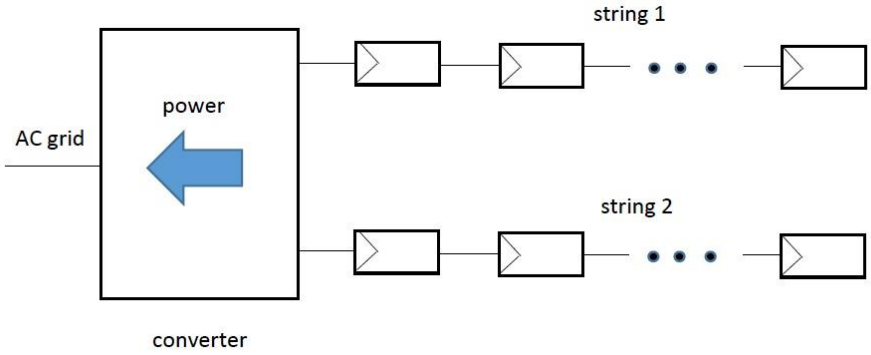


Fig. 4.4.3. Photovoltaic installation

Consider the situation where a single panel contains 6x12 cells. Such a panel generates a DC voltage of $32.4\,V$ in case individual cells generate $0.45\,V$. In general, a number of panels are connected in series in order to obtain a larger voltage level. The number of series connected panels depends on the panel type (i.e. the number of cells) and the converter type (including the allowed DC voltage range at the input of the converter). A series connection of panels is called a photovoltaic string (or solar string or briefly string). In general, the capital investment, the roof area and the rated power of the converter are important parameters when determining the number of strings which will be connected.

It is a good practice to spread the total number of photovoltaic panels over the strings in order to have identical strings. For instance, this implies an installation containing 17 panels is seldomly used. Indeed, when spreading 17 panels over two strings, a string of 8 and a string of 9 panels are obtained. By connecting two different strings in parallel, a decrease of the overall efficiency of the installation occurs.

Fig. 4.4.3 visualises a photovoltaic installation containing two strings which inject their power in the AC grid using a power electronic converter. Such an installation is often owned by private households in Western Europe.

4.4.2. The voltage current characteristic of a photovoltaic panel

Measuring the voltage current characteristic

In general, a photovoltaic panel contains a number of photovoltaic cells connected in series. Fig. 4.4.4 visualises typical voltage current characteristics of a photovoltaic panel (at a constant temperature of $25°C$) with different solar irradiances. For instance an incident solar power of $500 - 1000\ W/m^2$ corresponds with clear and sunny weather, an incident power of $120 - 500\ W/m^2$ corresponds with sunny partially clouded weather, an incident power of $50 - 120\ W/m^2$ corresponds with really clouded weather.

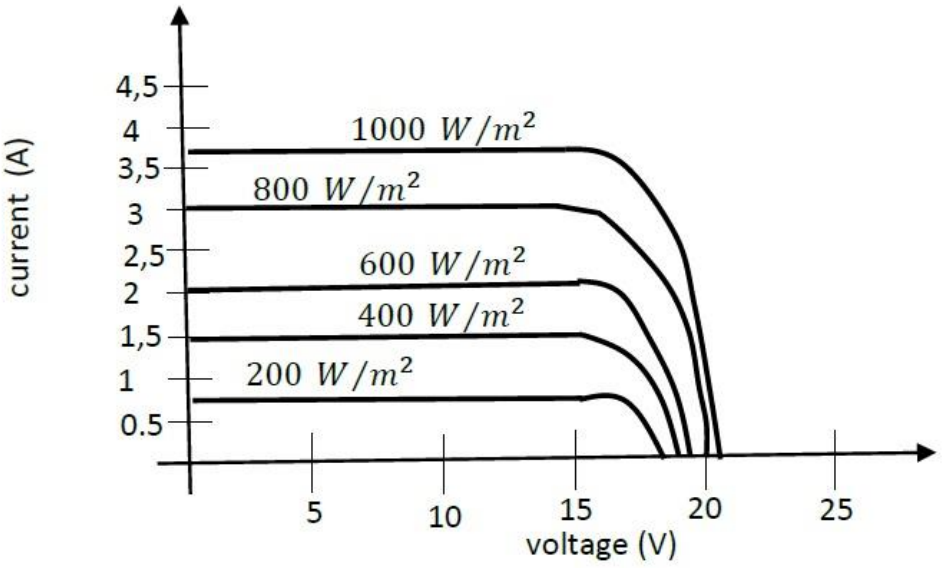


Fig. 4.4.4: Voltage current characteristics of a photovoltaic panel

As Fig. 4.4.4 visualises, the voltage current characteristic strongly depends on the solar irradiance. The higher the power available in the solar radiaton, the higher the generated voltage and (especially) the higher the generated current.

Such voltage current characteristics can be measured by loading the cell with an adjustable load. As the load resistance decreases, the voltage over the load decreases and the current increases. When short circuiting the cell, no voltage is obtained and the short circuit current I_{SC} (SC = Short Circuit) is obtained as visualised in Fig. 4.4.5.

As the load resistance increases, the voltage over the load increases and the current decreases. When the load resistance is very high, almost no current will flow. When considering such a no load situation, the open circuit voltage level U_{OC} (OC = Open Circuit) is obtained as visualised in Fig. 4.4.5.

The electrical power is maximum in the "maximum power point" indicated in Fig. 4.4.5. In this situation a current I_{MPP} and a voltage U_{MPP} account for a maximum power $P = U_{MPP}I_{MPP}$. When combining Fig. 4.4.4 and Fig. 4.4.5, it is clear the maximum power $P = U_{MPP}I_{MPP}$ strongly depends on the solar irradiance.

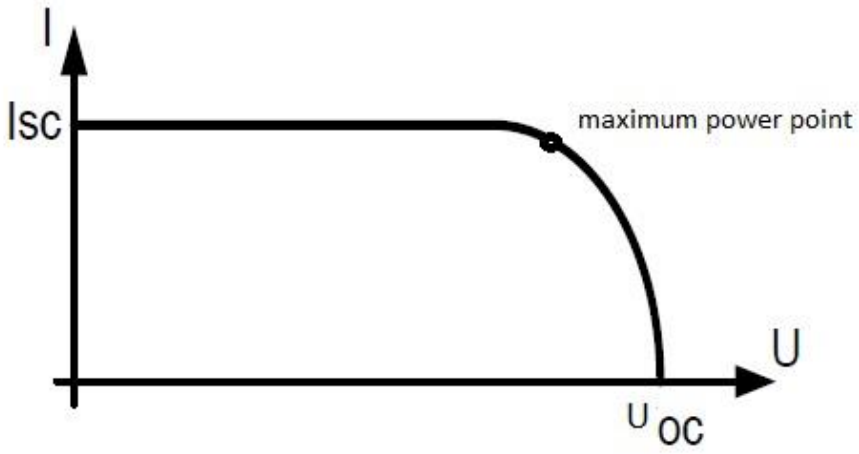


Fig. 4.4.5: The voltage current characteristic of a photovoltaic panel

A number of solar cells are connected in series to obtain a solar panel and a number of solar panels are connected in series to obtain a string. This implies solar panels and strings have voltage current characteristics which equal the shape visualised in Fig. 4.4.5 (although the total available voltage level is higher).

A lot of properties are valid when considering a single solar cell, a solar panel, a string or even a number of strings connected in parallel. Due to this reason, we will sometimes talk about a photovoltaic generator.

Standard conditions

When measuring the voltage current characteristic, it is important all measurements are performed at standard conditions. These standard conditions allow to compare the results provided by different constructors and different laboratories. These standard conditions require a temperature of $25\ °C$ and the lighting must be performed using a standard irradiance of $1000\ W/m^2$ with a standard spectrum A.M. 1.5. The notation A.M. (Air Mass) which equals 1.5 indicates the spectral content of the light has the spectral content of the sunlight at a latitude of $48.5°$ (where the distance travelled by the sunlight equals 1.5 times the height of the atmosphere).

The Maximum Power Point

The working point where the maximum power is extracted from the solar cell is the "Maximum Power Point" abbreviated as MPP. The rectangular red shaded area on Fig. 4.4.6, which represents the generated power, is maximized. The maximum power point on the voltage current characteristic of a single solar cell is obtained at a voltage level of approximately $0.45\ V$ to $0.5\ V$.

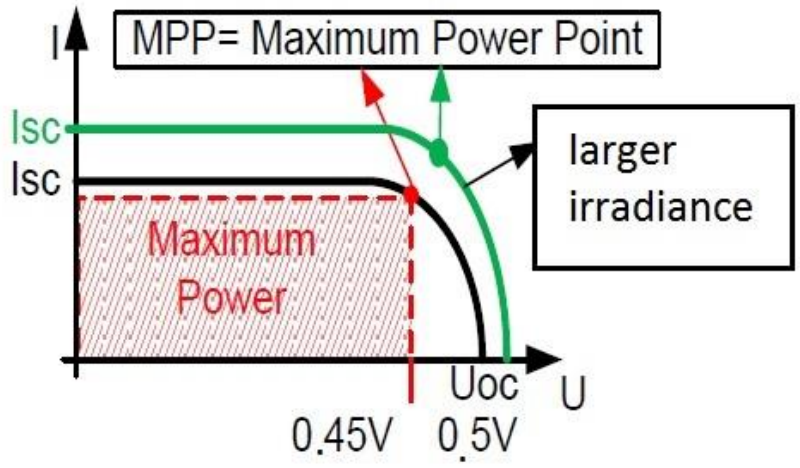


Fig. 4.4.6. The Maximum Power Point

The converter will, using a built in Maximum Power Point Tracker, take care the photovoltaic generator operates in the maximum power point (in case of one single solar cell this optimal voltage is situated between $0.45\ V$ and $0.5\ V$). The number of series connected solar cells, determines the total voltage of the maximum power point. When considering a string of ten solar panels with each containing 36 cells, a voltage between $162\ V$ and $180\ V$ is obtained.

The voltage level and the current in the maximum power point is noted as U_{MPP} and I_{MPP}. When considering Fig. 4.4.6, the green characteristic is obtained

when the solar irradiance is higher which changes U_{MPP}, U_{OC}, I_{MPP} and I_{SC}. The rectangle respresenting the generated power is larger when considering the green characteristic. Indeed, in case of a larger solar irradiance a larger power is generated at the Maximum Power Point.

The power curve

Not only the voltage current characteristic, but also the power curve has a typical shape. The power curve is obtained by considering the generated power $P = U.I$ as a function of the generated voltage U.

In Fig. 4.4.7, the power is visualised by the green curve. The area of the red shaded rectangle corresponds with the value on the vertical axis given by the green power curve. Notice this maximum power is obtained at the maximum power point of the voltage current characteristic.

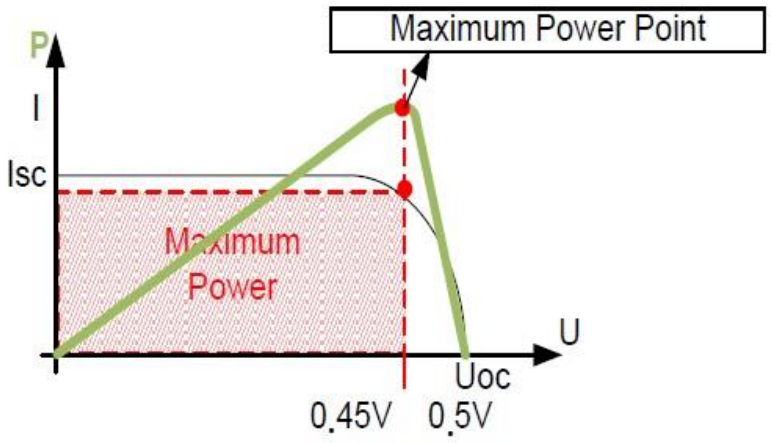


Fig. 4.4.7. The power curve

The nominal power of a solar panel is expressed as Watt peak or Wp. This respresents the maximum power the solar panel is able to generate at standard conditions. The data available in Table 4.4.1 also give this maximum power for a single square meter of solar panel area.

4.4.3. Technical data of a photovoltaic panel

A solar panel of the Multisol 200S type contains 8x12 solar cells. A single solar cell has a $U_{MPP} = 0.49\ V$. By connecting such 96 solar cells in series, each solar cell has the same $I_{MPP} = 4.42\ A$.

The voltage and the current mentioned in Fig. 4.4.8 are the U_{MPP} voltage ($47.5\ V$) and the I_{MPP} current ($4.42\ A$) generated by the Multisol 200 solar panel

in the maximum power point at standard conditions. This corresponds with the earlier mentioned maximum power Wp.

The short circuit current I_{SC} and the open circuit voltage U_{OC} are also mentioned. The dimensions, the mass and the termperature coefficients are mentioned too. The negative temperature coefficient for U_{MPP} indicates that U_{MPP} decreases as the temperature rises. As the temperature rises, the generated output voltage, the output power and the efficiency decrease (the impact of the temperature on I_{MPP} is small).

Table 4.4.1. Main data of Multisol 200A and 200S solar panels

Technical data Multisol 200			
Type		200A	200S
Peak power	Wp	$200 \pm 5\%$	$210 \pm 5\%$
Peak power per m^2	Wp/m^2	116	122
MPP-voltage	V	47.5	47.5
MPP-current	A	4.21	4.42
Open circuit voltage	V	57.5	57.7
Short circuit current	A	5.01	4.60
Temperature coefficient (U_{MPP})	$V/°C$	-0.19	
Temperature coefficient (I_{MPP})	$A/°C$	0.0024	
Dimensions	mm	1075x1600x42	
Mass	kg	20.0	

4.4.4. The photovoltaic installation at a private household

The convertor converts the DC input voltage to an AC voltage. This AC voltage has a frequency of $50\ Hz$ and allows to inject the power in the public low voltage grid. The convertor contains one single or a number of MPP trackers which force the operating point of the photovoltaic panels to the maximum power point. Fig. 4.4.9 visualises an example of a converter (a Sunny Boy) constructed by SMA (see e.g. http://www.sma-uk.com/).

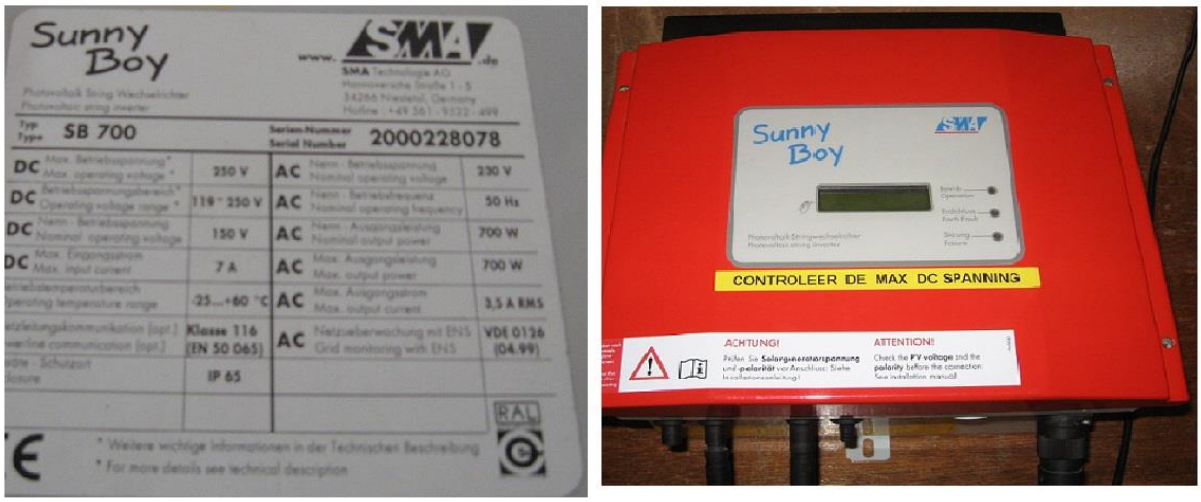


Fig. 4.4.9: Plate of the Sunny Boy converter

When considering a converter, the plate on the converter mentions the most relevant information. The maximum AC output power and the maximum AC output current are two relevant limits. When considering the Sunny Boy of Fig. 4.4.9, the maximum output power equals $700\ W$ (which is the rated power) and the maximum output current equals $3.5\ A$ in case of a $230\ V$ grid voltage. In case the invertor injects a sufficiently high power in the grid, the power factor is satisfactory high. In case only a limited power is injected into the grid, the power factor can be considerably lower.

The plate of the inverter also mentions information concerning the DC input voltage and the DC input current. This information is important when designing the photovoltaic generator.

The minimum length of a photovoltaic string

The allowed range of the DC input voltage of the converter determines the number of solar panels which will be connected in series to obtain a string. The plate in Fig. 4.4.9 mentions a voltage range between $119\ V$ and $250\ V$ which is a broad range. In case of an appropriate choice of the number of solar panels, this allows the installation to operate at different irradiances. Indeed, different irradiances imply a broad range of generated voltages.

At night, it is normal that U_{MPP} and even U_{OC} are lower than minimum required DC input voltage of the inverter. Indeed, at night no power will be generated and injeced into the grid. During the day, as much power as possible must be injected into the grid. In case the number of series connected solar panels

is too low, a larger solar irradation is needed to reach the minimum required voltage level of $119\ V$.

In case the number of solar panels is too small, or in case the number of solar cells in a single solar panel is too small, it occurs (too) often that the converter is unable to inject power due to a low DC input voltage. Although this accounts for energy losses, there is no danger of damaging the converter.

Fig. 4.4.10 and Fig. 4.4.11 visualise (red shaded) the working area of the converter. There is a minimum and a maximum DC input voltage. There is also a maximum DC input current and an upper limit for the power. In blue, the voltage current characteristics of the photovoltaic generators are visualised and the corresponding power curves are visualised in green. When considering the lowest voltage current characteristic, the U_{MPP}voltage of the photovoltaic generator is smaller than the minimum required DC input voltage. The inverter will operate at its minimum input voltage since this voltage level is lower than the U_{OC} voltage of the photovoltaic generator. Since the operating voltage is higher than the U_{MPP} voltage, the photovoltaic generator does not generate the maximum available power. In case of a higher solar irradiance, the higher voltage current characteristic is valid implying it is possible to operate at the maximum power point.

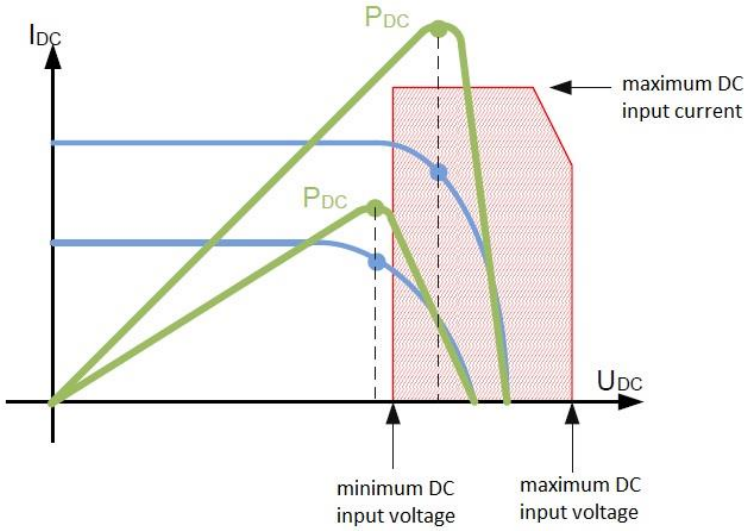


Fig. 4.4.10. The DC input voltage range

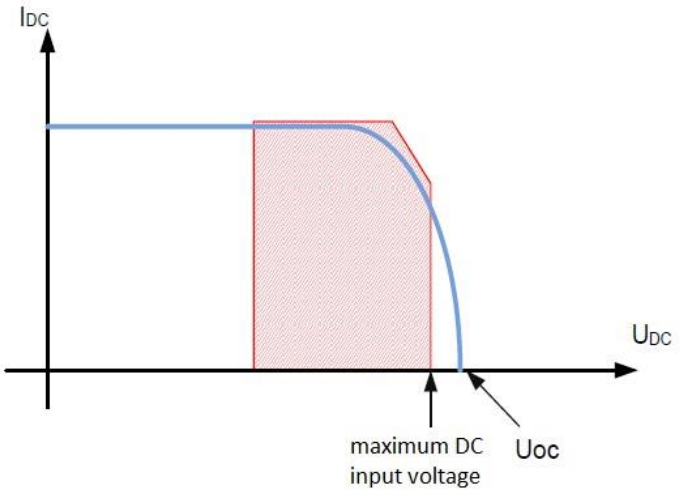


Fig. 4.4.11: The working area of the converter

The minimum number of photovoltaic panels which must be connected in a string depends on U_{MPP}. When considering the Sunny Boy in Fig. 4.4.9, and when using a solar panel containing 54 solar cells each having an $U_{MPP} = 0.45\,V$ ($U_{MPP} = 24.3\,V$ for each panel), at least five solar panels are needed. When considering these five solar panels, the photovoltaic generator is able to inject its power into the grid in case of standard conditions. When the U_{MPP} voltage decreases, for instance due to a lower solar irradiance, no operation in the maximum power point is possible.

The maximum length of a photovoltaic string

In case a larger number of solar panels are mounted in series (which also requires a larger financial investment), also in case of a lower solar irradiance the minimum required DC voltage can be obtained which allows the converter to inject the power into the grid. Notice, when considering a larger number of solar panels, care is needed in order to avoid the maximum allowable DC input voltage is not exceeded. The open circuit voltage U_{OC} of the photovoltaic generator must be lower than the maximum DC input voltage. It is important to satisfy this condition since an overvoltage at the input nodes, as visualised in Fig. 4.4.11, might damage the converter.

In case the voltage at the DC input of the inverter is too high, the large majority of the converters generate an error message. It is important to disconnect the photovoltaic generator from the converter in order to avoid damage.

The maximum output voltage U_{OC} approximately equals $0.6\,V$ for a single solar cell when considering standard conditions. When considering a solar panel

having 54 solar cells, this implies a voltage of $32.4\,V$. This means, when considering the data on the plate of the Sunny Boy in Fig. 4.4.9, a maximum of seven solar panels can be connected in series in one single string. This implies the U_{OC} voltage of the photovoltaic generator will not exceed the maximum allowed input voltage of the inverter (in this case, the minimum temperature of $-10\,°C$ must be considered since this mimimum temperature accounts for the maximum U_{OC}).

The generated power and the power ratio

The area of the solar cell has no impact on the generated voltage level. The area of the solar irradiation determines the amplitude of the generated current and the associated power level. The maximum allowed DC input current (7 A when considering the Sunny Boy of Fig. 4.4.9) determines the maximum allowed area of the solar cells. Possibly, more than one string can be connected in parallel at the input of the converter but the maximum input current limits the number of strings which can be connected in parallel.

Fig. 4.4.12 visualizes the upper limits for the DC input current and the DC input power. By operating outside the maximum power point of the photovoltaic generator, the power and the current are limited to levels allowed by the converter. Although the converter is operating at its maximum power, this is not the case for the photovoltaic generator.

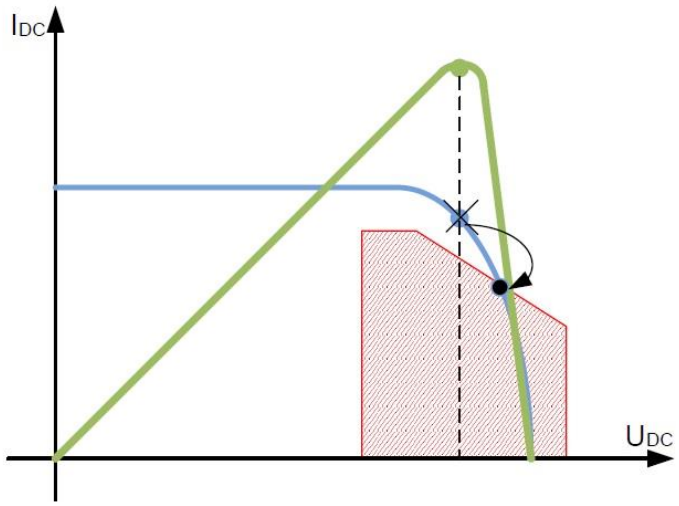


Fig. 4.4.12. The power limit

The power ratio is also frequently used when designing the photovoltaic generator and the converter (inverter). The power ratio equals the ratio between the maximum allowed DC input power of the converter (inverter) and the rated power of the photovoltaic generator.

When the power ratio is too small, the photovoltaic generator will not be able to operate at its full power in case of high solar irradiances. This implies power losses. At the other hand side, the installation will remain operational in case of lower solar irradiances.

When the power ratio is too high, in case of a lower solar irradiance the photovoltaic generator will not be able to inject its power into the grid. Indeed, the working point is far away from the maximum power point. At the other hand, the installation will operate properly in case of higher solar irradiance.

It is clear there exists an optimal choice for the power ratio. In countries where a lower solar irradiance is common, and a high solar irradiance does not occur very often, it is often useful to choose for a somewhat underdimensioned converter i.e. to take a power ratio lower than 100%. In reality, a power ratio of 90% is frequently chosen.

4.4.5. Blocking diodes and bypass diodes

Blocking diodes

In case a photovoltaic panel (or a number of photovoltaic panels) is used to load a battery, it is important to mount a blocking diode in series with the photovoltaic panel (Fig. 4.4.13). Due to this blocking diode, the photovoltaic panels are able to charge the battery but the polarity of the current cannot change. This is important at night or when having rainy weather when the battery voltage is higher than the voltage generated by the photovoltaic panels. Due to the blocking diode, a charged battery is not able to discharge and the photovoltaic panels do not behave as an electrical load.

Blocking diodes are also needed when a number of strings are connected in parallel. In case a first string generates a larger voltage than a second string, it is important to avoid the first string generates a current which flows through the second string which behaves as an electrical load. This situation can be avoided by placing a blocking diode in series with each string as visualised in Fig. 4.4.13.

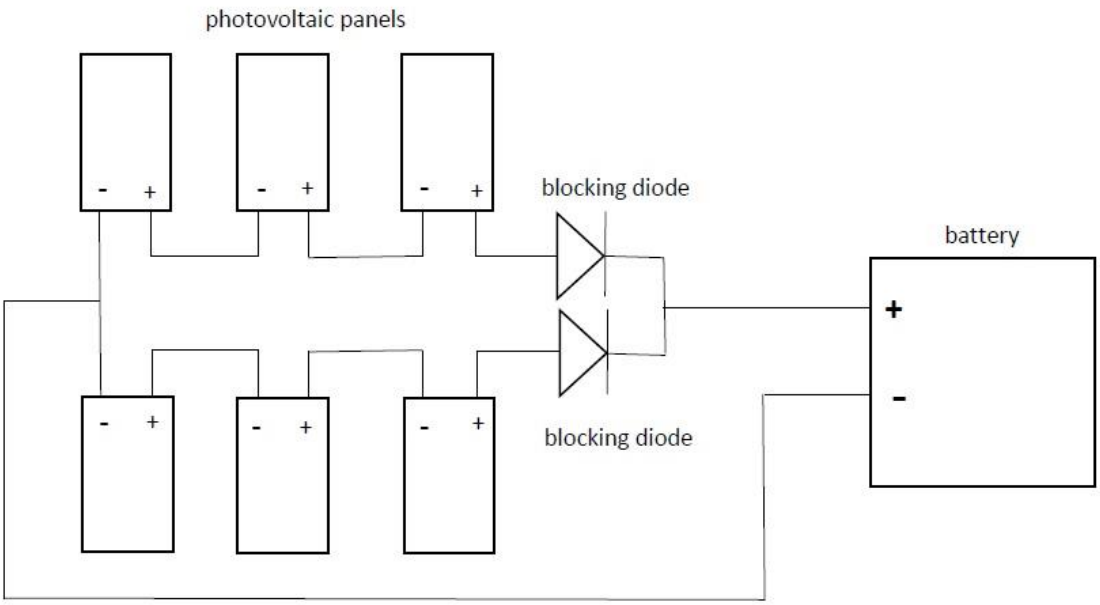


Fig. 4.4.13: Blocking diodes

Bypass diodes

Blocking diodes are connected in series with the photovoltaic panels. In a practical installation, also bypass diodes are used. These bypass diodes are connected in parallel with the photovoltaic panels. In a normal situation, no current is flowing in these bypass diodes as visualised in Fig. 4.4.14. In case a photovoltaic cell is shaded or broken, it mainly behaves as a resistance blocking the current in the other cells or panels connected in series. To avoid this phenomenon, a bypass diode is connected in parallel allowing the current to flow as visualised in Fig. 4.4.15.

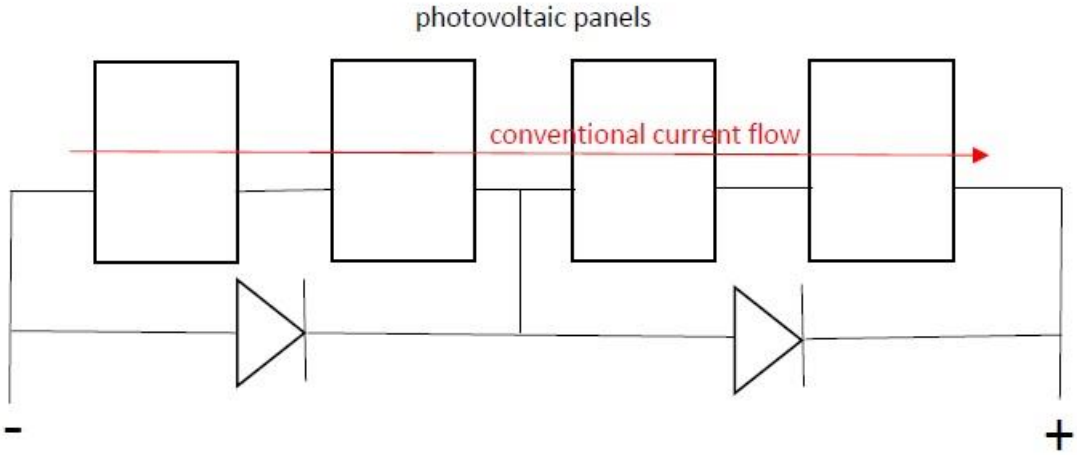


Fig. 4.4.14. Normal situation without influence of the bypass diodes

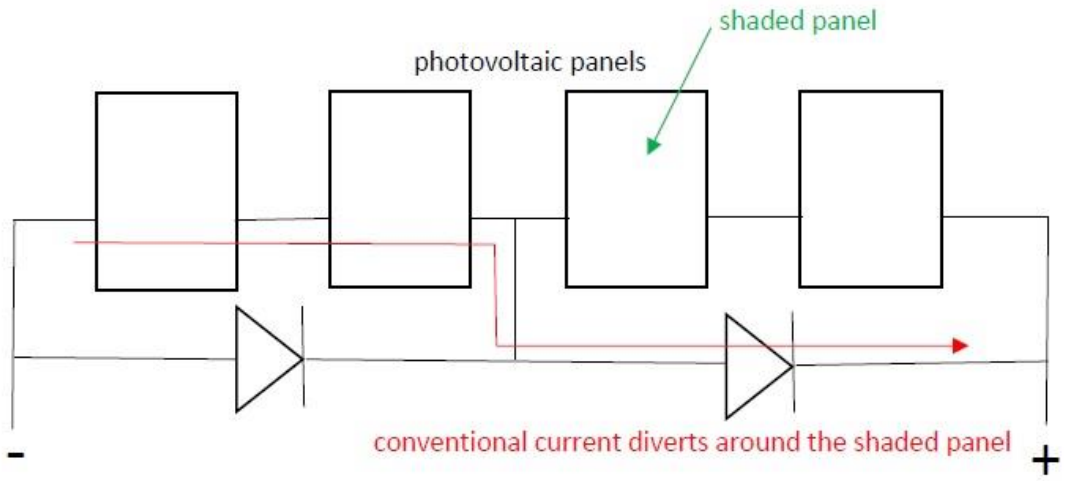


Fig. 4.4.15. The use of a bypass diode in case of a shaded panel

Depending on the manufacturer, these bypass diodes are mounted in the photovoltaic panel itself or they are mounted in the junction boxes used to connect the different panels. In case the bypass diodes are mounted in junction boxes, it is possible to replace them when necessary.

4.4.6. Grounding

Electrical devices are commonly grounded in order to guarantee a safe operation of the device. A distinction can be made between equipment grounding and system grounding. Legislations concerning the grounding of all kind of electrical devices and installations depend on the country. Despite these country dependent differences, there are a number of general physical principles.

When considering equipment grounding, the electrically conducting housing of the device is connected with the earth. This avoids (or reduces) the voltage difference between this housing and the ground (earth) which eliminates or reduces the voltage across the human body when someone touches the housing. This protects the human body against an electrical shock when touching the housing; even when a current-carrying conductor comes into contact with the housing (e.g. due to damaged electrical insulation) of the electrical device.

Grounding a photovoltaic installation: approach 1

Consider Fig. 4.4.16 where a string of photovoltaic panels generates a DC voltage. Using an inverter (converter), this DC voltage is converted into an AC voltage and the generated power will be injected into the public AC grid. The

active conductors have been drawn in solid lines whereas the electrical connections used for grounding have been drawn in dashed lines.

Notice the housings of the photovoltaic panels are connected with each other and grounded using a grounding electrode (equipment grounding). Such a grounding electrode is most often a rod or a ring which ensures an electrical contact with the earth.

When considering system grounding, a conductor of e.g. a two-wire electrical system is connected to ground. Notice in Fig. 4.4.16 that the negative conductor of the DC bus is grounded with the so-called DC grounding electrode. In Fig. 4.4.16 there is also an additional AC grounding electrode used to ground the neutral conductor of the AC grid at the AC side of the inverter (converter). In Fig. 4.4.16, the DC grounding electrode and the AC grounding electrode are also used to ground the housing of the inverter (equipment grounding). Notice also the bonding conductor connecting the DC grounding electrode and the AC grounding electrode.

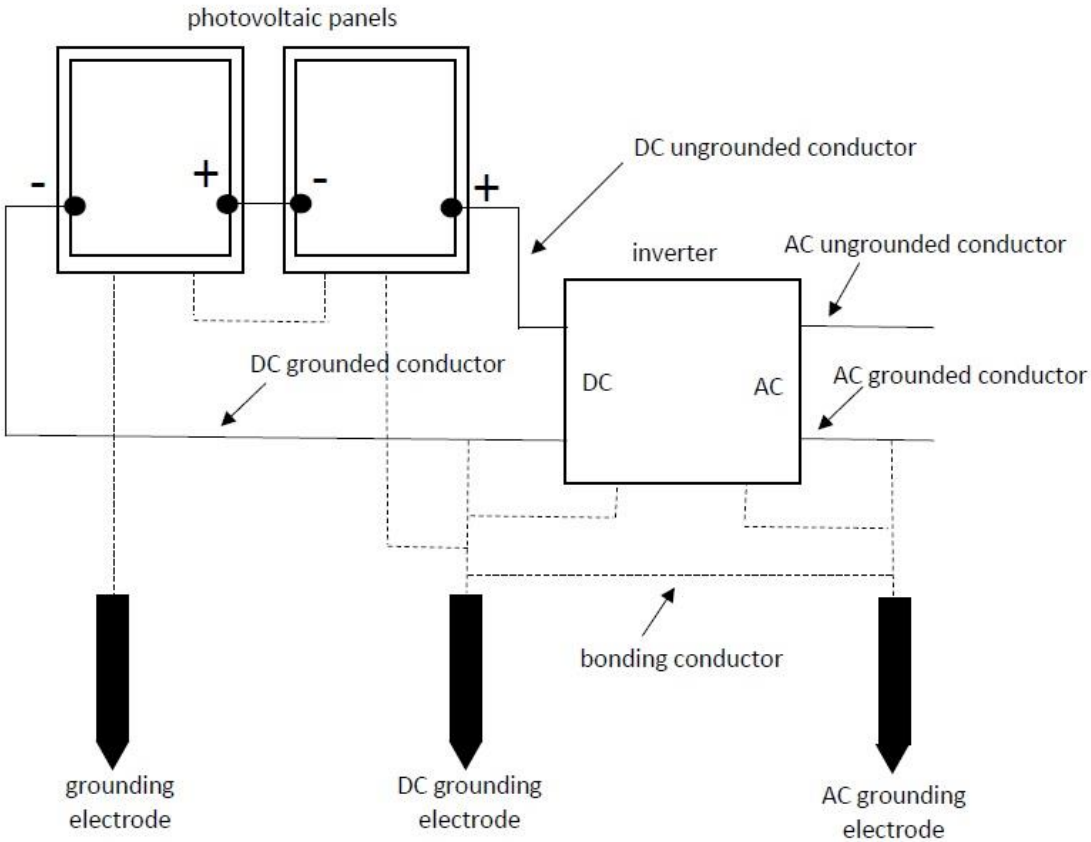


Fig. 4.4.16. Grounding using separate electrodes

Grounding a photovoltaic installation: approach 2

An alternative to the situation in Fig. 4.4.16 is given in Fig. 4.4.17. The equipment grounding of the housings of the photovoltaic panels does not change. In Fig. 4.4.17, the DC grounding electrode has been omitted (which reduces the installation cost). The AC grounding electrode is also used to ground the negative conductor of the DC bus (and additionally the housings of the photovoltaic panels). The AC grounding electrode is still used to ground the housing of the inverter (converter) and to ground the neutral conductor of the AC grid at the AC side of the inverter.

In case the photovoltaic panels are equipped with so-called double insulation, the grounding of the housings can be omitted. Notice however, a grounding can still be useful to drain electrical loads to the earth in case of a lightning strike.

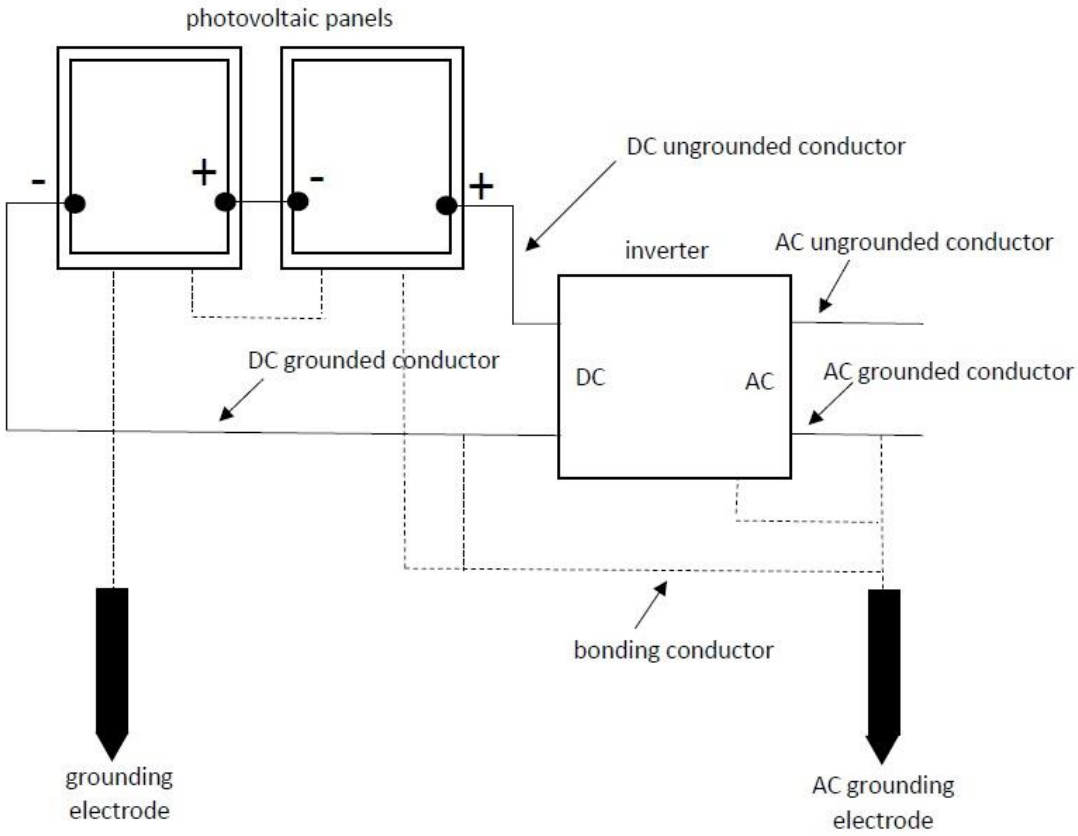


Fig. 4.4.17. Grounding using common electrodes

References

1. G. Boyle, Renewable Energy: Power for a Sustainable Future, The Open University, Oxford University Press, 2012.
2. J. P. Dunlop, Photovoltaic Systems, American Technical Publishers, Orland Park, Illinois, 2010.
3. J. Fraden, Handbook of Modern Sensors: Physics, Designs and Applications, Springer, London, 2010.
4. S. Franco, Design with operational amplifiers and analog integrated circuits, Mc Graw Hill International Editions, New York, 1998.
5. Koltun, M.M. Optics and metrology of solar cells / M.M. Koltun - M .: Science, 1984. - 280 p.
6. Zi, S. Physics of semiconductor devices: in 2 V. / S. Zi; transl. from English V.A. Gergel, V.V. Rakitin; by ed. R.A. Suris. - M .: Mir, 1984. - V. 2. - 456 p.
7. Terrestrial Photovoltaic Measurement Procedures, Technical Memorandum 73702, NASA, Cleveland, Ohio, 1977.
8. Fahrenbruch, A. Solar cells: theory and experiment / A. Fahrenbruch, R. Büb; transl. from English by ed. M.M. Koltun. - M .: Energoatomizdat, 1987. - 280 p.
9. Solar cells based on semiconductor materials (review). / V.F. Gremenok, M.S. Tivanov, V.B. Zalessky // Alternative energy and ecology. - 2009. - № 1 (69). - p. 59–124.
10. Shockley W., Queisser H.J. Detailed Balance Limit of Efficiency of p-n Junction Solar Cells // Journal of Applied Physics. 1961. Vol 32. P. 510-519.
11. Baruch P., De Vos A., Landsberg P.T., Parrott J.E. On some thermodynamic aspects of photovoltaic solar energy conversion // Solar Energy Materials and Solar Cells. 1995. Vol. 36. P.201-222.
12. Planck M. Distribution of energy in the spectrum // Annalen der Physik. 1901. Vol. 4. P. 553-563.
13. Dresselhaus M.S. Solid state physics. Part II. Optical Properties of Solids, MIT, 2001.
14. Garg H.P., Prakash J. Solar Energy: Fundamentals and Applications, Tata McGraw-Hill, 1997.
15. Abrams Z.R., Gharghi M., Niv A., Gladden C., Zhang X. Theoretical efficiency of 3rd generation solar cells: Comparison between carrier

multiplication and down-conversion // Solar Energy Materials and Solar Cells. 2012. Vol. 99. P.308-315.

16. Calculation the ultimate efficiency of p-n-junction solar cells taking into account the semiconductor absorption coefficient / M. Tivanov, A. Moskalev, I. Kaputskaya, P. Zukowski // Przegląd Elektrotechniczny (Electrical Review). – 2016. –R. 92, Nr. 8. – Pages 85-87.

17. Solar Energy International, Photovoltaics: Design and Installation Manual, New Society Publishers, Gabriola Island, Canada, 2004.

18. J. C. Wiles Jr, Photovoltaic System Grounding, Solar America Board for Codes and Standards, http://www.solarabcs.org.

19. Information is also obtained from several Wikipedia webpages (in Dutch or English): https://www.wikipedia.org/

Chapter 5. Optical waveguides

5.1. Optical waveguide basics

5.1.1. Optical waveguide structure

Optical waveguides are devices providing confinement of the light propagation direction, due to the total internal reflection effect at the interface of two media with different refractive indices, and transmission of light signals over long distances. Generally, a waveguide consists of a core made of the material with a higher refractive index and an environment (cladding) with a lower refractive index. When light is propagating in a medium with a higher refractive index compared to the environment, the total internal reflection effect is realized.

Let us consider the operation principle of optical fibers in some detail. The principal optical phenomena governing the mechanisms of light propagation in optical fibers are the effects of light *reflection* and *refraction* at the interface of two media having different refractive indices. Fig. 5.1.1 shows the geometrical constructions illustrating the essence of the involved phenomena.

Light propagating from the medium with a lower refractive index n_2 to the medium with a higher refractive index n_1 is deflected as a perpendicular to the interface between these two media (dashed line in Fig. 5.1.1, *a*). As this takes place, the angle of refraction Θ_r is smaller than the incidence angle Θ_i.

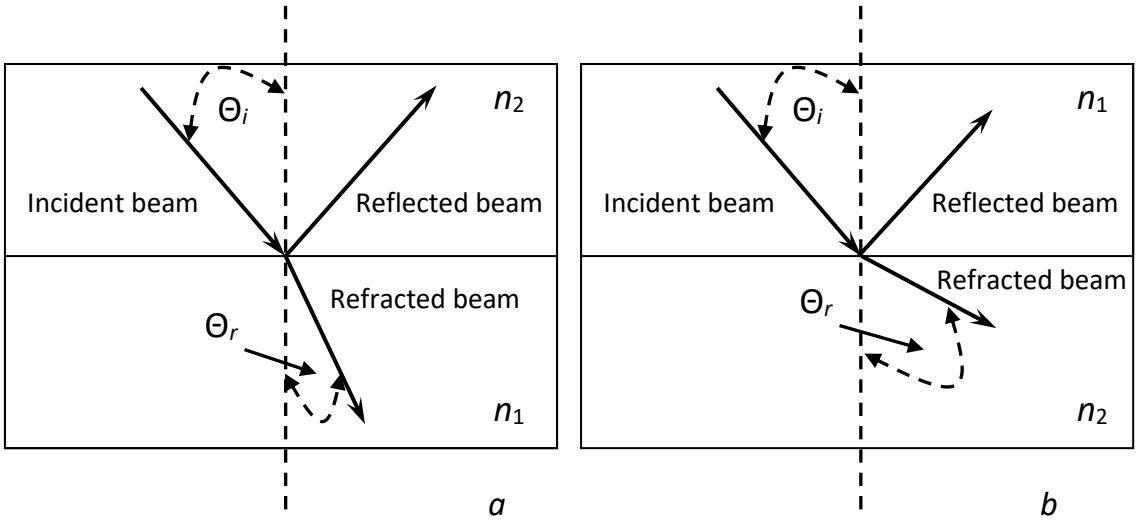


Fig. 5.1.1. Reflection and refraction at the interface of two materials ($n_2 < n_1$)

To the contrary, on passage from the medium with a higher optical density (n_1) to that with a lower density (n_2), the angle of refraction Θ_r is greater than the incidence angle Θ_i, and the beam is deflected toward the interface between the two media (Fig. 5.1.1, *b*). As the light beam incidence angle is increased, the

refraction angle approaches 90° and the refracted beam is grazing along the interface (Fig. 5.1.2, *a*). With further increase in the incidence angle, light no longer penetrates the material with a lower density n_2, being totally reflected from the interface of the two media (Fig. 5.1.2, *b*), i.e., the *total internal reflection* effect occurs. In this case the angle of reflection is always equal to the incidence

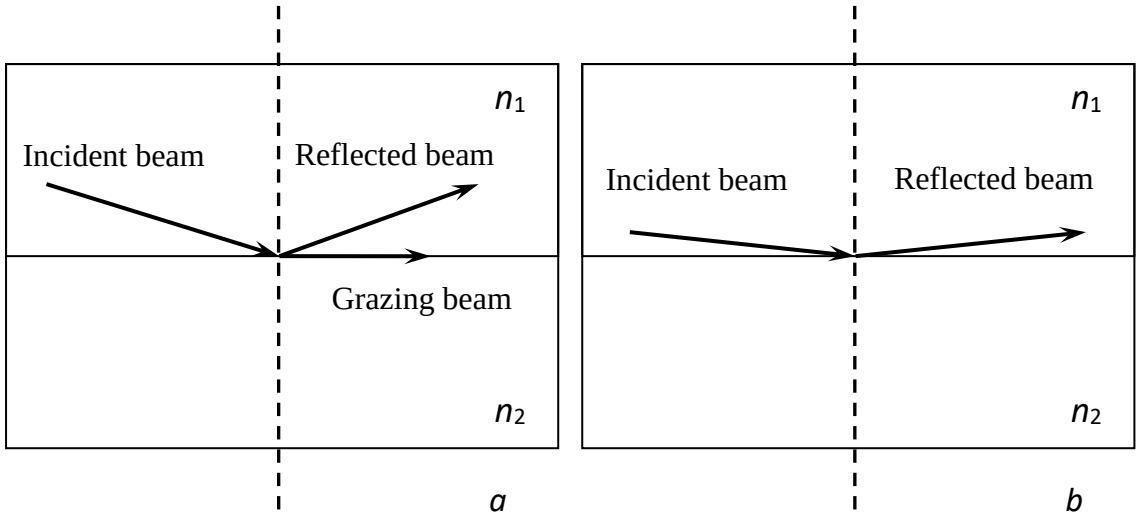


Fig. 5.1.2. The critical incidence angle (a) and the total internal reflection phenomenon (b) ($n_2 < n_1$)

angle.

The relation of the angles, at which light is refracted from the interface of the two media, is prescribed by the Snell law as follows:

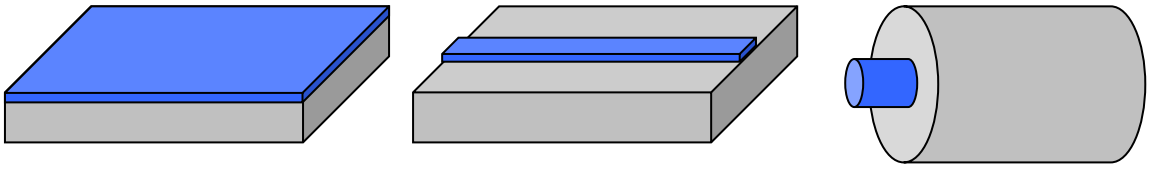
$$n_1 \sin\Theta_i = n_2 \sin\Theta_r, \qquad (5.1.1)$$

where the angles Θ_i and Θ_r are as indicated in Fig. 5.1.1. Proceeding from this law, we can determine the critical angle that, according to the condition $\Theta_r = 90°$, equals

$$\Theta_C = \arcsin(n_1/n_2). \qquad (5.1.2)$$

When the incidence angle is above the critical angle Θ_C, light is totally reflected. This effect underlies the operation principle of fiber-optical data transmission systems.

Fig. 5.1.3 schematically shows the devices used to implement the waveguide light propagation. A planar waveguide (Fig. 5.1.3, *a*) confines the light propagation to one, vertical, direction only, whereas a channel waveguide (Fig. 5.1.3, *b*) – to two directions. Optical fiber (Fig. 5.1.3, *c*) – the most common waveguide type –also confines the light propagation to two spatial directions.



Planar waveguide Channel waveguide Optical fiber

a *b* *c*

Fig. 5.1.3. Different types of waveguides: planar (a), channel (b), and optical fibers (c)

As a rule, optical fibers used for data transmission have small cross-sectional dimensions. Typical diameters are as follows: from 4 to 100 μm for the core, 125 μm – for the cladding. In the dimensions of an optical fiber the core diameter (in micrometers) is indicated first and then the optical cladding diameter is specified, e.g., 50/125. Generally, optical fiber is surrounded by an additional coating (buffer) to protect it against any mechanical damage and environmental effects – this coating is of no importance for waveguide propagation of light. The majority of fiber-optical systems employ infra-red light, with the wavelengths exceeding those of the visible light and ranging from 800 to 1500 nm (or from 0.8 to 1.5 μm). The materials commonly used in fiber-optical systems are most transparent just for infer-red radiation, offering lower energy losses.

5.1.2. Classification of optical fibers

The characteristics of optical fibers largely determine their applications. Most often optical fibers are classified by the material used to produce this fiber, by the refractive index profile or the mode structure of radiation propagating in optical fiber. Let us consider these characteristics in detail.

All optical fibers are subdivided into two main groups: single-mode fiber (SMF) and *multimode fiber* (MMF).

Ordinary, the core diameter of **single-mode fibers** is from 7 to 10 micron. We distinguish between *step index single mode fibers*, so-called *standard fibers* (SF), *dispersion-shifted single mode fibers* (DSF), and *non-zero dispersion-shifted single-mode fibers* (NZDSF).

The core diameter **of multimode fibers** is generally 50 micron according to the European standard and 62.5 micron according to the North American and Japanese standards. These fibers may be subdivided into *step index multimode fibers* and *graded index multimode fibers*.

Materials used. Most often, optical fibers are manufactured of glass and plastic materials. The core and cladding of glass fibers is made of glass. Such glass consists of superpure silicon dioxide or fused silica. The impurities are usually added to modulate the refractive index of high-purity glass: germanium and phosphorus – to increase the refractive index, boron and fluorine – to lower the refractive index. Plastic fibers have plastic core and cladding; such fibers are characterized by rather high attenuation of an optical signal and by the limited transmission band. However, their low cost and simplicity of use attract the developers of short-distance fiber-optical communication lines. A plastic fiber is, as a rule, employed for waveguide propagation of red light over the wavelength range close to 660 nm. Combined fibers consisting of a glass core and a plastic cladding are also available. Besides, some optical amplifiers are produced with the use of fibers having rare-earth elements as impurities.

Refractive index profile. Fiber waveguides are also classified according to the cross-sectional refractive index profile. There are two main profile types: step-index and smoothed graded-index. Fig. 5.1.4 demonstrates waveguides which differ in their refractive index profiles. As seen, in the cases *a*) and *b*), the refractive index profile is of the step-index type, i. e., there is a step-wise change from refractive index of the core to that of the optical cladding. Note the difference

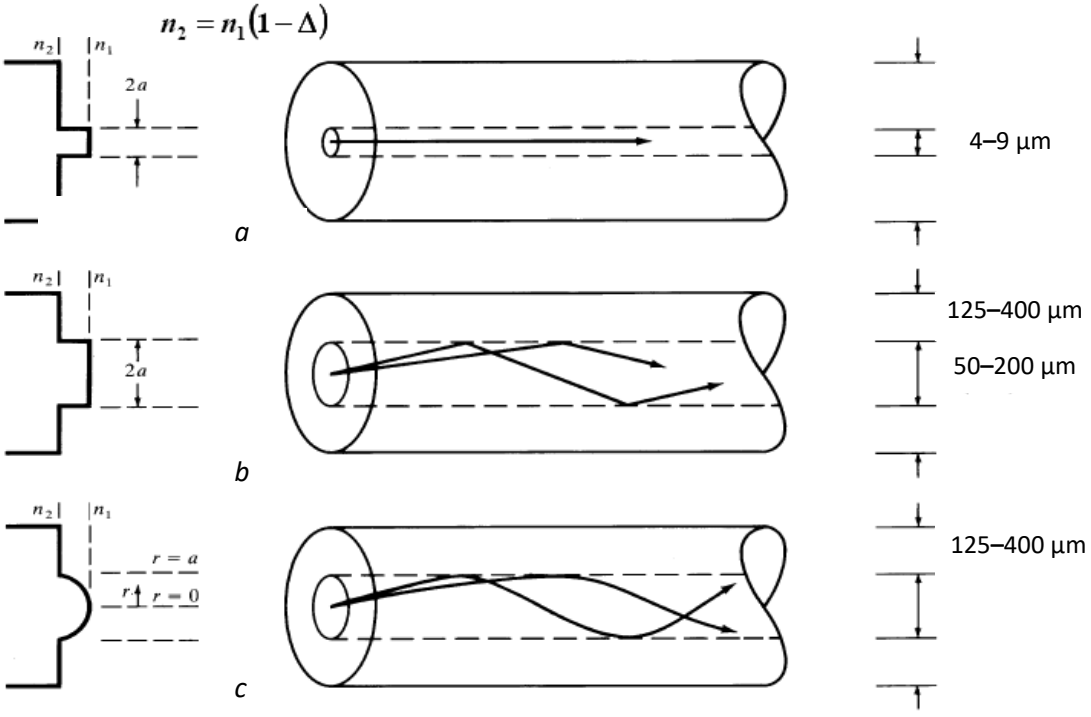


Fig. 5.1.4. Types of fiber-optical waveguides: step-index single-mode (a); multimode (b); graded-index multimode (c)

in diameters of the core in Figs. 5.1.4, *a* and 5.1.4, *b*. A graded-index fiber, or a graded-profile waveguide, is shown in Fig. 5.1.4, *c*. A refractive index of such a waveguide is varying smoothly from the core center to its edges. The beams propagating in such a waveguide are bent in the direction of the refractive index gradient. Most common are the fibers with the refractive index varying by the parabolic law: $n(r) = n_1\left[1 - 2\Delta r/a^2\right]^{1/2}$, where $\Delta = (n_1 - n_2)/n_1$ – relative difference in refractive indices.

Apart from optical data transmission systems, optical fibers find applications in other fields: gyroscopes; image transmission systems for spectroscopy and endoscopy; generation and transmission of high-power laser radiation; supercontinuum generation. Waveguides may be more complex in their structure in accordance with the specific problems solved: doubly clad fibers; polarization-maintaining fibers; hollow-core fibers; microstructured fibers.

5.2. Waveguide modes

The *mode* of an electromagnetic field propagating in a medium or within some optical device is understood as a stable spatial configuration of the field self-sustained in the process of propagation. Such configuration should satisfy a system of Maxwell equations for the electric-field strength vectors $\vec{E}$, magnetic field $\vec{H}$, and the vectors of electric $\vec{D}$ and magnetic $\vec{B}$ induction:

$$\begin{aligned} &\mathrm{rot}\,\vec{H} - \frac{1}{c}\frac{\partial \vec{D}}{\partial t} = 0, \quad \mathrm{rot}\,\vec{E} + \frac{1}{c}\frac{\partial \vec{B}}{\partial t} = 0, \\ &\mathrm{div}\vec{D} = 0, \qquad\qquad \mathrm{div}\vec{B} = 0. \end{aligned} \tag{5.2.1}$$

Besides, the characteristics of electric and magnetic fields are related to the medium parameters (material equations): $\vec{D} = \varepsilon\vec{E}$, $\vec{B} = \mu\vec{H}$, where ε – dielectric and μ – magnetic permeability of the medium.

Mathematically, the notion of the electromagnetic field mode is associated with some of the solutions for the reduced system of differential equations that is derived with regard to the boundary conditions determined by the conditions of light reflection from the medium boundaries. Equations (4.1), written in the assumption of the absence of currents and charges in the spatial region under study, adequately describe the propagation patterns of optical-range electromagnetic waves in dielectrics. At the same time, it is more convenient to study the electromagnetic wave propagation processes with the use of a wave

equation that is derived from Maxwell equations by means of the following simple mathematical transformations:

$$\Delta\vec{E} - \frac{\varepsilon\mu}{c^2}\frac{\partial^2\vec{E}}{\partial t^2} = 0, \qquad (5.2.2)$$

where $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ – Laplacian. Similar equations are also valid for the magnetic field vector $\vec{H}$.

Plane harmonic waves are the simplest solutions for the wave equation in a free space (or an optically homogeneous medium)

$$\vec{E} = \vec{E}_0 \exp(i\vec{k}\cdot\vec{r} - i\omega t), \quad \vec{H} = \vec{H}_0 \exp(i\vec{k}\cdot\vec{r} - i\omega t), \qquad (5.2.3)$$

where $k = \omega/v$ – wave number; $v = c/n$ – speed of light; $n = \sqrt{\varepsilon\mu}$ – refractive index; ω – circular frequency of an electromagnetic wave. When these waves propagate in a medium, only the field oscillation phase $\beta = kn$ at the specific point is varying. In this way plane waves are the simplest *modes* of an electromagnetic field in a homogeneous medium. Note that in the case of nonmagnetic media (generally considered as a medium for optical waveguides) magnetic permeability of the medium is $\mu = 1$.

An electromagnetic field in laser beams is mainly concentrated along the longitudinal axis of the beam, rapidly going to zero in the transverse directions. The above-mentioned Gaussian beams represent another example of the *modes* of an electromagnetic field. Owing to their symmetry about the transverse coordinates, such beams describe the simplest modes of an electromagnetic field not only in the free space but also in some optical systems, e.g., in optical cavities.

Let us consider the propagation of an electromagnetic field in optical waveguides. First, we consider the planar waveguide structure (Fig. 5.2.1). The captured light beams forming guided modes of the core should meet specific phase conditions for the description of self-consistent fields. The beams within the core correspond to standing waves formed on wave interference. For propagation at angle Θ with respect to the axis, the wave vector has the following axial component:

$$\beta = n_1 k_0 \cos\Theta, \qquad (5.2.4)$$

where $k_0 = \omega/c$ – wave number in the vacuum.

This axial component of the wave vector is called the waveguide mode propagation constant. The relation

$$n_{ef} = \beta/k_0 = n_1 \cos\Theta \qquad (5.2.5)$$

is referred to as the *effective index of refraction* for the waveguide mode and gives the beam delay in a waveguide as compared to its free-space propagation. The beam path length and hence the propagation time is dependent on the incidence angle, being different for different modes. The effective length of optical fiber for every mode equals

$$L_{ef} = Ln_{ef}, \tag{5.2.6}$$

where L – fiber length.

Fig. 5.2.1 shows a typical path of the beam *ABCD* that may be used to describe a homogeneous plane wave propagating in a planar optical waveguide. The dashed lines correspond to the phase front of a homogeneous plane wave travelling from *A* to *B* and from *C* to *D*, respectively. After the two successive reflections (e.g., at the points *B* and *C*), a phase of the wave should be retained to form the self-consistent field distribution corresponding to the particular fiber mode. Provided this phase condition is not met, the interfering waves are quenching each other. Let us consider a phase front including the points *B* and *E*. When a wave is travelling from *E* to *F*, a front of the wave *BE* is transformed to the phase front *FC*. Constructive interference necessitates that, with regard to two reflections at the points *B* and *C*, the phase difference between the two paths (*BC* and *EF*) be equal to $2m\pi$, where m – integer.

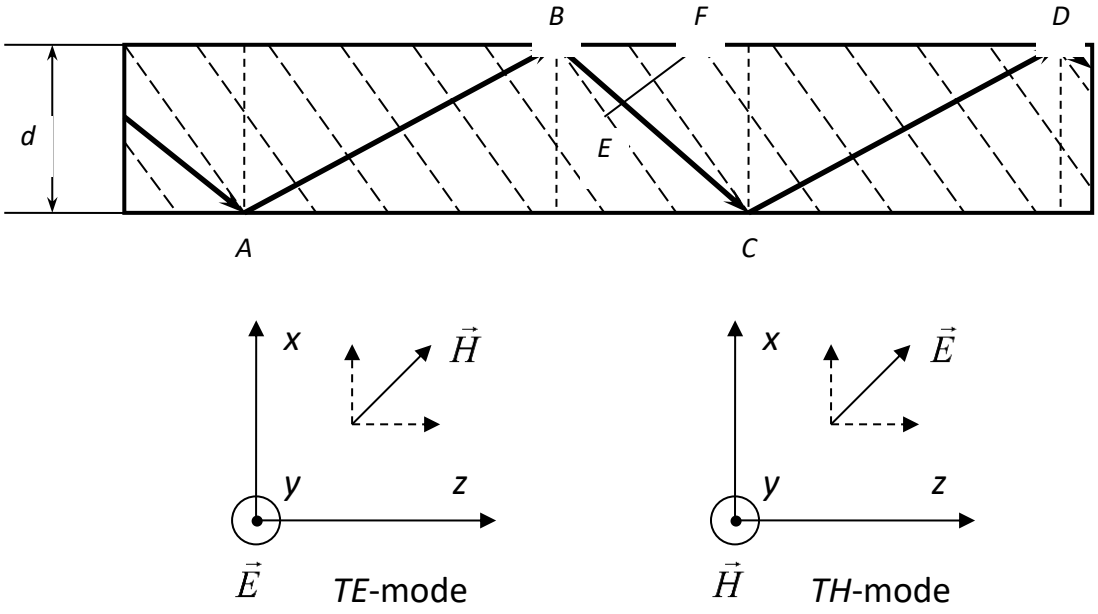


Fig. 5.2.1. Formation of planar waveguide modes

From the viewpoint of a wave theory, we can find possible configurations of an electromagnetic field for a planar waveguide by solving Maxwell equations with due regard for the waveguide geometry. Considering transversality of electromagnetic waves (the vectors $\vec{E}$ and $\vec{H}$ are mutually perpendicular) in the stationary mode, the Maxwell equations break down into two independent systems as follows:

$$ikH_x - \frac{\partial H_z}{\partial x} = -i\omega\varepsilon E_y, \quad -kE_y = \omega\mu H_x, \quad \frac{\partial E_y}{\partial x} = i\omega\mu H_z; \tag{5.2.7}$$

$$ikE_x - \frac{\partial E_z}{\partial x} = i\omega\mu H_y, \quad kH_y = \omega\varepsilon E_x, \quad \frac{\partial H_y}{\partial x} = -i\omega\varepsilon E_z. \tag{5.2.8}$$

In this case system (5.2.7) describes the so-called transverse-electric (TE) waves, whereas system (5.2.8) – transverse-magnetic (TM) waves (see Fig. 5.2.2).

Spatial distributions of the lower-order TE modes are demonstrated in Fig. 5.2.2. As seen, the electric field strength in this case is varying harmonically within the core and exponentially – within the cladding. Number of zeros in the amplitude along the transverse axis x (perpendicular to the propagation direction z) determines the mode index: TE0, TE1, TE2, etc. A similar situation is observed for TM mode too. Note that, due to the fixed width of a waveguide, a

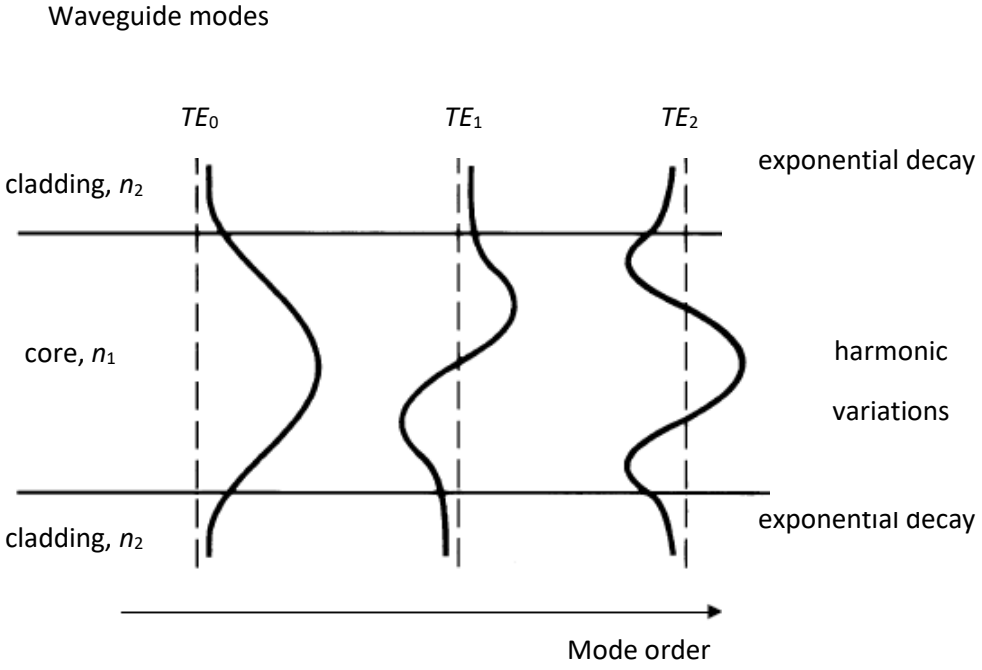


Fig. 5.2.2. Transverse distribution of the mode field for a planar waveguide

number of the modes propagating in the waveguide is limited. A decrease in the width leads to the decreased number of the modes propagating in the waveguide. Decrease in the thickness results in the reduced number of the modes propagating in the waveguide. By selection of a particular thickness, we can get a single-mode waveguide, where only TE_0- и TM_0 modes can exist.

Calculation of all possible spatial distributions for an electromagnetic field in the core of a circular (fiber) waveguide is a much more complex problem. Even from the viewpoint of beam analysis, it is clear that the total internal reflection condition may be fulfilled not only for meridional beams travelling along the central axis of a fiber, but also for asymmetric beams propagating along the fiber and bypassing its central axis. Their path represents a spiral turning about the central axis. Propagation of an oblique beam in the core of a fiber waveguide is shown in Fig. 5.2.3.

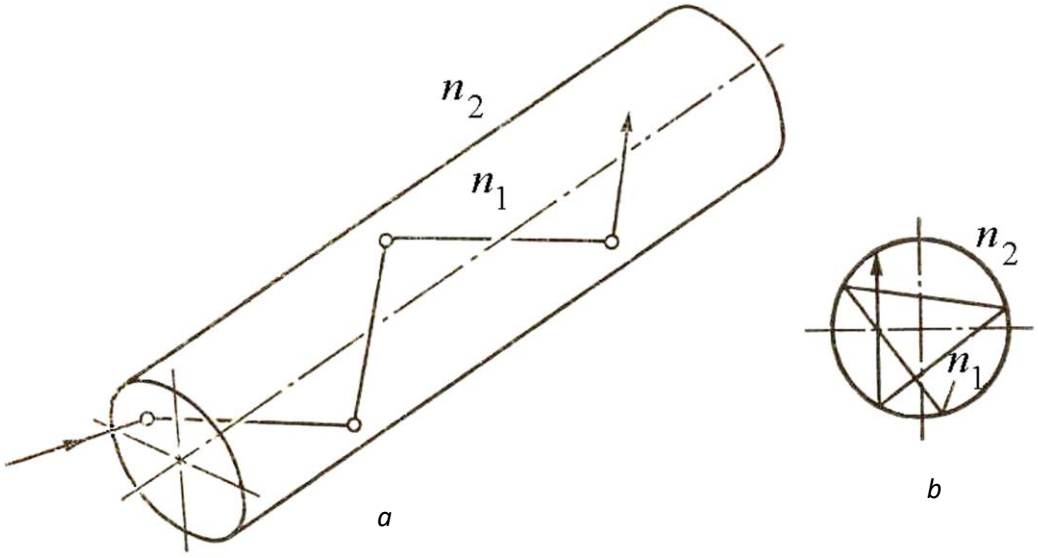


Fig. 5.2.3. Propagation of an oblique beam in a circular dielectric waveguide:
a – beam propagation along the broken spiral path; b – beam path projection onto the 125 μm fiber end

The trapped light beams forming a guided mode of the core should meet the specific phase conditions for description of self-consistent fields. These beams in the core correspond to standing waves formed by the superimposed waves. For the formation of the standing wave pattern in the azimuthal direction, the number of intersections with the boundary circle must be an integer.

A more accurate description of radiation propagation in fiber waveguides may be obtained with the use of wave analysis based on solutions of Maxwell equations. In a cylindrical coordinate system (r, φ, z) Maxwell equations are transformed as follows:

$$\frac{1}{r}\frac{\partial H_z}{\partial \varphi} - \frac{\partial H_\varphi}{\partial z} = -i\omega\varepsilon_j E_r, \quad \frac{\partial H_r}{\partial z} - \frac{\partial H_z}{\partial r} = -i\omega\varepsilon_j E_\varphi,$$

$$\frac{1}{r}\frac{\partial}{\partial r}\left(rH_\varphi\right) - \frac{1}{r}\frac{\partial H_r}{\partial \varphi} = -i\omega\varepsilon_j E_z; \tag{5.2.9}$$

$$\frac{1}{r}\frac{\partial E_z}{\partial \varphi} - \frac{\partial E_\varphi}{\partial z} = i\omega\mu H_r, \quad \frac{\partial E_r}{\partial z} - \frac{\partial E_z}{\partial r} = i\omega\mu H_\varphi,$$

$$\frac{1}{r}\frac{\partial}{\partial r}\left(rE_\varphi\right) - \frac{1}{r}\frac{\partial E_r}{\partial \varphi} = i\omega\mu H_z. \tag{5.2.10}$$

In these equations the quantities with the subscript $j=1$ belong to the waveguide core region, whereas those with the subscript $j=2$ – to the surrounding cladding. The transverse field components E_z and H_z satisfy the wave equation

$$\frac{\partial^2 U}{\partial r^2} + \frac{1}{r}\frac{\partial U}{\partial r} + \frac{1}{r^2}\frac{\partial^2 U}{\partial \varphi^2} + (k^2 - \beta^2)U = 0, \tag{5.2.11}$$

with the solution of the form

$$U = F(r)e^{im\varphi}, \tag{5.2.12}$$

where m – integer. The functions $F(r)$, in turn, represent solutions of the Bessel differential equation

$$\frac{\partial^2 F}{\partial r^2} + \frac{1}{r}\frac{\partial F}{\partial r} + \left(\frac{u^2}{a^2} - \frac{m^2}{r^2}\right)F = 0, \tag{5.2.13}$$

where $u^2 = a^2(k_1^2 - \beta^2)$; a – radius of the fiber core. In this way the modes of a cylindrical waveguide are represented in the form of the superimposed Bessel functions.

Some of the solutions of equations (5.2.9)–(5.2.13) are given in Fig. 5.2.4. In this geometry the mode indices have dual numbering. The first index indicates a number of the field variations for the azimuthal angle φ in accordance with formula (5.2.12). The second index, pointing to the field nodes in the radial direction, is the radial order of a particular mode and corresponds to the Bessel function that is a solution of (5.2.13).

As distinct from planar waveguides, propagating waves may be attributed to TE or TM modes in the case $m=0$ only. In the beam pattern these axially symmetric modes are formed by meridional beams. Any nonmeridional beam can form solely the asymmetric mode with $m \neq 0$. In this case only superposition of

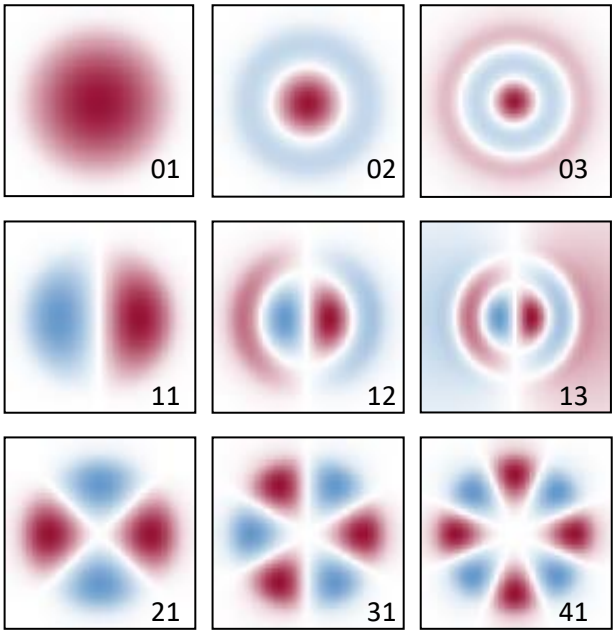


Fig. 5.2.4. Electric field strength distribution in some transverse modes of optical fiber

the E_z- and H_z longitudinal components of an electromagnetic field can meet all the boundary conditions. The modes lose their transversality to become hybrid modes. Hybrid modes are demoted as HE_{mp} and EH_{mp} modes. The first index corresponds to an integer number of field variations with respect to the angular component, i.e. denotes the azimuthal mode order. The second index indicates the order of the origination of HE_{MP} and EH_{MP} modes with the increased parameter

$$V = k_0 a (n_1^2 - n_2^2)^{1/2} = \frac{2\pi a}{\lambda} NA, \qquad (5.2.14)$$

where λ – light wavelength; a – fiber core radius; $NA = \sqrt{n_1^2 - n_2^2}$ – numerical aperture of a fiber. Axially-symmetric modes are denoted as TM_{0p} and TE_{0p} modes. HE_{11} is a lower-order hybrid mode representing the principal mode of a fiber.

The quantities β, derived by solving of a dispersion equation for the modes of a cylindrical optical waveguide, represent the propagation constants or the phase constants of these modes. The solutions for different values of the normalized frequency $V = k_0 a (n_1^2 - n_2^2)^{1/2}$ are shown in Fig. 5.2.5, where $n_{ef} = \beta / k_0$ – effective index of refraction.

As seen in Fig. 5.2.5, each mode is associated with a cutoff frequency (wavelength), i.e., a minimal frequency (maximal wavelength) of light, with which the mode propagates in a waveguide of a given size at the specified ratio of refractive indices of the core and of the cladding. The vertical line in Fig. 5.2.5

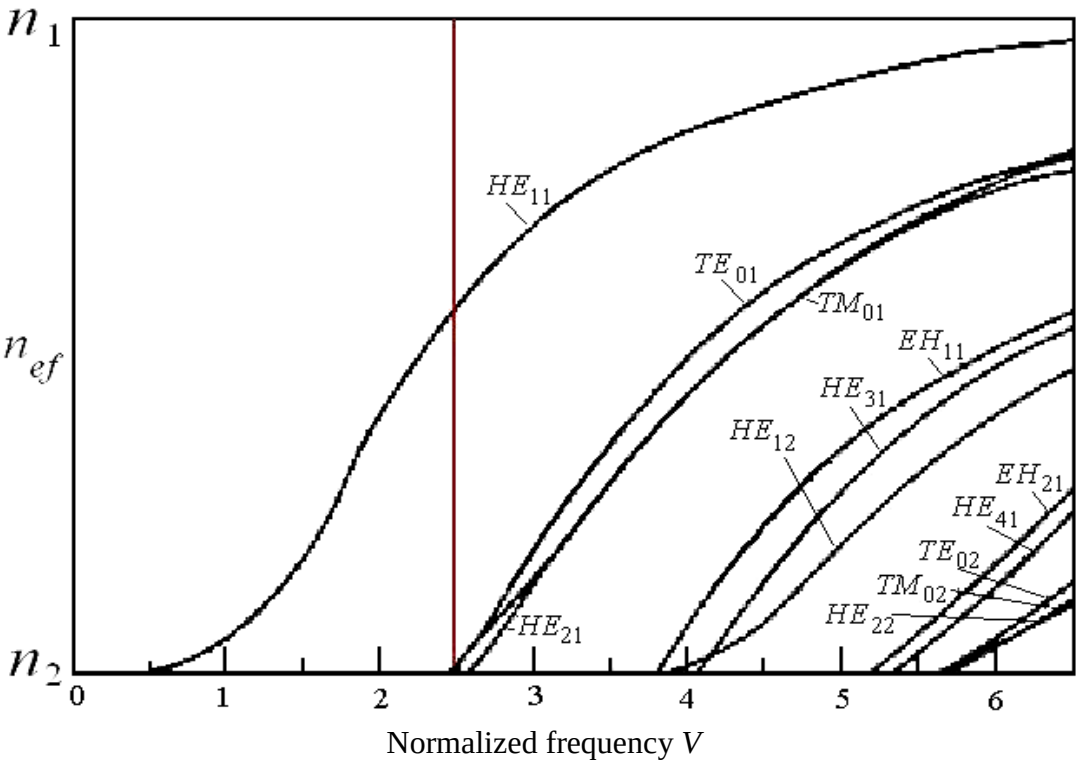


Fig. 5.2.5. Effective index of refraction for modes in a circular optical waveguide as a function of the normalized frequency

indicates the critical frequency of the first higher mode and the single-mode region of a waveguide. A number of the guided modes dependent on the core diameter, wavelength, and aperture of a fiber is determined as

$$N \approx 0{,}5V^2. \qquad (5.2.15)$$

This expression is valid for the majority of the guided modes. According to the guided mode number, fibers are subdivided as follows: single-mode $N=1$ (HE_{11}mode) with the core diameter $2a \approx \lambda$; few-mode $N \le 20$ at $2a > \lambda$; multimode $N > 20$ at $2a >> \lambda$. Multimode fibers have sufficiently large diameters of the core, whereas the diameter of single-mode fibers is comparable with the wavelength.

To illustrate, when the refractive index of the core is $n_1 = 1{,}45$ and of the cladding – $n_2 = 1{,}445$, the numerical aperture $NA = \sqrt{n_1^2 - n_2^2} \approx 0{,}12$, and hence, according to Fig. 4.5 and formula (4.14), the single-mode regime necessitates fulfillment of the condition $2\pi a/\lambda \le 20$. Then, for the wavelength λ = 1.55 μm (near IR region), the single-mode regime is realized when the core diameter is $2a <$ 10 μm. At the same time, at the wavelength 650 nm (red spectral region) a

fiber of the diameter 10 μm is a few-mode fiber. As a rule, the core diameter of multimode fibers ranges from 50 to 100 μm.

Numerous modes are excited when radiation is propagating in a multimode fiber. A field at the fiber output is determined by interference of different-order waveguide modes, each of which has radial and (or) azimuthal symmetry. Light fields interfere with the phases depending on the propagation rate of different modes and on the length of the fiber segment under consideration. A structure of the formed interference field is complex and depends on the number of interacting modes and on the cross-sectional coordinate. With a sufficiently great number of the modes (N>20), the pattern is similar to the speckle structure well known in holography. Fig. 5.2.6 illustrates the spatial intensity distributions at the output of a multimode optical fiber.

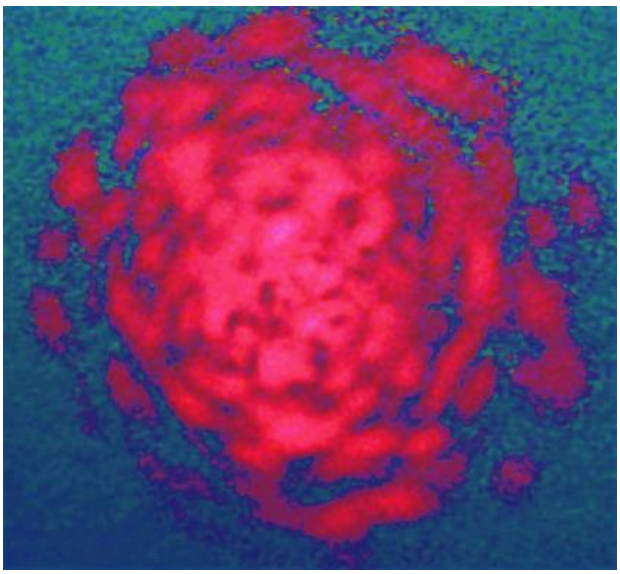


Fig. 5.2.6. Spatial intensity distribution at the output of the 50–100 μm multimode fiber

The speckle structure of radiation at the output of a fiber contains all information about the coherent mode propagation process. Specifically, when deformation of a fiber is due to external factors, this is exhibited by the field distribution – separate spots (speckles) become shifted and deformed. This effect is governed by changes in the condition of total internal reflection of multimode radiation propagating in the fiber core. As a result, phase and geometrical parameters of the output radiation are also varying. High sensitivity of the speckle pattern at the output of a fiber to its mechanical or thermal deformations is effectively used in fiber-optical detectors, though interpretation of interference patterns for many modes is a rather complex problem. It should be noted that contrast of the speckle structure on coherent mode excitation is fairly high and,

generally, the speckle pattern is depolarized. The speckle number is approximately equal to the number of propagating modes. The speckle-structure spot size decreases with an increase in the normalized intensity *V*, i. e., with the increased number of excited modes $N \approx 0{,}5V^2$. Considering that the beam cone at the output $2\Theta_{0C}$ is dictated by the numerical aperture of a fiber $NA = \sin\Theta_{0C}$ and a number of speckles corresponds to the number of propagating modes *N*, the angular speckle size is approximately given by the formula

$$\varphi_S \approx 2NA/\sqrt{N}. \tag{5.2.16}$$

From this formula, with due regard for expressions (5.2.14) and (5.2.15), the angular speckle size is easily found as follows:

$$\varphi_S \approx \lambda/2a, \tag{5.2.17}$$

where a – fiber core radius. It is important that speckles are chaotically positioned, differ in sizes, and may be overlapping. The derived formula (5.2.17), that is classical in a speckle theory, gives a size of a separate speckle for the structures arising on coherent illumination of randomly inhomogeneous objects.

Note that the effect of mode interference is most marked for short-length communication lines when using coherent laser sources in the case of imperfect joints of the fiber ends – the speckle structure at the output of one fiber is partly diaphragmed by the collective aperture of another fiber. In the process the energy loss is insignificant but at minor mechanical or thermal loads the speckle structure is considerably varying to cause fluctuations in the amount of the energy accepted by the second fiber. When the fiber length is great, intermode dispersion causes a relative delay between different modes that can exceed a coherence time. Disturbance of mutual coherence of the waveguide modes causes lowering of the spackle structure contrast or disappearance of the structure at the background of spatially homogeneous output radiation.

5.3. Systems for radiation coupling in optical fiber

5.3.1. Radiation propagation in fiber waveguides

Fig. 5.3.1 shows the longitudinal cross-section of a two-layer model for optical fiber. A light beam incident on the fiber end from the air with the refractive index $n_0 = 1$ and at the angle Θ_0 to the axis is refracted at the angle Θ. From Snell's law is follows that

$$\sin\Theta_0 = (n_1/n_0)\sin\Theta = n_1 \sin\Theta. \tag{5.3.1}$$

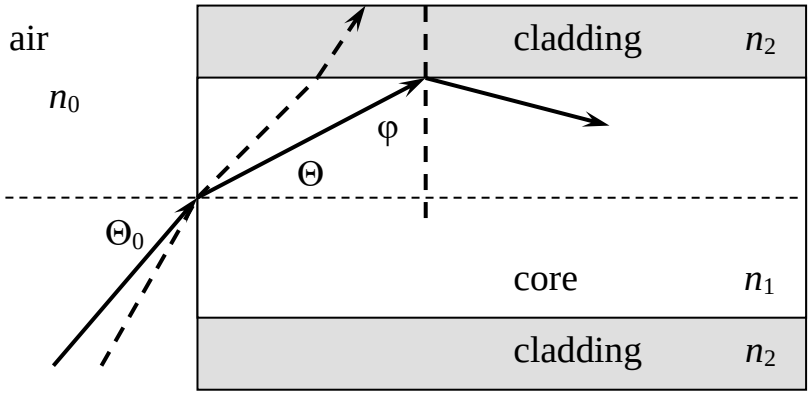


Fig. 5.3.1. Light propagation in a bilayer fiber

Until the angle Θ remains lower than the critical angle Θ_C of total internal reflection from the interface «core – cladding», determined by the relation

$$\cos\Theta_C = \sin\varphi_C = n_2/n_1, \tag{5.3.2}$$

the core confines this beam. As a result, all the beams incident on the face surface of the fiber cladding at the angle Θ_0, that is lower than Θ_{0C} for which we have

$$\sin\Theta_{0C} = \frac{\sqrt{n_1^2 - n_2^2}}{n_0} = \sqrt{n_1^2 - n_2^2}, \tag{5.3.3}$$

are confined within the cladding due to total internal reflection. The quantity $\sin\Theta_{0C}$ represents a numerical aperture (*NA*) of the fiber cladding, i. e. we have

$$NA = \sin\Theta_{0C} = \sqrt{n_1^2 - n_2^2}\,. \tag{5.3.4}$$

Numerical aperture designated as *NA* is a very important characteristic of a fiber. The fiber with a high value of *NA* has a high gathering power, whereas the fiber with a low value of *NA* accepts only a low-divergence beam.

Light propagation in optical waveguide is demonstrated for the so-called *meridional* beams passing through the central axis of a fiber (Fig. 5.3.1). Other beams, referred to as *asymmetric* beams, travel along the fiber beyond its central axis. Their path represents a spiral turning about the central axis. The same critical angle Θ_C, determined by expression (5.3.3), and the same numerical aperture, determined by expression (5.3.4), is valid not only for meridional but also for all the beams accepted by the core.

Apart from total internal reflection, the light beams captured by the waveguide core should also satisfy the positive interference condition that consists in the following: after two sequential re-reflections from the walls, the

corresponding waves should be in phase and interference-amplified on overlapping. Only when waves satisfy this phase condition, they form the self-consistent distribution of a field of the guided waves (waveguide modes). When this condition is not satisfied, the waves are interfering to quench themselves and disappear. Thus, out of a continuum of light beams, within the total reflection angle the guided modes are formed only by limited numbers of the beams with discrete angles.

5.3.2. Radiation coupling in fiber waveguides.

Optical radiation coupling devices in fiber waveguides should provide input into fiber for the greatest part of incident radiation from a light source. The efficiency of radiation injection into a fiber is significantly dependent on the characteristics of the waveguide itself: *numerical aperture* and refractive index distribution profile in the transverse cross-section of the fiber. The numerical aperture *NA* (5.3.4) determines the beam cone that may be accepted by the fiber. Note that the input geometry is often characterized by the *beam capture (acceptance) angle* $= 2\Theta_{0C}$. A cone of the beams accepted by an optical fiber is

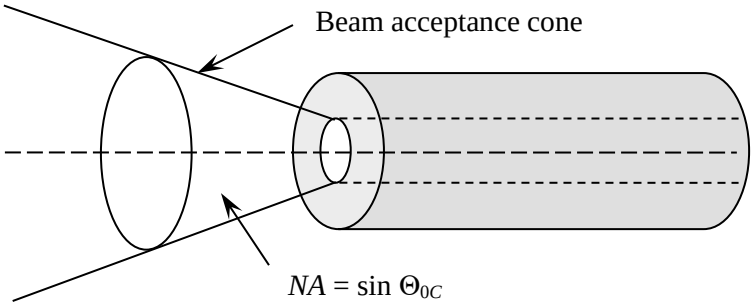


Fig. 5.3.2. Cone of beams accepted by optical fiber

given schematically in Fig. 5.3.2.

To determine a part of the power from a light source (laser, light-emitting diode, and the like) that may be injected into an optical fiber, we use the excitation efficiency parameter. In the case of light sources close positioned to the waveguide end, this parameter is reduced to the following form:

$$\eta = W_2 / W_1 = R\left(S/S_0\right) NA^2 = R\left(a/a_0\right)^2 NA^2, \qquad (5.3.5)$$

where W_1 – total power radiated by a source; W_2 – total power injected into a fiber; $S_0 = \pi a_0^2$ – radiating surface area of the source (cross-section of the light

beam incident on the waveguide); $S = \pi a^2$ – area of the fiber core; R – coefficient including the loss by Fresnel reflection from the fiber end. It should be noted that expression (5.3.5) is obtained only for *meridional* beams. But, as noted above, in fibers many beams have spiral propagation paths. The total coupling factor with regard to *asymmetric* beams is growing by the value dependent on the fiber type.

As seen from formula (5.3.5), to attain maximal efficiencies of radiation coupling into optical fiber, the geometrical sizes of a radiation source and of the fiber core diameter must be matched. When the source size is considerably in excess of the core diameter, the greatest part of light is lost (Fig. 5.3.3.)

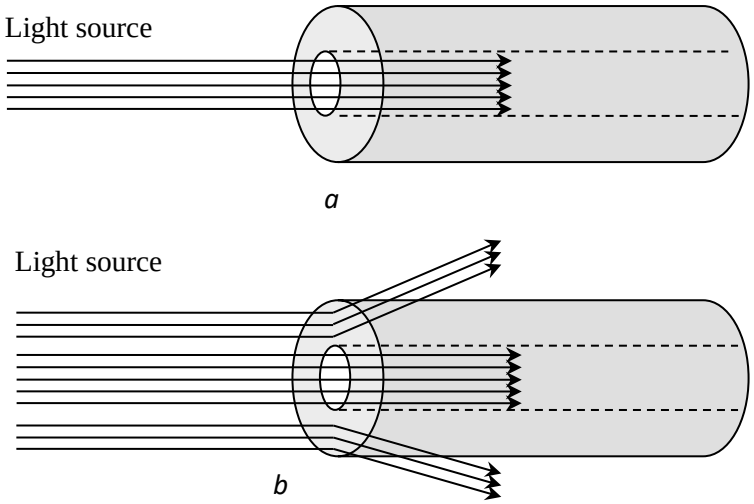


Fig. 5.3.3. Efficiency of radiation coupling into optical fiber
a – light source is well matched with the fiber;
b – light source is not matched with the fiber

Actually, the area of the radiating surface of a light source is much greater than the area of the waveguide end. Because of this, the coupling efficiency may be improved by the use of lenses, embedded dielectric cones, and other additional optical elements or by means of spherically melted fiber ends.

A simple system projecting an image of the light-source surface on the fiber end consists of a lens positioned between the source and the fiber. The position and the size of this lens should provide maximal radiation acceptance from the source. Spherical lenses increase the coupling factor due to the increased numerical aperture. As the coupling factor (5.3.5) is proportional to the squared numerical aperture, by means of a spherical lens we can greatly augment the power injected into the fiber. From the viewpoint of geometrical optics, radiation coupling is most effective with the use of long-focus lenses offering lower radiation divergence in the focal plane. However, the focused spot with the use of long-focus lenses in practice may be greater than the core diameter.

It should be noted that the coupling efficiency is maximal when the fiber end is positioned perpendicular to the source. When a fiber is tilted with respect to the source, on the one hand, its effective core area decreases (critical cut-off angle Θ_{0C} is lowered) and, on the other hand, Fresnel radiation loss by scattering from the fiber end increases. As a result, the coupling efficiency of radiation is significantly lowered.

5.3.3. Input of Gaussian light beams into waveguide

Laser generated radiation represents narrow beams with the transverse dimension much greater than the wavelength. Such light beams are well described by a theory of Gaussian beams, the amplitude of which in the cross-sectional plane is varying under the Gauss-Hermit law as $E \sim \exp\left[-r^2/w^2\right]$ (w – light beam radius), whereas the phase surface in the process of the beam propagation is curved due to light diffraction.

The intensity distribution in the cross-sectional direction of a Gaussian beam of the power P is given by the following formula:

$$I(r,z) = \frac{2P}{\pi w(z)^2} \exp\left[-2\frac{r^2}{w(z)^2}\right], \qquad (5.3.6)$$

and the light beam intensity at a distance from the axis $r = w$ comes to $1/e^2 \approx 13{,}5\,\%$ of the maximal value. When a light beam propagates in a free space, it is subjected to diffraction, retaining its form (i.e., still remaining the Gaussian beam) but changing its amplitude, radius, and the wavefront curvature.

A complex amplitude of the electric field strength for a monochromatic light beam with the wavelength λ, propagating along the axis z, is varying as follows:

$$E(r,z) = E_0 \frac{w_0}{w(z)} \exp\left(-\frac{r^2}{w(z)^2}\right) \exp\left(-i\left[kz - \arctan\frac{z}{z_R} + \frac{kr^2}{2R(z)}\right]\right), \qquad (5.3.7)$$

where E_0 – maximal amplitude of the light beam ($r = 0$, $z = 0$); w_0 – light beam radius $z = 0$ (waist radius); $k = 2\pi/\lambda$ – wave number; $z_R = \pi w_0^2/\lambda$ – Rayleigh length determining the distance for which a Gaussian beam is propagating without considerable divergence; $R(z)$ – wavefront curvature radius. The electric-field strength amplitude oscillating in time we derive multiplying the expression for the

spatial component (5.3.7) by the temporal component $\exp(i\omega t)$, where $\omega = 2\pi c/\lambda$ – light-wave circular frequency.

Assuming that at the point $z = 0$ a beam has a flat phase front and its radius is w_0, the Gaussian beam radius varies depending on the distance as

$$w(z) = w_0\sqrt{\left(1+\left(z/z_R\right)^2\right)}. \qquad (5.3.8)$$

In this case the beam wavefront curvature radius is described by the expression

$$R(z) = z\sqrt{\left(1+\left(z_R/z\right)^2\right)}. \qquad (5.3.9)$$

Another very important parameter of a Gaussian beam is the far-field divergence determined by the formula

$$\psi = \lim_{z\to\infty}\left(2w(z)/z\right) = 2\lambda/\pi w_0, \qquad (5.3.10)$$

demonstrating that divergence of a light beam is the greater the longer the radiation wavelength or/and the lower the beam radius.

To excite a particular stable structure of the light beam (mode) in a fiber waveguide, it is required to match the spatial structure of the incident Gaussian beam and the fiber mode. Fig. 5.3.4 shows such a lens transformer. When a minimal spot radius of one of the beams is w_0 and of the other – w_1, a lens with the focal length

$$f \geq f_0 = \pi w_0 w_1/\lambda \qquad (5.3.11)$$

transforms the field distribution in one beam to the corresponding field distribution in the other beam, and the cross-sections with minimal beam sizes are positioned at the distances

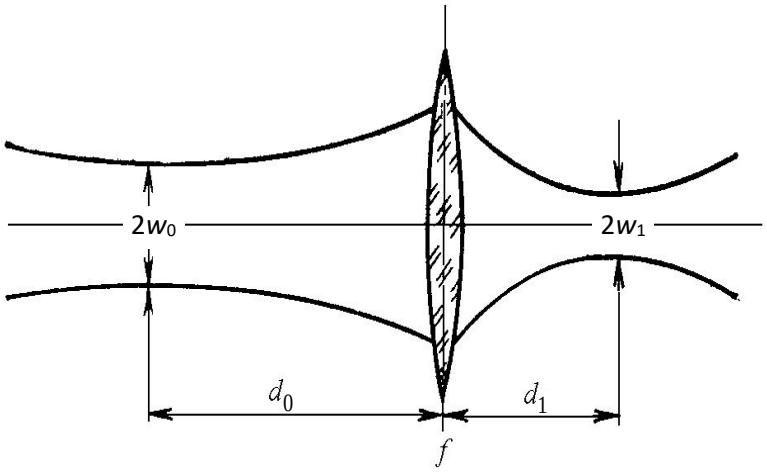


Fig. 5.3.4. Transformation of the Gaussian beam mode with the help of a matching lens.

$$d_0 = f \pm (w_0/w_1)\left(f^2 - f_0^2\right)^{1/2};$$
$$d_1 = f \pm (w_1/w_0)\left(f^2 - f_0^2\right)^{1/2} \qquad (5.3.12)$$

from the central plane of the matching lens. A lens with the minimal focal length f_0, according to relation (5.3.11), should be positioned exactly at the center between minimal cross-sections of these two beams.

The highest efficiency of radiation excitation in a fiber is attained for the approximately equal sizes of the focused beam waist and of the fiber core $w_1 = a$. At the same time, a size of the waist we can determined from formulae (5.3.10), (5.3.11), given the divergence of laser radiation $w_1 = f_0\psi/2$. In this way, based on the divergence of incident radiation and the fiber core size, one can calculate an optimal focal length of a lens.

Fig. 5.3.5 shows a scheme for coupling in optical fiber of the collimated light beam with the help of a focusing lens. Note that light at the output of the fiber core represents a divergent beam with the divergence corresponding to the numerical aperture. As the input and output parameters of the fiber are identical, the exit aperture is the same as the entrance one.

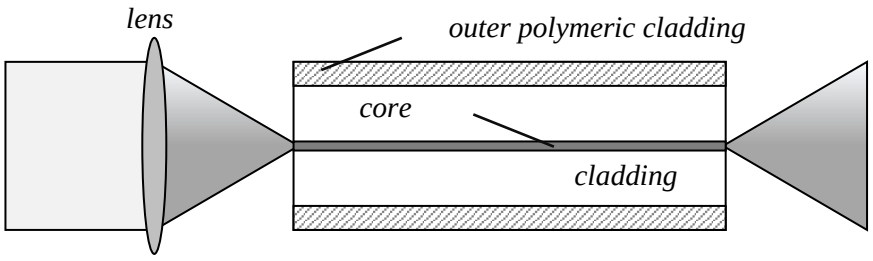


Fig. 5.3.5. Simple scheme of optical radiation coupling in fiber

5.4. Fiber-optical data transmission systems

5.4.1. General characteristics of fiber-optical data transmission systems

Fiber-optical communication lines (FOCL) are data transmission systems based on dielectric optical waveguides (optical fibers). Apart from fiber optics, FOCL technologies envelope the electronic transmitting and receiving equipment, signal standardization, and network topologies. Being an important component of FOCL, optical fibers offer much promise as a medium for transmission of large

data flows to long distances. Among the advantages of optical fibers, we can name the following:

- *broad band* due to an extremely high carrier frequency ($\nu \sim 10^{14}$ Hz) at a low noise level in a fiber-optical cable makes it possible to increase the transmission band by the use modulated signals with low code redundancy. Owing to FOCL, information is transmitted at the rate of about a few terabit per second;
- *low optical-signal attenuation in a fiber* (0.2–0.3 dB/km at the wavelength 1.55 μm) allows for the development of FOCL 100 km long and more without intermediate signal amplification;
- *high noise immunity* due to the use of the electromagnetic-interference immune dielectric material for the production of fibers;
- *protection against unauthorized access* because the data transmitted by fiber-optical communication lines are hardly intercepted without destructive procedures;
- *electrical safety* –fibers contribute to fire and explosion safety of networks (no sparking) that is of particular importance for chemical and petrochemical plants, for the technological processes associated with enhanced risk;
- *low weight and small dimensions* – fiber-optical cables have lower weight and smaller dimensions as compared with copper cables having the same transmission capacity. For example, 900-pair telephone cable with a diameter of 7.5 cm may be replaced by one fiber 0.1 cm in diameter. Even covered with protective cladding, tape armor including, the fiber has a diameter of 1.5 cm;
- *low cost* as fiber is manufactured of quartz that is based on silicon dioxide, widespread and inexpensive material compared to copper;
- *durability* as service life of FOCL is no less than 25 years.

Disadvantages of FOCL are mainly associated with rather expensive cross equipment, as electric signals should be transformed rapidly to optical signals and vice versa, and with a necessity for servicing the optical lines which involve intermediate amplifiers, multiplexers, and signal repeaters. However, these disadvantages are insignificant compared with the above-mentioned advantages of optical fibers – their applications in communication systems are ever widening.

Any FOCS includes a number of the required components:

- *optical transmitter* provides transformation of the input electrical (digital or analog) signal to the output light (digital or analog) signal;
- *optical detector* offers inverse transformation of the input optical pulses to the output electrical-current pulses;

- *multiplexer/demultiplexer* is used for combination or separation of information channels. Multiplexers and demultiplexers are operable both in temporal and frequency regions, may be electrical and optical;
- *regenerator* reconstructs the form of the optical pulse distorted in the process of propagation in a fiber. Regenerators may be both optical and opto-electrical, realizing transformation of an optical signal to the electrical signal, its restoration, and again transformation to the optical signal;
- *amplifier* offers amplification of the signal power. Amplifiers, optical and electrical, realize optoelectronic and electron-optical transformations of a signal;
- *fiber-optical cable* with optical fibers as light-bearing elements;
- *optical connectors and couplers* enable commutation of optical channels.

Optical fibers used for signal transmission by FOCL are characterized by the two very important parameters: attenuation and dispersion.

5.4.2. Dispersion

Dispersion – time mismatch between the spectral and mode components of an optical signal. We transmit by optical fiber not only the optical energy but also the information signal. Light pulses, a series of which determines the information flow, are spreading in the process of propagation. When broadening is considerably large, the neighboring pulses are overlapping to make impossible their separation on reception.

Dispersion is defined as the squared difference of the pulse lengths at the output and at the input of a waveguide

$$\tau(L) = \sqrt{t_{out}^2 - t_{in}^2}. \qquad (5.4.1)$$

As a rule, dispersion is normalized on a 1 km basis and is measured in ps/km. As pulse broadening is proportional to the communication line length, dispersion limits a length of the regeneration section, at which a signal is propagating in a fiber without any systems of transformation.

On the other hand, pulse broadening determines the transmission band of the transmitted signal ΔF

$$\Delta F \approx \tau^{-1}. \qquad (5.4.2)$$

Transmission band represents a measure of the fiber ability to transmit a certain information content per unit time.

The lower dispersion, the broader the transmission band and the higher information flow is transmitted by the fiber. Many manufacturers in their specifications indicate the transmission band ΔF in terms of MHz/km. The

indicated quantity gives the product of the transmission band by the fiber length used – we can transmit a low-frequency signal (low information content) over a long

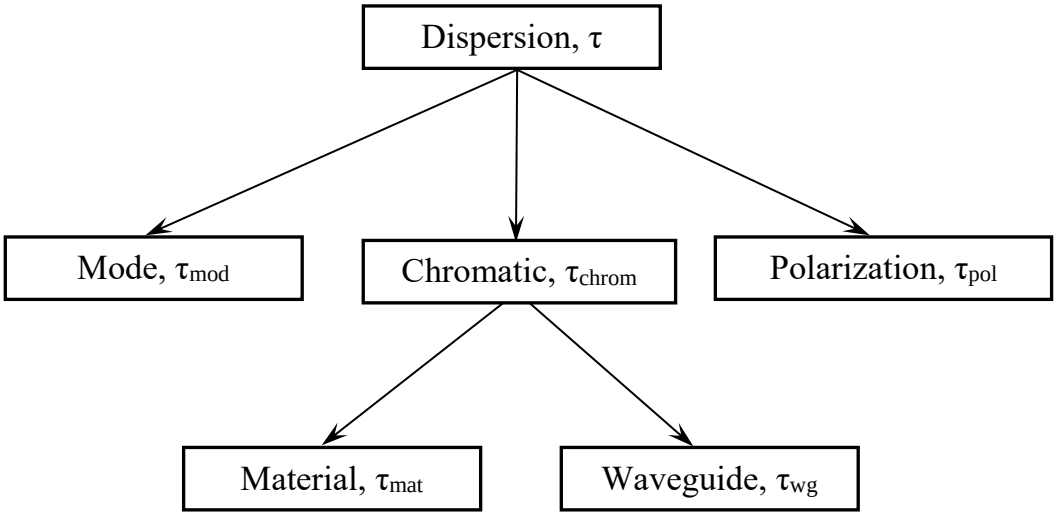


Fig. 5.4.1 Dispersion types

distance or a higher-frequency signal – to a shorter distance.

The following dispersion types are distinguished for optical waveguides: mode (intermode) dispersion, polarization, and chromatic dispersion subdivided into the material and waveguide types (Fig. 5.4.1).

Mode (intermode) dispersion exists only in a multimode fiber, occurring due to different propagation rates of the beams (with differing modes) which form a light pulse. Different modes arrive to the fiber exit at different times, leading to the pulse broadening. As seen in Fig. 5.4.2, *a*, modes of different orders are propagating at different angles with respect to the waveguide axis, covering different distances at the same time. Obviously, this results in pulse spreading in the process of its propagation along the fiber.

Mode dispersion may be lowered due to a decrease of the core diameter *d*, suppression of the higher-order modes or due to the use of a graded-index fiber with the refractive index smoothly increasing from the core edges to its center (Fig. 5.4.2, *b*). In this case the propagation conditions for the axial and aperture beams become different. The geometrical path of an aperture beam is longer than that of the axial one, though its optical path length is not very different as propagation of light in the off-axial regions with a lower refractive index is faster than along the axis. This is the way to adjust transmission times for different beams in an optical fiber. Decreasing the beam diameter and suppressing the

higher-order modes, one can go to the single-mode regime that, evidently, is not associated with mode dispersion (Fig. 5.4.2, *c*).

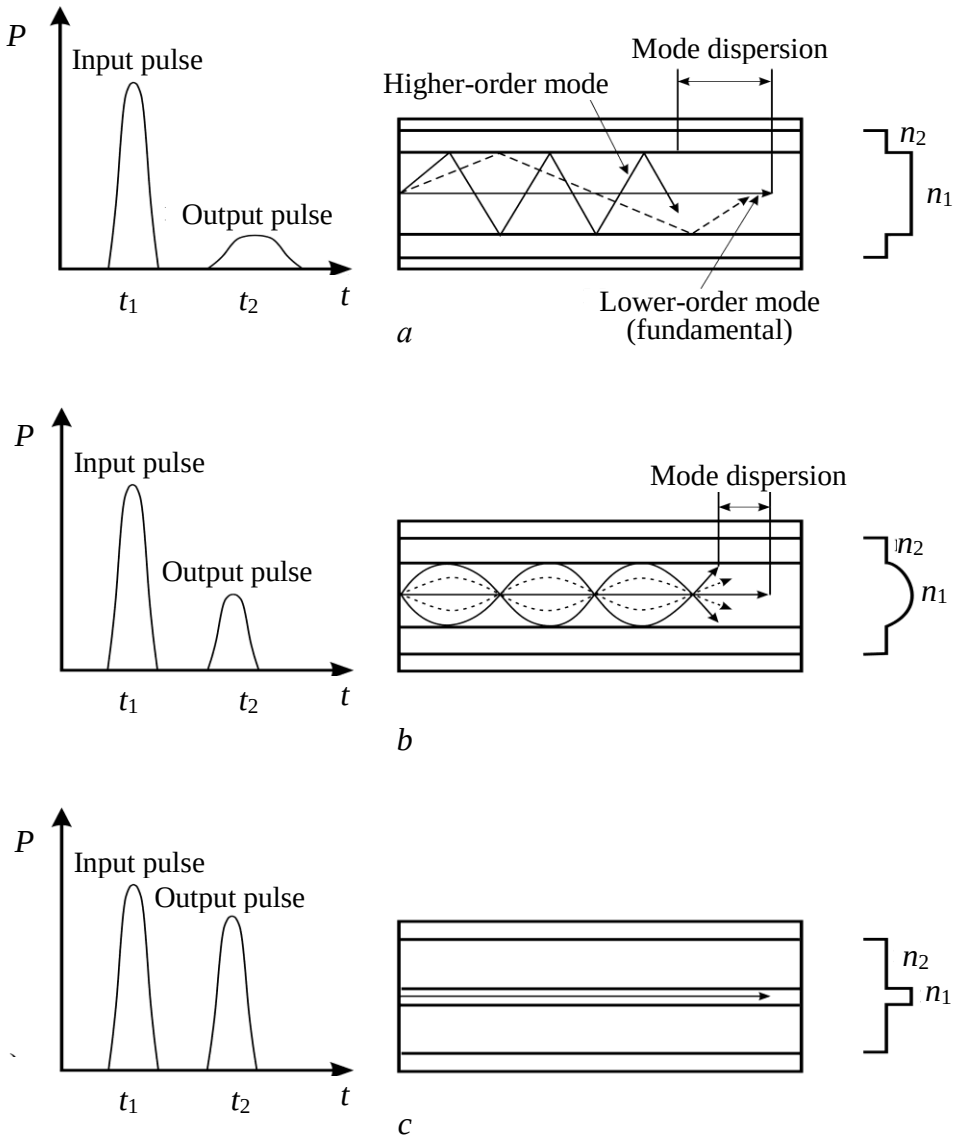


Fig. 5.4.2. Character of light propagation in optical fiber with different refractive index profiles:
a – step-index multimode fiber; *b* – graded-index multimode fiber;
c – step-index single-mode fiber

Dispersion in the case of a step-index multimode or graded-index multimode fiber with a parabolic profile of the refractive index may be calculated as follows:

$$\tau_{\text{mod step}}(L)\cdot L=\begin{cases}\dfrac{n_1\Delta}{c}L, & L<L_{\text{c}},\\[2ex] \dfrac{n_1\Delta}{c}\sqrt{L\cdot L_{\text{c}}}, & L>L_{\text{c}};\end{cases} \qquad (5.4.3a)$$

$$\tau_{\text{mod grad}}(L)\cdot L=\begin{cases}\dfrac{n_1\Delta^2}{2c}L, & L<L_c,\\ \dfrac{n_1\Delta^2}{2c}\sqrt{L\cdot L_c}, & L>L_c,\end{cases} \qquad (5.4.3b)$$

where L – fiber length; n_1 – refractive index of the core; $\Delta=(n_1^2-n_2^2)/2n_1^2$ – relative difference in refractive indies of the core and of the cladding; c – speed of light in free space; L_c – intermode coupling length (for the step-index fiber $L_c \sim 5$ km, for the graded-index fiber ~ 10 km).

Variations in the dispersion law with the increased fiber length are due to inhomogeneities inherent in a real fiber. These inhomogeneities lead to the intermode interactions and to the energy redistribution within the modes. When $L > L_c$, the steady-state regime is set and in radiation all the modes are present in a certain proportion. Generally, with the use of multimode fibers, lengths of communication lines between the active devices are not longer than 2 km that is considerably below the intermode coupling length. Because of this, one can use for these fibers a linear dependence of dispersion on the fiber length.

In formulae (5.4.3) for step-index and graded-index fibers note the difference in dependences of dispersion on the relative difference in refractive indices of the core and of the cladding Δ, that is commonly on the order of 0.003. Due to quadratic dependence of the intermode dispersion of a graded-index fiber on the difference in the refractive indices Δ, dispersion of the graded-index fiber is significantly lower than that of the step-index fiber, making it more preferable in communication lines.

Chromatic dispersion includes the material and the waveguide components and is associated with propagation both in single- and multimode fibers. However, most marked this dispersion is in a single-mode fiber due to the absence of intermode dispersion.

Material dispersion results from the dependence of the refractive index of a fiber on the fiber length. Material dispersion is determined by the electromagnetic interaction of a wave with bound electrons of the medium material that, as a rule, is nonlinear in character. Also, the expression for dispersion of a single-mode fiber includes the differential dependence of the refractive index on the wavelength

$$\tau_{\text{mat}}\left(\Delta\lambda, L\right)=\Delta\lambda\cdot L\cdot\frac{\lambda}{c}\left[\Gamma\frac{d^2n_1}{d\lambda^2}+\left(1-\Gamma\right)\frac{d^2n_2}{d\lambda^2}\right], \qquad (5.4.4)$$

where Γ – factor of the waveguide mode localization in the fiber core determined

by the following relation:

$$\Gamma = \frac{P_1}{P_1 + P_2} = 1 - \frac{u^2}{V^2}\left[1 - \chi_l(v)\right], \tag{5.4.5}$$

where $\chi_l(v) = \dfrac{K_l^2(v)}{K_{l-1}(v)K_{l+1}(v)}$; P_1 and P_2 – waveguide mode power in the core and in the cladding, respectively; $V^2 = u^2 + v^2 = \left(2\pi a/\lambda\right)\sqrt{n_1^2 - n_2^2}$ – normalized frequency; a – fiber core radius; u and v – normalized transverse wave numbers of a mode in the core and in the cladding as found from the wave equation

$$u\frac{J_m(u)}{J_{m\pm1}(u)} = \pm v\frac{K_m(v)}{K_{m\pm1}(v)}, \tag{5.4.6}$$

where $J_j(u)$ and $K_j(v)$ – j-order Bessel and Macdonald functions, respectively.

Since the main part of the mode power is accumulated in the core of a fiber, the expression for material dispersion is approximately given as

$$\tau_{\text{mat}}\left(\Delta\lambda, L\right) = \Delta\lambda \cdot L \cdot \frac{\lambda}{c} \cdot \frac{d^2 n_1}{d\lambda^2} = \Delta\lambda \cdot L \cdot M\left(\lambda\right), \tag{5.4.7}$$

where $\Delta\lambda$ – spectral width for a light source; λ – radiation wavelength; $M\left(\lambda\right) = \dfrac{\lambda}{c} \cdot \dfrac{d^2 n_1}{d\lambda^2}$ – specific material dispersion that is ordinary measured in ps/(nm · km).

In Fig. 5.4.3, showing the spectral dependence of a specific material dispersion for quartz glass, it is seen that over the wavelength range 1000–1600 nm the material dispersion $M(\lambda)$ almost linearly decreases from +60 to –40 ps/(nm · km), going to zero at the wavelength 1270 nm. The wavelength associated with zero value of the specific material dispersion $M(\lambda)$ is referred to as zero-dispersion wavelength λ_{0D} for a bulk medium. In the region of wavelengths above 1270 nm material dispersion is negative, i.e., the refractive index is growing with the wavelength. The inverse situation is observed in the spectral region $\lambda < 1270$ nm, where dispersion is positive.

Material dispersion is a major type of dispersion in single-mode fibers. On the contrary, in multimode fibers material dispersion is neglected because mode dispersion in these fibers is most substantial.

Waveguide dispersion. Dispersion of real waveguides is distinguished from that of a bulk medium by the presence of the waveguide structure varying the effective refractive index of the mode. As a result, a special waveguide component of dispersion arises. Most profound this effect is exhibited in single-mode fibers.

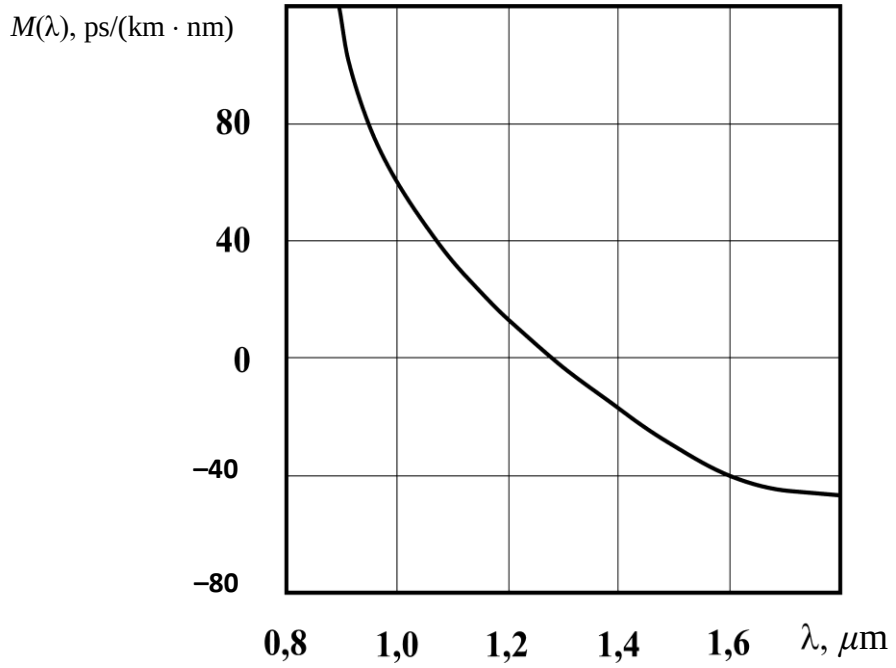


Fig. 5.4.3. Specific material dispersion as a function of the wavelength

The expression for waveguide dispersion includes refractive indices of the core n_1 and of the cladding n_2 as well as the cross-sectional wave numbers of the mode in the core u and in the cladding v determining the normalized frequency $V = \sqrt{u^2 + v^2}$:

$$\tau_{\text{wg}}\left(\Delta\lambda, L\right) = \Delta\lambda \cdot L \cdot \frac{2n_1\Delta}{c\lambda}\left(\frac{m_1}{n_1}\right)^2 V\frac{d^2(BV)}{dV^2} = \Delta\lambda \cdot L \cdot N(\lambda), \qquad (5.4.8)$$

where $N\left(\lambda\right) = \frac{2n_1\Delta}{c\lambda}\left(\frac{m_1}{n_1}\right)^2 V\frac{d^2(BV)}{dV^2}$ – specific waveguide dispersion; $\Delta = (n_1^2 - n_2^2)/2n_1^2$ – relative difference in refractive indices of the core and of the cladding; $m_1 = n_1 - \lambda\frac{dn_1}{d\lambda}$ – group refractive index; $B = (v/V)^2$.

The spectral dependence of specific waveguide dispersion of a single-mode quartz fiber is demonstrated in Fig. 5.4.4. As seen, the specific waveguide dispersion $N(\lambda)$ is always over zero.

Considering similar forms of formulae (5.4.7) and (5.4.8), the resultant coefficient of specific chromatic dispersion is introduced that is determined as $D(\lambda) = M(\lambda) + N(\lambda)$. Due to the inclusion of waveguide dispersion, the zero-dispersion wavelength is shifted from 1270 nm (for a bulk medium) to

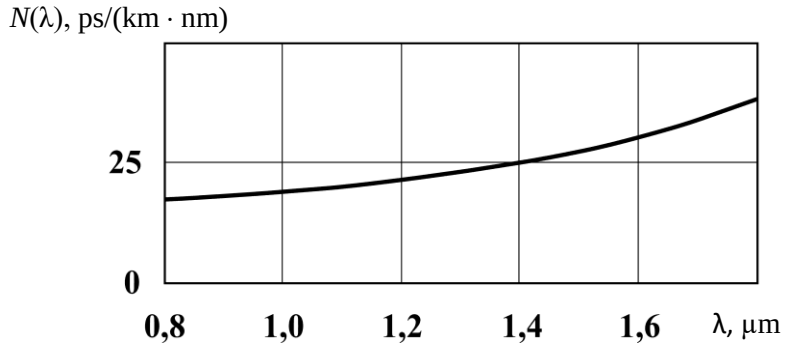


Fig. 5.4.4. Specific waveguide dispersion of a quartz fiber as a function of the wavelength

1310 ± 10 nm (for a single-mode step-index fiber). The fact that the resultant dispersion $D(\lambda)$ is going to zero at this wavelength explains a wide use of radiation sources with the wavelength 1310 nm in fiber-optical communication systems.

Chromatic dispersion is connected to specific chromatic dispersion by the simple relation $\tau_{chrom}(\lambda) = D(\lambda) \cdot \Delta\lambda$. Chromatic dispersion may be reduced by the use of more coherent radiation sources (e.g., laser transmitters, $\Delta\lambda \approx 2$ nm) and of the working wavelength that is closer to that of zero dispersion.

Dispersion-shifted fibers. From the viewpoint of dispersion parameters, the available single-mode fibers are subdivided into three main types: standard step-index (dispersion-nonshifted) fibers (*SF*), dispersion-shifted fibers (*DSF*), and nonzero dispersion-shifted fibers (*NZDSF*).

All three types of fibers are very similar in attenuation over the most common ranges 1310 and 1550 nm but they differ in the chromatic dispersion characteristics. *SF* is optimized for dispersion in the region 1310 nm. *DSF* is fully optimized for operation in the region 1550 nm, both for attenuation and dispersion. However, when the zero-dispersion wavelength (1550 нм) is within the working range of an erbium amplifier, the developing nonlinear effects (first of all, four-wave mixing) lead to an abrupt increase of noise during the propagation process of a multichannel signal. Fibers of the type *NZDSF* have been created to overcome the disadvantages of *DSF* exhibited during operation with multiplex optical signals. A wavelength of zero dispersion is removed beyond the working range of erbium amplifiers in use to lower nonlinear effects during transmission of multiplex signals.

Polarization mode dispersion τ_{pmode} occurs due to different propagation rates of the two polarization mode components. The specific dispersion coefficient T is normalized on a 1 km basis with the units (ps/km$^{1/2}$), the dispersion τ_{pmode} growing

with the distance as $\tau_{pmode} = T \cdot L^{1/2}$. Due to its low value, polarization dispersion is developing only in a single-mode fiber when the broad-band signal with a very narrow spectral band (0.1 nm and lower) is transmitted (transmission band 2.4 Gbit/s and higher). In this case chromatic dispersion becomes comparable to polarization mode dispersion.

The major reason for the development of polarization mode dispersion is nonroundness of the core profile in a single-mode fiber associated with manufacturing imperfections during the fiber pulling or cable production process and also with dynamic deformations of fibers under mechanical load or temperature variations. An electrical field of a light wave is represented as superposition of two polarization states (two linear polarizations with orthogonal vectors or two circular polarizations with opposite senses of rotation). In the ideal isotropic fibers with such a partition both components are propagating at the same rate, the resultant pulse length on passage through a medium remains the same as at the fiber input. In a fiber with the anisotropic profile the two differing effective refractive indices are associated with two orthogonal linear polarizations. This results in different group propagation rates of polarization modes and in time delay of signals at the fiber output. As in conventional FOCL polarization is not distinguished by photodetectors, the resultant signal is broadened. Such broadening may be lowered in fibers of large length as polarization axes occur randomly, varying their orientation along the fiber length. As a result, a fast mode goes to the slow one, and vice versa. This mode coupling leads to equalization of their propagation times and to lowering of polarization mode dispersion. Also, it should be noted that in fibers with marked birefringence a weak coupling of polarization modes allows for their use to lower the beats due to the energy transfer from one mode to another.

Proceeding from the above, the resultant dispersion of a fiber takes the following form:

$$\tau^2 = \tau_{mod}^2 + \tau_{chrom}^2 + \tau_{pol}^2 = \tau_{mod}^2 + \left(\tau_{mat} + \tau_{wg}\right)^2 + \tau_{pol}^2 . \qquad (5.4.9)$$

In ordinary working conditions of a single-mode fiber polarization dispersion is sufficiently low and may be neglected in calculations of the total dispersion. Mode dispersion is also absent because in such fibers there is propagation of only one HE_{11} mode or, as noted above, of two modes with two different polarization states. Consequently, pulse broadening in a single-mode fiber is actually determined only by chromatic dispersion

$$\tau = \tau_{chrom} = \tau_{wg} + \tau_{mat} . \qquad (5.4.10)$$

As regards a multimode fiber with a large diameter of the core, its polarization as well as waveguide dispersion is low and hence the total dispersion is represented by the following expression:

$$\tau = \sqrt{\tau_{\text{mod}}^2 + \tau_{\text{mat}}^2} \qquad (5.4.11)$$

In a multimode fiber with the step-index refractive-index profile the mode dispersion τ_{mod} dominates over the material dispersion τ_{mat}; the reverse situation is the case in a graded-index fiber. This is explained by the fact that in a multimode graded-index optical fiber τ_{mod} decreases due equalization of propagation times for different modes.

Comparing dispersion characteristics of different fibers, we can note that single-mode fibers offer the best performance; dispersion is most marked in multimode fibers with a step-index profile of the refractive index.

5.4.3. Radiation attenuation in FOCL

One of the general properties of optical waveguides is attenuation of their guided modes. The guided modes lose the power for absorption in a material and scattering from its inhomogeneities. Signal attenuation in fiber waveguides is governed by the mechanisms associated with (*a*) the material properties; (b) waveguide geometry; (c) presence of fiber junctions; (d) availability of the cladding and fiber jacket; (e) outgoing modes.

To determine attenuation in optical fibers quantitatively, we use the integrated fiber attenuation quantity including all the types of loss and given by the formula

$$\alpha = 10 \lg \frac{P_{in}}{P_{out}} \ \text{дБ}, \qquad (5.4.12)$$

where P_{in} – optical power arriving to the fiber input; P_{out} – power output from the fiber.

Besides, we can use the attenuation coefficient for a fiber that is given by the relation

$$\alpha_L = \frac{1}{L} 10 \lg \frac{P_{in}}{P_{out}} \ \text{dB/km}, \qquad (5.4.13)$$

where L – waveguide length in kilometers.

Radiation attenuation due to material properties.

The materials used in waveguides are transparent in the ranges, for which the effective light sources or sensitive photo detectors are available, the loss in fiber itself being at minimum. At the present time three major ranges corresponding to the lowest attenuation are used. The so-called first, second, and third transparency windows are associated with the central wavelengths 0.85, 1.3, and 1.55 μm. The spectral dependence of the loss by absorption in a material is shown in Fig. 5.4.5.

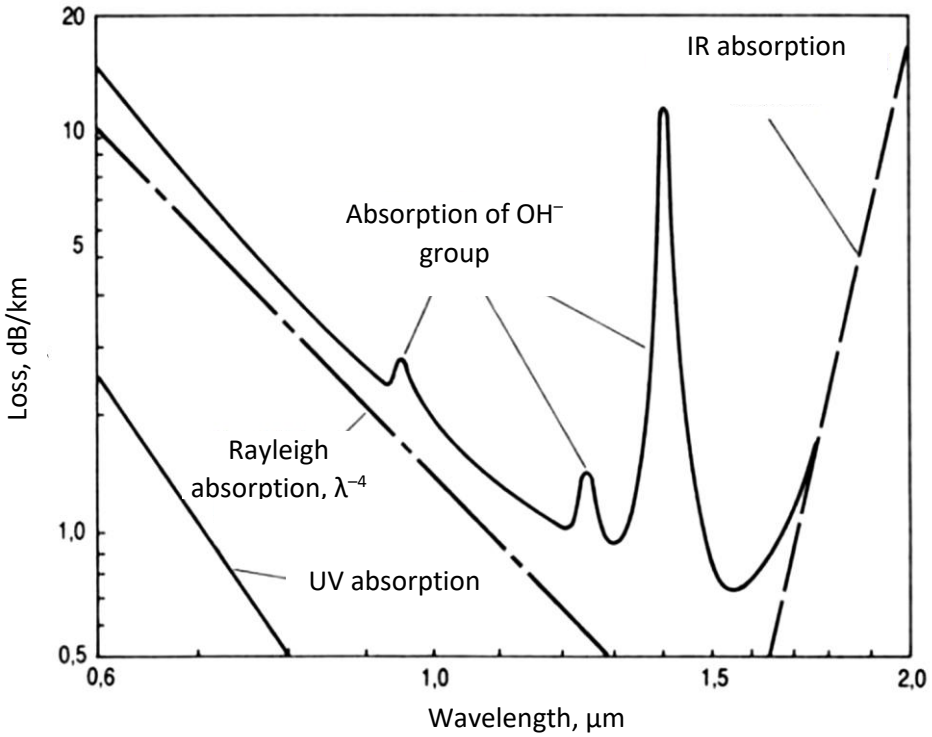


Fig. 5.4.5. Spectral dependence of the material loss by absorption in the case of germanium-silicate single-mode optical fiber

The main cause of the power loss in a fiber is absorption and scattering of the energy. Attenuation of absorption is determined by intrinsic absorption of the optical fiber material and absorption by impurities and atomic defects. Impurities of metal ions (*Fe*, *Cu*, *V*, *Cr*) and of the hydroxyl group OH^- lead to a drastic increase of attenuation in some regions of the spectral range. They are responsible for absorption maxima at the wavelengths 0.95, 1.25, and 1.39 μm.

To guarantee low losses, the manufacturers of fibers should maintain the concentration of these ions at a level of one part per billion. The atomic defects and absorption induced by them arise under the effect of high-intensity radiation or due to thermal treatment of a waveguide. Intrinsic absorption is observed in the

ultraviolet and infrared spectral regions. As seen in Fig. 5.4.5, intrinsic absorption of glass is actually determining the loss at the wavelengths above 1.7 μm.

Another cause of the loss is radiation scattering in a fiber. Attenuation by scattering may be of two types: linear scattering and nonlinear scattering. Linear scattering includes the intrinsic (Rayleigh) scattering and Mie scattering. Rayleigh scattering is due to the density fluctuations, low as compared with the wavelength, inevitable in the process of glass production. As inhomogeneities in their magnitude are much lower than the wavelength, the loss by Rayleigh scattering is growing with a decrease in the wavelength in proportion to $1/\lambda^4$, being independent of the light intensity, as seen from the following expression for the attenuation coefficient:

$$\alpha_p \approx \frac{8\pi^2}{3\lambda^4}\left(n^2-1\right)KT\chi, \tag{5.4.14}$$

where K – Boltzmann constant; T – absolute temperature; n – refractive index; χ – compressibility. For the typical parameters of glass, the loss due to Rayleigh scattering comes to 2 dB/km for 850 nm, 0.4 dB/km for 1300 nm, and 0.dB/km for 1550 nm.

Mie scattering is due to scattering from inhomogeneities having the sizes comparable with the light wavelength. The fabrication methods of glass are so perfect that such defects can be avoided, in general the resultant loss in the fiber is due to Rayleigh scattering.

Nonlinear scattering includes the stimulated Raman scattering and the stimulated Brillouin scattering exhibited at high power levels in the form of radiation with shifted wavelengths and of radiation propagating in the counter direction. When signals are transmitted in a single-mode fiber to long distances, these phenomena determine the upper limit for the transmitted power.

Comparison of the contributions to losses by absorption and by scattering demonstrates that the main contribution into the loss in the process of IR radiation propagation in a fiber is made by intrinsic absorption of the material itself and by absorption on the impurities and atomic defects. Attenuation of the typical single-mode fiber with the refractive index structure 50/125 (core diameter 50 μm, cladding – 125 μm) is 3 dB/km for 850 nm and 1 dB/km for 1300 nm. In the third transparency window (1.55 μm) the loss of single-mode fibers comes to 0.7 dB/km.

Currently, there is a tendency for the development of even more “transparent”, fluozirconate, fibers with the theoretical limit on the order of 0.02 dB/km at the wavelength 2.5 μm. As demonstrated by the laboratory studies,

these fibers can form the basis for the creation of communication lines offering data transmission at the rate about 1 Gbit/s to the distances up to 4600 km without the use of intermediate amplification systems.

Typical values of the considered characteristics of fibers used in FOCL are given in the Table 5.4.1.

Table 5.4.1. – Typical characteristics of cabled optical fibers

Single mode fibre	**10/125 µm**	**SM / OS2 (ITU-T G.652.D)**
Mode field diameter (MFD)	1310 nm	9,3 +/- 0,5 µm
Mode field eccentricity -Installation cables		≤ 1,0 µm ≤ 0,5 µm
Cladding diameter - Installation cables		125 +/- 2 µm 125 +/- 1 µm
Cladding ellipticity		≤ 2 %
Fibre attenuation	1310 nm 1550 nm	≤ 0,40 dB/km ≤ 0,25 dB/km
Zero dispersion range		1300...1324 nm
Dispersion coefficient - Dispersion at	 1550 nm	≤ 0,093 ps/nm2/km ≤ 18 ps/nm/km
Cut-off wavelength - Installation cables		≤ 1260 nm 1180...1250 nm
Polarization mode dispersion		≤ 0,5 ps/√km

Multi mode fibre	**50/125 µm**	**OM3**
Core diameter		50 +/- 3 µm
Core ellipticity		≤ 6 %
Core eccentricity		≤ 3 µm
Cladding diameter		125 +/- 2 µm
Cladding ellipticity		≤ 2 %
Fibre attenuation	850 nm 1300 nm	≤ 2,7 dB/km ≤ 0,8 dB/km
Bandwidth	850 nm 1300 nm 850 nm	≥ 1500 MHz x km (LED) ≥ 500 MHz x km (LED) ≥ 2000 MHz x km (Laser)
Numerical aperture, NA		0,200 +/- 0,015

Loss associated with waveguide geometry.

In optical waveguides the loss due to their geometry violation is associated with the following factors: a) irregularities caused by the uneven interface «core – cladding», variations in sizes of the core section, and deviations of the fiber section from the ideal form of coaxial cylinders; b) fiber microbendings associated with the fiber jacket and cable fitting; c) fiber twisting and bending in the process of cabling (Fig. 5.4.6).

The loss by fiber bendings is associated with walk-off or emission of the guided modes. The frequency of walk-off modes is lower than the cut-off frequency – the loss is due to partial penetration into the fiber cladding and the

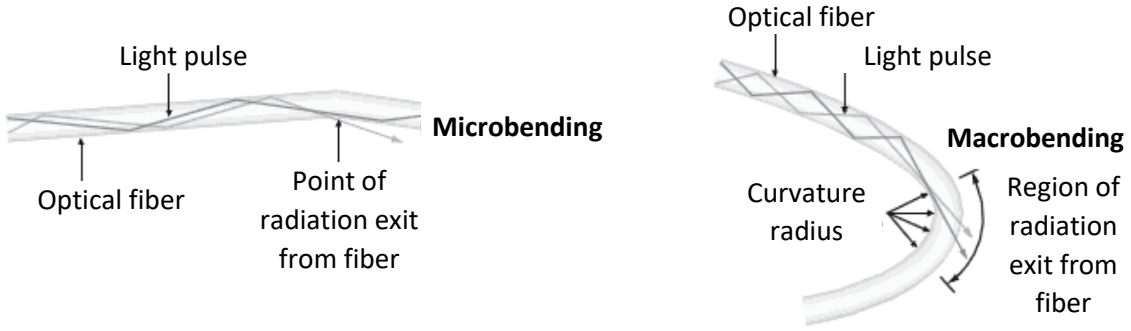


Fig. 5 4.6. Loss in fiber associated with micro- and macrobending

modes fail to propagate to long distances. In multimode fibers different groups of the modes are subjected to different attenuation as their fields penetrate the cladding to a variable degree. Statistical fluctuations of the properties along the fiber length lead to the energy exchange between the modes. In this way the energy may be transferred to the modes with a high attenuation coefficient, the walk-off modes including.

The relationship between the loss by bendings of a fiber and the waveguide parameters or bending radius is described by the exponential function; the loss becomes intolerable the moment when the bending radius lowers to the critical values. Attenuation of an optical signal as a function of the fiber bending radius R for the fixed fiber parameters and wavelength is given by the formula

$$\alpha_R = A\, R^{-1/2} \exp(-BR), \tag{5.4.15}$$

where $A = \dfrac{\pi^{1/2} u^2}{2a^{1/2}(u^2+v^2)v^{3/2}} K_1^{-2}(v)$ and $B = \dfrac{4v^3\Delta}{3a(u^2+v^2)}$; a – fiber core radius; $\Delta = (n_1^2 - n_2^2)/(2n_2^2)$ – quantity determined by the difference in refractive indices of the core n_1 and of the cladding n_2; $u = ka\sqrt{n_1^2 - n_{ef}^2}$, $v = ka\sqrt{n_{ef}^2 - n_2^2}$ – cross-sectional wave numbers of the fundamental mode in the core and in the cladding, respectively; n_{ef} – effective refractive index of the waveguide mode; $k = 2\pi/\lambda$ – wave number; $K_1(v)$ – first-order Macdonald function. The relation for the powers of an optical signal before (P) and after (P_R) bending is as follows:

$$P_R = P \exp(-\alpha_R z), \tag{5.4.16}$$

where z – bending arc length (for the full turn of a fiber $z = 2\pi R$).

Loss in optical connectors.

Apart from the loss of optical waveguides themselves, in fiber-optical lines there are some other types of losses. First of all, we should name the coupling loss and the loss at the junctions of separate fiber segments.

It is impossible to pullout a one-piece fiber from the source to the detector. Ordinary, the technological length of a fiber is a few kilometers. Even with welding of waveguides, mobility of the local optical network is provided only by means of cross equipment. To offer multiple and easy connection, the fibers are terminated by optical connectors. Considering micron sizes of the present-day fibers, termination of the fiber by connectors is a challenging job. Attenuation due to releasable connections is influenced by several factors which may be subdivided into three groups: a) internal factors caused by manufacturing defects of optical fibers; b) external factors associated with manufacturing defects of an optical connector itself; c) systematic factors involving the mode distribution in an optical fiber.

Internal factors.

When there is a need to connect two optical fibers, it is assumed that they are identical. Actually, this is not always the case. In the process of production some deviations of the geometrical parameters of optical fibers from the nominal ones are inevitable. Accuracy of the geometrical parameters is of great importance for optical fibers both in the case of releasable and welded connections. Let us consider the influence of nonidentity of optical fibers.

Loss associated with the aperture mismatch (*NA*). Numerical aperture (*NA*) – angle between the optical axis of an optical fiber and one of the generating lines of the light cone, within the bounds of which the light incident on the optical fiber end propagates according to the total internal reflection law. The numerical aperture dependent on refractive indices of the fiber core (n_1) and cladding (n_2) is calculated as

$$NA = (n_1^2 - n_2^2)^{1/2}. \qquad (5.4.17)$$

Loss in the power of an optical signal due to the difference in numerical apertures of the connected optical fibers occurs when a numerical aperture of the transmitting optical fiber is greater than that of the receiving fiber. This loss is calculated by the following formula:

$$\alpha_{NA} = 10 \lg (NA_T/NA_R)^2, \qquad (5.4.18)$$

where NA_T – numerical aperture of the transmitting fiber and NA_R – of the receiving one.

Difference in diameters of the cores of optical fibers. The loss is the case when the core diameter of the transmitting fiber is greater than that of the receiving fiber because some part of the optical power propagates in the cladding of the receiving optical fiber. This loss is determined from the formula

$$\alpha_d = 10 \lg (d_R / d_T)^2, \qquad 5.4.19)$$

where d_T – diameter of the transmitting fiber; d_R – of the receiving fiber.

Besides, sometimes the loss is due to the mismatch in sizes of optical claddings, at that axes of the fibers are decentered.

Nonconcentricity and nonroundness (out-of-roundness) of the core and of the cladding. Also, the loss may be caused by nonconcentricity of the fiber core positioning within the optical cladding. In the ideal case, axes of the core and of the cladding must be coincident. Mismatch associated with nonconcentricity is determined by the intercentral distance between the core and the cladding.

Nonroundness (out-of-roundness) of the core in an optical fiber leads to the same effect as the differing core diameters. Nonroundness of the fiber differently influences the attenuation magnitude in a connector at two variants of the connector engaging. In the first case oval cores are coaxial, whereas in the second case a situation is possible when the loss in optical power is maximal – turning of the connected optical fibers at an angle of 90° with respect to each other. This effect is especially appreciable in releasable optical connectors without reference slots because attenuation varies with every subsequent engagement depending on the fiber position.

The above-mentioned cases are peculiar to all optical fibers though in the process of manufacturing the geometrical parameters are tightly controlled. In the last few years the manufacturing techniques of optical fibers have been improved so that the loss in releasable optical connectors associated with the geometrical parameters of the fiber is lowered considerably.

External factors.

The connectors themselves also contribute to the loss in connection. If the central axes of two fibers are brought into coincidence insufficiently, the loss occurs even in the absence of variations in the characteristics of fibers. Let us consider four main causes for the loss in a connector we should control.

Lateral displacement. A fiber should be positioned in the connector exactly along its central axis. The loss is inevitable when the fibers are not coaxial (the central axes are not coincident). Permissible mismatch in this case decreases with a decrease of the fiber size. To illustrate, a displacement of 10 % leads to the loss about 0.5 dB. In the case of a fiber with the core diameter 50μm a relative

displacement of 10 % means that the real displacement is at a level of 5 μm, corresponding to the displacement in every connector by 2.5 μm. It is clear that the lateral displacement control is especially difficult in small-diameter fibers. The majority of the manufacturers of releasable optical connectors aim at limiting this displacement to a level below 5 % of the core diameter.

Spacing between surfaces of connected optical fibers. Connection of two fibers with a small spacing is prone to losses of the two types. The first is associated with Fresnel reflection due to the difference in refractive indices of the fibers and of the spacing medium (usually air). Fresnel reflection takes place both at the output of the first fiber and at the input to the second fiber. In glass fibers separated by the air gap the loss by Fresnel reflection comes to about 0.34 dB. We can lower Fresnel losses considerably by the use of a liquid with the matched refractive index. Such a liquid is an optically transparent medium or gel, with the refractive index close to that of glass.

The second type of losses in multimode fibers is associated with losing of the higher-order modes on transition of light through the spacing and at the entrance to the core of the second fiber. The loss in these cases is dependent on the numerical aperture *NA* of the fibers. A fiber with a high value of *NA* cannot have such a great interfiber spacing at the same level of losses as in the fiber with a lower value of *NA*.

Most of the modern releasable optical connectors have spring-attached ceramic ferrules to ensure physical contact of optical fibers with fixed pressure. This makes it possible to avoid the air gap and to provide the adequate physical contact without the risk of their damaging.

Angular axis misalignment. Chips of the connected optical fibers must be perpendicular to the fiber axes and parallel to each other. A level of the loss is growing with an increase in the misalignment angle. Similar to the previous case, the loss is dependent on numerical apertures of the connected fibers. But in this case the effect is opposite to that with the interfiber spacing – a great value of the numerical aperture *NA* can compensate, to some or other extent, for the angular misalignment of axes.

By the adequate use of a connector, the angular misalignment is practically excluded as the manufacturers usually control perpendicularity of the surface chip relative to the axis of optical fiber.

Chip surface quality. The surface of a chip should be smooth and free of defects (e.g., cracks or scratches). Rough surface of the fiber end is destructive

for the geometrical pattern of light beams, induces their scattering. As a result, attenuation in a releasable connection is higher.

5.4.4 Ferrules of optical connectors

As demonstrated in the preceding section, several factors contribute to the loss in connected optical fibers. Both waveguides must be aligned precisely and should be in close contact. For the protection of fragile fibers subjected to repeated alignment, their terminations have ceramic, plastic or steel ferrules. The majority of these ferrules are cylindrical in form and their diameter is 2.5 mm. Sometimes they are conical, and the *LC*-type connector has a ferrule 1.25 mm in diameter.

Of particular interest is a form of the ferrule faces. Their treatment requires certain workmanship. The simplest variant of the face is flat. This form is associated with a high return loss as the probability of the air gap in the neighborhood of waveguides is high – even roughness of the nonfunctional face surface is important. Because of this, uneven faces with the rounding radius about 10–15 mm are commonly used. Good centering ensures close contact of waveguides, most probably without the air gaps. Most promising is the use of rounded (at an angle of several degrees) faces. Rounded faces are less dependent on deformations arising on engagement of connectors and, as a consequence, such ferrules can withstand more engagements: from100 to 1000.

Engagement of optical connectors. Basically, engagement of two optical connectors of cross equipment is realized according to the following scheme. Outlet is a platform for connectors. The connectors are fixed in such a way that the axes of their ferrules be centered, parallel, and closely pressed. The outlets are usually mounted into patch panels or plugs of mounting boxes. Specifications of the optical connectors of various types and from different manufacturers are somewhat differing but within certain limits.

At the present time more than 20 types of releasable optical connectors are standardized over the whole world. Most widely used are optical connectors of the types *FC*, *ST*, and *SC* (Fig. 5.4.7). In technical literature they are often referred to as the first generation connectors. Besides, the *LC*- and *FDDI*-type connectors are also common.

FC-type connector. The first connector with the 2.5-mm ceramic ferrule was a connector of the *FC* type (Fig. 5.4.7, *a*). With this type the reliable connection is ensured even in the presence of vibrations due to the casing with a thread but this is not convenient for quick disengagement. It is required to turn the threaded head several times, before the connection becomes released or recovered. In some

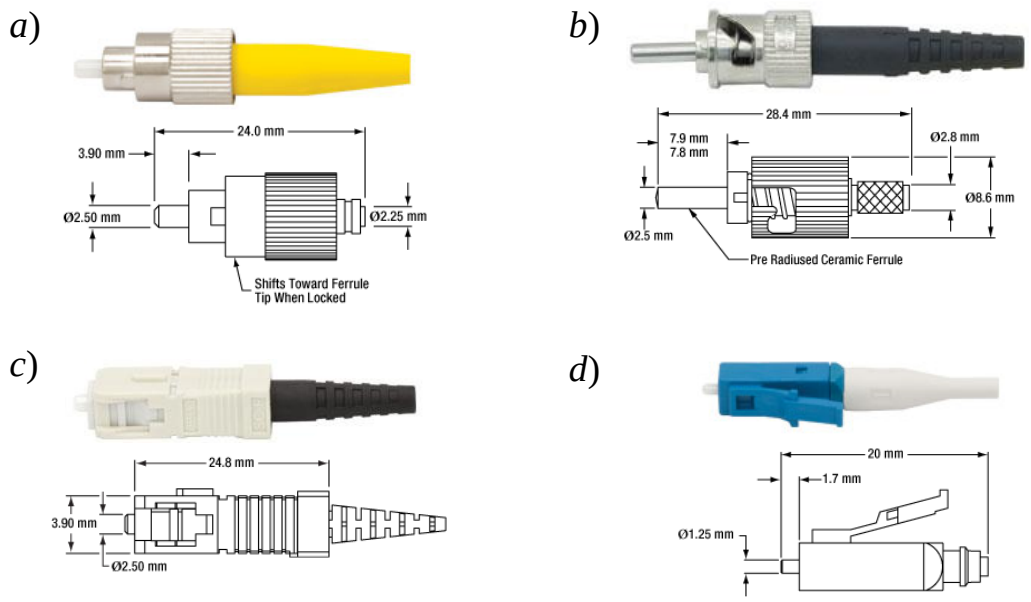


Fig. 5.4.7. – Most widely used optical connectors types

connectors the heads have turning ferrules. Such connectors with a small turning angle of the ferrule ensure that a tip finds the same position within the adapter. The tip is not revolving relative to the casing or to the adapter to minimize any variations in the connection characteristics associated with concentricity or off-roundness of both the tip and the fiber. Such a ferrule is useful when there is a need to minimize the loss for connection. Ferrules in the first connectors had flat faces but later the manufacturers have decided to use the inserts with the rounded outer surface to lower the effect of backward reflection.

ST-type connector. Connectors of the ST-type are also manufactured with a 2.5-mm ceramic ferrule but are distinguished by a quick-disconnected pin system offering the connection retention (Fig. 5.4.7, b). Such a type is preferable when there is no need for protection against vibration, e.g., in office. It is widely used in local networks, interior cable systems, testing equipment, etc. For mating or unmating of the quick-disconnected system, it is sufficient to move the connector head a quarter of a turn. The pressing mechanism ensures stable position of the head in the adapter: in the case of repeated connections one fiber retains its orientation relative to the other. A weak point of the ST-technology is rotational motion of the casing when the connector is engaged or disengaged. Because of this motion, for one connector the operating space should be large enough, that is of particular importance for multiport cabling systems. Moreover, the tip protrudes from the base structure for 5–7 mm and may be service contaminated.

SC-type connector. To eliminate disadvantages of *ST*-connectors, the technology of *Subscriber Connector (SC*) is used. The casing profile is

rectangular in form. Mating/unmating of the connector is realized by the progressive motion along the guides and the connector is fixed with the locks. Cylindrical ceramic ferrule 2.5-mm in diameter has a slightly curved end face (some models have bevel faces). The ferrule is almost entirely covered by the casing and hence is less subjected to clogging as compared to the *ST*-structure. Without rotational motion, ferrules are pressed more safely. Sometimes *SC*-connectors are used in the duplex variant. The connectors may be paired with the help of holders; special clips are used to batch the cases. As a rule, connectors with a single-mode fiber are blue, whereas with a multimode fiber – gray in color.

LC-type connectors. Connectors of the *LC* type represent a small-size variant of *SC*-connectors. Their casing has a rectangular cross-section. The structure based on plastic has a lock similar to those used for modular connectors in systems with copper cables. Because of this, mating of the connector is similar. The ferrule is made of ceramics and has a diameter of 1.25 mm. The connectors may be of both multi- and single-mode variants. These products are applicable in multiport optical systems.

FDDI-type connectors. To connect a duplex cable, we can use not only paired *SC*-connectors. Most common are the *FDDI*-type connectors. They are made of plastic and have two ceramic ferrules. To exclude malconnections, the outlet profile of a connector is asymmetric. The *FDDI* technology provides for four types of the ports used: *A*, *B*, *S,* and *M*. The problem associated with identification of some outlets is solved by the use of special plugs distinguished by their color or alphabetic code. Connectors of this type are mostly used for engagement of terminal equipment to optical networks.

SMA 905 connectors are used for large core, multimode fibers with cladding diameters ranging from 125 to 1580 μm. These connectors are threaded like FC connectors. Industrial lasers, optical spectrometers, military; telecom multimode Ferrule diameterTyp. 3.14 mm The ferrule is made of steel and has an enlarged diameter of 1.25 mm.

5.4.5 Optical-radiation attenuation measuring methods

Among numerous methods for attenuation measuring, the most widely used methods are as follows: *a*) signal comparison at the input and at the output of a fiber; *b*) fiber cleaving method; *c*) backward scattering method (reflectometry).

The first method is the most direct approach to measuring attenuation of optical radiation. To calculate the attenuation factor α_L, we can use formula (5.4.13) involving the signal power at the input P_{in} and at the output P_{out} of a fiber,

and the fiber length L. This method suffers from disadvantage that the radiation coupling loss is unaccounted. To minimize the error, attenuation is measured at two segments of a fiber differing in their lengths ($L_1 > L_2$) for the same signal levels at the input. Attenuation is estimated by formula (5.4.13) for the length $L = L_1 - L_2$, but we take power values at the output of the second and of the first fiber segments as the input and output powers, respectively.

In fact, the second method is rather similar. Radiation is injected into the input end (flat and perpendicular to the fiber axis) of an optical fiber. In the process the radiation source and the output end of the fiber are rigidly fixed to avoid variations of the radiation coupling conditions in the process of measurements. We take a fiber of the specified length L_0 and measure the optical power P_1 at the output end of the optical fiber. Then, with the use of the fiber cleaving technique, a segment of the length $L_1$is detached. The end face of the remaining fiber, having the length $L_2 = L_0 - L_1$, should also be flat and perpendicular to the axis of the optical fiber – this is controlled by means of a special microscope. If a quality of the end face is inadequate, cleaving is repeated and the face quality is controlled. Provided the end face quality is adequate, the optical power at the fiber output P_2 is recorded. The attenuation factor α_L in a fiber of the length L_1 is determined by formula (5.4.13); the input power is designated as P_2 and the output power – as P_1.

The advantage of these methods is that they require no special devices and may be realized with the use of standard recording units. However, the difference is that in the first method measurements are conducted for different fibers, whereas the second method is tedious enough and results in fiber damaging.

Reflectometry is the most exact method based on measurements of inverse Rayleigh scattering in the time domain. To realize this task, a periodic sequence of optical pulses is injected into the optical fiber. The low-intensity radiation scattered within the fiber returns to the input end. The greater the delay between the input and output pulses, the larger the distance covered by light through the fiber. Time scanning the output radiation, we can obtain a noise signal of scattered radiation, with the average exponentially lowering in time. Repeated accumulation of the results enables one to lower a noise level and to construct a plot for attenuation as a function of the measured fiber length; this plot is called the optical reflectogram. A reflectogram is informative not only about attenuation but also about the optical fiber length or distance to the local inhomogeneities (e.g., to the place of fiber damage). A need for expensive equipment is the main disadvantage of this method.

***Diagnostics of fiber-optical communication lines*.**

Installation and servicing of fiber-optical lines is impossible without several measurements. The most typical measurement at the installation stage is measuring of attenuation of the whole line or of its particular segments (at the welding places, mechanical joints, etc.). During the starting-up and adjustment procedures the optical power levels at the transmitter output and at the receiver input are measured and the error ratio is recorded. When some problems are reveled, an optical reflectometer is used for diagnostics of the line.

The operation principle of optical reflectometer is based on analysis of the reflected optical pulses introduced into the optical fiber. When propagating along the fiber-optical lines, light pulses are subjected to reflection and attenuation by optical inhomogeneities and due to the medium absorption. The characteristic formed on the basis of the obtained data is called the reflectogram. An analysis of the reflected pulses makes it possible to determine the length of a fiber-optical line and signal attenuation for this line including the loss by connectors and couplers.

The advanced optical reflectometer offers solution of the following problems:

- testing of FOCL in the automatic mode when a reflectometer independently finds the optimum parameters for measurements, presenting the data in the form of a reflectogram and a detailed table;
- estimation of the optical line length and of the distances to the points of optical inhomogeneities (welding places, commutation points, and the like);
- calculation of attenuation in the line, of the values for its return loss and reflected signal;
- visualization of damages in FOCL.

It is important that a reflectometer offers the possibility to determine in a fiber a distance to the location of inhomogeneities which may be associated with fiber breaking or with structural changes in the fiber. Besides, comparing the current reflectogram with the earlier obtained or reference reflectogram, one can immediately detect deviations in the line parameters possible in the course of time.

5.5. Fiber-optical sensors

Fiber-optical sensors are used to measure properties of the environment. The sensors can measure any physical quantities characterizing the effect that changes the manner of light transmission by the fiber or changes the properties of light. A physical quantity is measured by the sensor based on the induced variations in the

radiation intensity, wavelength, phase, polarization and spectral distribution of light transmitted through optical fiber. Besides, one can measure transmission time of an optical signal.

Owing to some characteristics of optical fibers, such as sensitivity to microbending, interferometric effects, variations in the refractive index, polarization, and in the fiber length, the effect of the fiber diffraction grating or the Sagnac effect (when light propagates in counter directions about the loop used to determine the rotation), they may be used for the development of sensors.

Mechanical or electronic sensors are used in many measuring applications but fiber-optical sensors have several advantages over them and are extensively used because they

- are compact;
- may be realized with the point and distributed sensitive area;
- offer the possibility of multiplexing the transmitted signals with the use of different wavelengths of light for every sensor or by measurements of delay times when light is transmitted through every sensor;
- are multifunctional;
- immune to electromagnetic interference;
- are not under the electric current and hence there is no risk of sparking, so the sensors are used in fuel tanks, in oil-extraction and -refining;
- enable realization of remote access;
- are resistant to the environmental conditions over a wide range of temperatures and pressures and exhibit high corrosion stability.

Also, fiber-optical sensors are used for the detection of various properties:

- temperature;
- pressure;
- movement;
- rate;
- acceleration;
- rate of rotation;
- vibration;
- acoustic vibrations;
- electrical and magnetic fields;
- mechanical stress;
- concentration of chemical substances.

Sensor types.

According to different criteria, optical-fiber sensors may be subdivided into the following types.

According to their spatial realization:

- point sensor: variations in measurements are detected only close to the sensor;
- multiplexed sensor: several localized sensors are arranged at certain intervals along the fiber length;
- distributed sensor: detection is distributed along the whole length of a fiber.

According to the form of the transmitted light-signal characteristic:

- sensors for measurements of the intensity variations;
- sensors for measurements of the frequency or wavelength variations;
- sensors detecting the phase modulation;
- sensors detecting the polarization state.

Sensors for measurements of intensity variations

These sensors detect variations in the intensity of light transmitted through the fiber, which correlate with variations in the pressure or temperature, etc. They are simple and inexpensive measuring units comprising only a simple source and a detector. Nevertheless, these sensors are sensitive to the power fluctuations of a light source, to the connector loss, to fiber bendings. Besides, one can detect variations of the mode power distribution in a multimode fiber.

Sensors for measurements of the frequency and wavelength variations

These sensors are more intricate. A signal in such systems is varying beyond the sensing region. The wavelength measuring method is highly sensitive and weakly dependent on losses of light by connections or on the source intensity variations.

Most common are fibers with Bragg grating. The characteristic wavelength of reflected light is determined by the grating period. The effect changing in a fiber the order of the Bragg grating and, accordingly, the detected wavelength is recorded.

Sensors detecting the phase modulation and polarization state are most complex and sensitive devices.

The phase modulation is recorded by interferometric methods. Polarization sensors are realized, e.g., with the use of the Faraday effect for measurements of magnetic fields and with recording of polarization rotation.

Disadvantages of these sensors are associated with variations in the phase or polarization state of radiation in the case of unintentional bending, stretching or twisting.

According to the region with modulated characteristics of transmitted light signal:

- *internal* when the measured quantity influences transmitted light within the fiber;
- *external* when the light travelling from the source to the detector leaves the fiber, modulation of the characteristics takes place outside the fiber, light reenters the fiber to be recorded by the detector.

Internal fiber-optical sensors.

Optical fibers are used as sensors for measurements of deformation, temperature, pressure, and other quantities by the intensity, phase, polarization, wavelength or transit time modulations in the fiber. Of particular importance is the fact that, if required, internal fiber-optical sensors give the opportunity of distributed sensing for very long distances.

In internal temperature sensors (temperature sensitive elements) the loss of fibers is decreasing with temperature variations. The voltage is easily revealed by the nonlinear optical effects in a specially doped fiber – these effects change polarization of light depending on the electric field voltage. The Sagnac effect is used to realize sensors for measurements of the deflection angle. For operations in conditions of strong magnetic fields (MRT) one can use fiber-optical headset (microphone and headphones).

High-frequency electromagnetic fields are detected by significant modulation of the phase in the presence of an external field due to the Faraday and Kerr effects.

Such sensors are sensitive, characterized by difficulties of multiplexing and lower requirements to connection losses; they are rather expensive.

External fiber-optical sensors use a fiber-optical cable (commonly multimode) to transmit modulated light from a nonfiber optical sensor or from an electronic sensor connected to the optical transmitter. The principal advantage of external sensors is their ability to reach the places otherwise inaccessible. They

may be used to measure internal temperature of electrical transformers, where other measuring methods are impossible due to extraordinary electromagnetic fields.

External sensors are used to measure vibrations, rotations, displacements, rates, accelerations, torques, and temperatures. These sensors are less sensitive, characterized by simplicity of multiplexing and higher requirements to the connection loss; they are easily used and less expensive.

Fiber-optical gyroscope

Fiber-optical gyroscope is used to locate the spatial position and to determine the angular rate. Its operation principle is based on the Sagnac effect. This effect is as follows: when light is emitted by the source into a closed immobile loop in counter directions, the phase difference after bypassing the loop is zero. When the plane, where the circuit is positioned, begins to rotate, the phases of counter-propagating beams differ by the quantity φ_S that is called the phase shift of Sagnac and is proportional to the angular rate of rotation Ω_p about the axis perpendicular to the loop.

$$\varphi_S = \frac{8\pi AN\Omega_p}{\lambda c} \tag{5.5.1}$$

AS follows from formula (5.5.1), the phase shift is proportional to the area A limited by the light beam propagation profile and to the number of the loop bypasses N. It is clear that realization of such a light profile necessitates the use of a single-mode fiber with a low attenuation factor. Light arrives from source *1* to coupler 2 (Fig. 5.5.1), where it is split into two counter flows, and travels

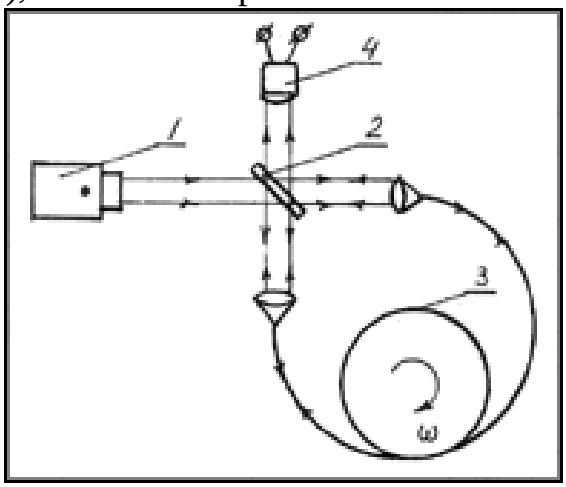


Fig 5.5.1 Schematic of fiber-optical gyroscope

through optical fiber coil *3*. After the fiber, counter flows of light meet at

photodetector *4* that detects the incursed phase due to rotation of the object with the mounted gyroscope, and the angular rate is calculated.

Among the advantages of fiber-optical gyroscopes, we can name their small dimensions and low weight, the absence of rotary mechanical elements contributing to reliability and almost instantaneous availability of the device, great dynamic range of the measured angular rates from 1 grad/h to 300 grad/s, high attainable sensitivity (precision) (above 0.1 grad/h).

Gyroscopic systems based on optical fibers are used in navigation, for guiding of ships, aircraft, tanks, missiles, etc.

References

1. D.J. Sterling, Jr., Technician's Guide to Fiber Optics. 2nd Ed., Delmar publishers Inc., 1993.
2. D. Marcuse. Optical waveguides. M., World., 1974.
3. G.P. Agrawal. Fiber-Optic Communications Systems, 3rd Ed. John Wiley & Sons, Inc., 2002
4. M. Adams Introduction to the Theory of Optical Waveguides, M., World, 1984.
5. L.M. Andrushko, I.I. Grodnev, I.P. Panfilov. Fiber-optic communication lines. M., Radio and communication. 1985.
6. M. Adams. Introduction to the theory of optical waveguides. M., World., 1984.
7. Fundamentals of the theory of fiber-optic communication. Ed. EAT. Dianova. M .: Owls. radio. 1980. 232 p.
8. R. Noé. Essentials of Modern Optical Fiber Communication, 2nd Edition, Springer-Verlag Berlin Heidelberg 2010.
9. H. Venghaus, N. Grote, Fibre Optic Communication. Key Devices. Springer-Verlag Berlin Heidelberg. 2012.
10. Information is also obtained from several Wikipedia webpages
11. www.helkamabica.com/optical-fibre-cables-overview
12. www.thorlabs.com

Chapter 6. Nanophotonics

6.1 Quantum and classical confinement effect

The problems of particles confinement in potential well is a key problem in quantum mechanics. The potential well presents some model, meaning and essence of which are considered in course of atomic physics [1], in the course of nuclear physics and physics of elemental particles. On the base of knowledge obtained in the frame of these courses, solving the Schrödinger equations for the simplest one-dimensional problems it is easy to generalize that the potential well form determines the character of energy eigen value spectra of given quantum system.

So, for the rectangular box with infinitely high walls or with walls of finite height the spectrum of energetic states diverges with increase of quantum number (fig. 6.1.1, *a*). For the wells with typical the Coulomb potential (fig. 6.1.1, *b*) the spectrum converges to limit value – dissociation energy. For the parabolic well the energy spectrum is equidistant (fig.6.1.1, *c*).

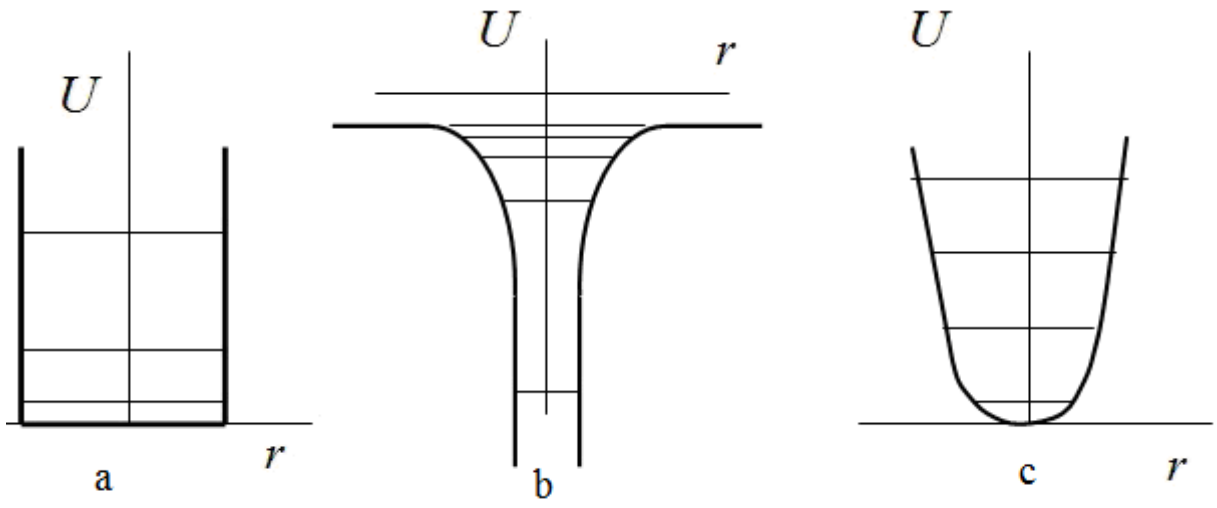


Fig. 6.1.1 – The models of rectangular (a), Coulomb (b) and parabolic (c) potential well with typical spectra of energetic states realized in such a wells

The choice of potential in a concrete quantum-mechanical problems on practice is carried out on the base of experimental data analysis. In nuclear physics there are a number of models such as shell model or the model of 5 –dimensional harmonic oscillator. In these models the form of potential well is specified then the results of Schrödinger equation solving is compared with real nuclear properties – the availability a stable nucleus with magic number of nucleons or collective surface vibrations of nuclear.

The situation is similar in physics of nanostructures. The simplest theoretical model is the spherically symmetric well with barriers of finitely height. This

problem is solving for the nonequilibrium charge carriers – the electron and the hole – in assumption of effective masses model.

In the frame of first lection let us consider how the effect of the size quantization may be explained. Taking into account the exact solution of the Schrödinger equation for a particle in an infinitely deep potential well, we do that from the first principles. We regard how to use the terminology of band theory of crystals and molecular orbitals theory for the characterization of electronic quantum states in nanostructures; how this problem is detailed for the cases of low-dimensional systems and structures with heteroboundaries.

The solution of problem of particle localization for the simplest case of one dimensional well with infinitely high walls is regarded in detail in course of atomic physics and quantum mechanics [1]. Let us present the results and remember what the basic conclusions may be formulated. So, for the particles with mass *m*, localized in one dimensional rectangular well with infinitely high walls and width equaled *a* (fig. 6.1.2), the spectrum of energy eigenvalues and eigenfunctions is given by the following expression

$$E_n = \frac{\pi^2 \hbar^2}{2ma^2} n^2, \; n=1, 2, 3,\ldots \tag{6.1.1}$$

$$\psi_n(x) = \sqrt{\frac{2}{a}} \sin\left(\pi n x / a\right), \; n=1, 2, 3,\ldots \tag{6.1.2}$$

Let us regard haw these results may be achieved from the first principles, from the Heisenberg uncertainty relation.

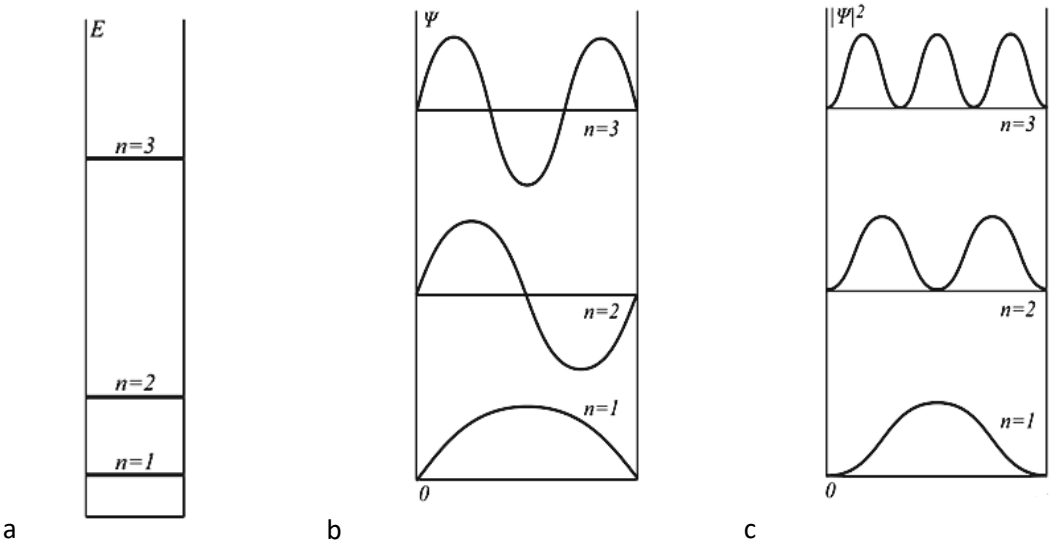


Fig. 6.1.2 – The spectrum of energy eigenvalues (a), eigenfunctions $\psi_n(x)$ (b) and $|\psi_n(x)|^2$ (c) in one dimensional potential well with infinitely high walls and width equaled *a*

Let us suppose that in the state with lowest energy and lowest possible pulse the uncertainty of particle coordinate is equal an order of magnitude to well size $\Delta x \approx a$. The pulse uncertainty is equal an order of magnitude to minimal pulse value $\Delta p \approx p_{\min}$, which due to Heisenberg uncertainty relation

$$\Delta x \cdot \Delta p_x \geq \hbar \qquad (6.1.3)$$

is equal to $p_{\min} = \hbar / a$. Then the minimal possible value of energy $E_{\min} = p_{\min}^2 / 2m = \dfrac{\hbar^2}{2ma^2}$ is consistent in order of magnitude with expression (6.1.1) at $n = 1$.

The basic conclusions which can be formulated on the base of (6.1.1) solutions convey the essence of so called *quantum sized effect*. Let us enumerate it.

1. The particle energy depends on localization region size *a*.

This is the quantum size effect. So as the size of box decrease two times, the energy spacing between neighboring levels increase 4 times as it is evident from expression (6.1.1).

2. The discreteness of energy spectrum inside the box depends on particle mass and on localization region size.

Indeed, at the same mass, for example $m = m_e$, but different well sizes a_1= 1 cm and a_2=0,1 nm, it may be made the following estimation. The $\Delta E_1 = 7{,}5 \cdot 10^{-15} n$ eV in first case, but $\Delta E_2 = 0{,}75 \cdot 10^{2} n$ eV in the second one. In other words, for the microscopic centimeter well the spacing between neighboring energy levels is much smaller than energy on these levels, the spectrum of energetic states can be considered a quasi-continuous. For the mesoscopic well the energy level values are comparable with discreteness and it can not be neglected.

It is important to note that discreteness can not be neglected for nanoparticles and therefore incorrectly employ the concept of energy bands as it is accepted in solid state physics.

3. Because of the uncertainty ratio, there is a lower limit for the energy values in the well.

4. Upon reflection from the well wall a standing de Broglie waves are aroused.

Really, in each state the well width is equal to an integral number of particle half wavelengths of de Broglie

$$n \frac{\lambda}{2} = a . \qquad (6.1.4)$$

The expression (6.1.4) evidently demonstrates that quantum sized effect is due to the wave properties of charge carriers and they and for them it is possible to enter the same characteristics for light quanta – light waves. Indeed, the probability of finding a particle at the vicinity of the center of the well at n=2, is equal to zero although on either side of the well center the probability find a particle is maximal and the same (fig.6.1.2, *c*). The behavior of the particle, which with the same probability density arises on both sides of the point at which it is never possible to find, is specified only for the particle-wave. In fairness, it should be noted, that there are also *classical sized effects*, which are developed in semiconducting films when its thickness are comparable with the Debye screening length, electron mean free path length or the diffusion length [2].

The subject of nanophotonics is interaction of light waves with nanosized structures. Wherein the electron localization range, as a rule, are comparable with the electron de Broglie wavelength. The spacing between neighboring energy level in such structures is comparable with the energy of light.

Expression (6.1.4) shows that the well size or size of localization range must be comparable with the particle de Broglie wavelength. This condition is well satisfied for the semiconductor nanoparticle materials of $A^{II}B^{VI}$ or $A^{III}B^{V}$ groups.

While for the noble metal nanoparticle de Broglie wavelength does not fit in the nanometer range and quantum-sized effect has not such vivid manifestation as for semiconductors. Indeed the electron effective mass in metal is closed to mass of free electrons, the Fermi energy of the order of a few eV, so de Broglie wavelength of the order of 0,1 – 1 nm. Therefore, even in very thin metal films of 10 nm thick the quantum-confinement effects usually are insignificant.

In semimetals and narrow-gap semiconductors the effective mass of electrons $m^* \approx 0{,}01m_e$ and carrier energy $E = 10^{-2}$ eV. De Broglie wavelength in that materials order of 100 nm, therefore quantum effects are clearly seen in crystals and films up to 100 ÷200 nm thickness. However, there are still a number of conditions which must be followed to the quantum size effect was observed in real systems. Let us enumerate these conditions.

Firstly, the discreteness of energy levels must be visible on the background of thermal noise.

$$\Delta E_{n,n+1} >> kT\ . \qquad (6.1.5)$$

For the electronic systems with a given Fermi energy E_F should be satisfied,

$$\Delta E_{n,n+1} >> E_F\ , \qquad (6.1.6)$$

which includes (6.1.5).

The requirement (6.1.6) as well as (6.1.4) show, that in metals the observation of quantum-sized effect is practically impossible.

Second, under the scattering of electrons with the characteristic momentum relaxation time τ and the related mobility μ in real systems it is necessary to

$$\Delta E_{n,n+1} >> \frac{\hbar}{\tau} = \frac{\hbar \cdot e}{m \cdot \mu}. \qquad (6.1.7)$$

This condition is equivalent to the requirement that the free path *l* was much more than the space of a charge carrier localization *a*

$$l >> a. \qquad (6.1.8)$$

Thirdly, the necessary condition is a high quality surface of real structures. Indeed, the role of atoms on the surface of the nanostructures as well as the role of surface itself with its natural defects and surface electronic states are increased in comparison with bulk materials in measure factor as

$$S / V \approx 1 / R,$$

where *S* – square of sample surface, *V*- sample volume, *R* – sample size. The availability of surface electronic states of Tamm or Shockley type, adsorption centres and various kinds of impurities on the semiconductor surface greatly complicates the description of real nanomaterials and reduces the efficiency of their operation.

In the case of thin films it comes to providing the two surfaces – the outer boundary of the film and the film–substrate interface. The interface between two different semiconductors is naimed *heterojunction*. These two semiconductors in close contact – *heterostructures.* The creating of a heterojunction is one of protection techniques, or surface passivation from natural defects, caused by chemical bonds breakage on the surface.

If the motion of charge carriers is limited in one, two or all three directions and the sizes of localization range are comparable with de Broglie wavelength of particle than the energy spectrum becomes partially or totally discrete and dependent on the size of localization range. Such a materials are called *nanomaterials*, production technologies – *nanotechnologies*.

If the motion of charge carriers is limited in one directon than that system belongs to the class of *low-dimensional* systems, namely *2D* dimension systems. An example is the so-called quantum films (fig.6.1.3, *a*). If the motion of charge carriers is limited in two directons than that system belongs to the class of *low-dimensional* systems, namely *1D* dimension systems. An example is the so-called quantum wires (fig.6.1.3, *b*). If the motion of charge carriers is limited in all three directons than that system belongs to the class of *low-dimensional* systems,

namely *0D* dimension systems. An example is the so-called quantum dots (fig.6.1.3, *c*).

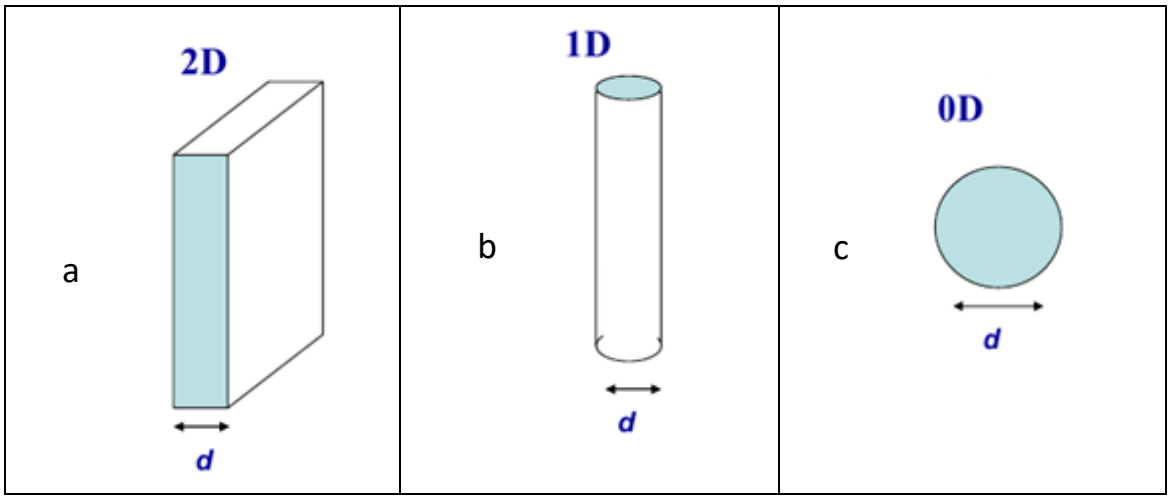


Fig.6.1.3 – Quantum film (a), quantum wire (б) and quantum dot (c)

The model of one-dimensional quantum well is well suited to describe the low-dimensional systems namely 2D (fig.6.1.3, a). In order to use the quantum model of the potential well, whish we have considered at the beginning of the paragraph, for the *1D* and 0D systems it is necessary to pass from one dimensional well to the two-dimensional or three-dimensional potential wells.

The mail conclusions 1-4, which are formulated on the base of solution (6.1.1) remained are also valid for the cases of two and three dimensional potential wells.

The spectrum of energy eigenvalues for real wells can be directly evaluated if in expression (6.1.1) replace the mass of the particle at its effective mass. Electron work function in order of magnitude is much greater than its thermal energy so the potential well in a real structure can be considered infinitely deep.

In the quantum film movement of charge carriers is limited only along one direction perpendicular to the surface, Z axis, for example. In film plane (XY plane) the charge carriers motion does not limited. You can determine the total energy of the particles as follows

$$E = E_n + \frac{p_x^2 + p_y^2}{2m}, \qquad (6.1.8)$$

where E_n – the energy of sized quantization, associated with the motion perpendicular to film plane, p_x and p_y - particle momentum component along the film. There are subband generated for each fixed value of *n*, because the energy can be changed from E_n to infinity. However, spontaneous transition from one

subband to another is impossible, as it is required to expend energy and momentum transfer perpendicular to the film to change the quantum state energy E_n. These *2D* dimensional systems are also called as two-dimensional degenerate electron gas.

In the quantum wires the total energy of the particle also is a discrete-continuous and is given by

$$E = E_{nm} + \frac{p_x^2}{2m}, \tag{6.1.9}$$

where E_{nm} - discrete set of energy states, corresponding to motion across quantum wire, p_x - momentum component along the wire. The electrons in *1D*-dimensional systems are seen as one-dimensional degenerate electron gas. In low-dimensional systems, *0D* or quasi zero-dimensional systems (quantum dots), the total energy of the system is quantized and depends in general on three quantum numbers $E = E_{nml}$.

6.2. Density of states and modified density of states in system of low dimensionality

The density of states concept is one of the fundamental concepts in quantum mechanics, optics and particle physics. From the density of states depends the probability of β-decay [3], knowing the density of states it is possible to estimate the total number of electrons in the metal. For photons of electromagnetic radiation in a wide range of fundamentally important characteristic is the equilibrium density of radiation, which is determined by photon density of states (radiation spectral density) and temperature follow to the Rayleigh-Jeans. In the language of wave mechanics, the photon density of states is the number of standing waves in a spectral range, enclosed in a unit volume of a cavity.

The density of photonic states is the characteristic of the electromagnetic field and does not depend on the presence or absence of the field source. The local density of photonic states can be directly measured using a near-field microscopy, which will be discussed below.

Let us find the number of quantum states, in which can be distributed particles whose energy does not exceed a certain value *E*, and define firstly this number for the case of an electron located in a three-dimensional potential well with impenetrable walls. As it was shown in the paragraph 6.1 (see. (6.1.1))

energy of the electron in a one-dimensional potential well is described by the expression

$$E_n = \frac{\pi^2\hbar^2}{2ma^2}n^2,\ \ n = 1,2,3,...$$

It is easy to prove that in the case of a three-dimensional well it is true the expression

$$E = \frac{\pi^2\hbar^2}{2m}\left(\left(\frac{n_1}{a_1}\right)^2 + \left(\frac{n_2}{a_2}\right)^2 + \left(\frac{n_3}{a_3}\right)^2\right),\ \ n_1, n_2, n_3 = \{1,2,3,...\}, \qquad (6.2.1)$$

where a_1, a_2 a_3 – the size of the potential well or side a cuboid, in which the particle is contained, quantum numbers n_1, n_2 and n_3 independently take positive integer values and correspond to quantum numbers, describing quantum states in three independent measurements. It follows from (6.2.1) that the electron energy does not change in a continuous manner, but discretely in all three dimensions. Let us consider the value of energy which significantly excesses the energy of the ground state ($n_1 = n_2 = n_3 = 1$). In the case of macroscopic well, as it is shown in 6.1, energy change from level to level $\Delta E << E$, so we can assume that the electron energy changes almost continuously (quasi-continuously).

Let us consider the phase space of quantum numbers, that is, three-dimensional space along three mutually perpendicular axes of which the quantum numbers n_1, n_2 and n_3 are plotted, (Fig. 6.2.1 *a*). Each point of this phase space corresponds to two states with different electron spin projection. Thus, each electron state in the phase space can be set by four quantum numbers n_1, n_2, n_3, and m_s, where m_s - the quantum number of the particle spin projection.

We introduce the notation: $\vec{r}$ the radius vector of a point (state) in phase space by using the following expression

$$r^2 = (a_2a_3n_1)^2 + (a_1a_3n_2)^2 + (a_1a_2n_3)^2 \qquad (6.2.2)$$

in order to the expression (6.2.1) give to form

$$E = \frac{\pi^2\hbar^2}{2m(a_1a_2a_3)^2}r^2. \qquad (6.2.3)$$

Let us express the radius in the phase space region as *r* terms through the energy *E* of the particles whose states are enclosed in this area

$$r = \frac{a_1a_2a_3\sqrt{2mE}}{\pi\hbar}. \qquad (6.2.4)$$

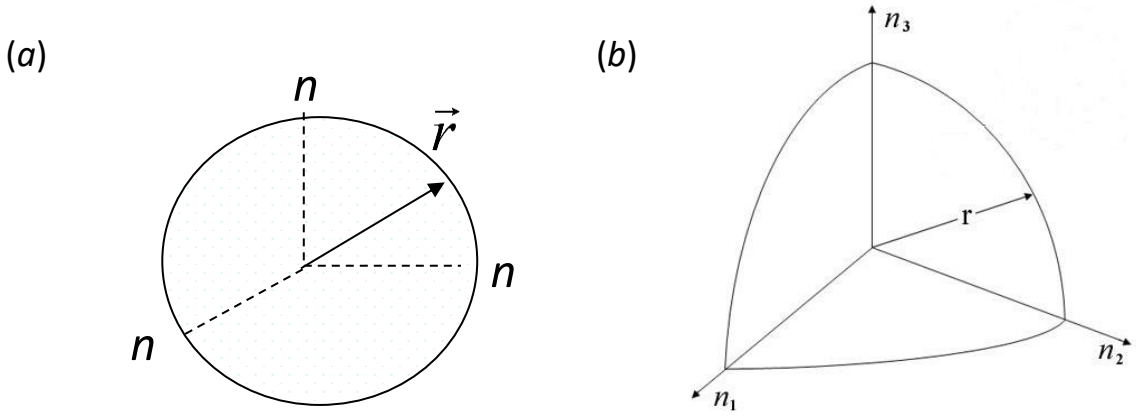


Fig. 6.2.1 – Three-dimensional phase space of integers n_i, including states with different energies, defined by expression (6.2.3)

All the really existing states, in accordance with the values of integers $n_1, n_2, n_3 = \{1,2,3,...\}$ are defined by points bounded in phase space octant (fig.6.2.1, *b*). In order to find the total number of states *G*, we need octant volume (i.e. 1/8 of a sphere volume) to divide into the volume ΔV, attributable to one state (one point in the phase space), and to multiply the resulting expression by a factor of 2, equaled to the number of possible electron spin projections $m_s = \pm 1/2$,

$$G = \frac{1}{8} \cdot \frac{4}{3} \cdot \pi r^3 \cdot \frac{1}{\Delta V} \cdot 2 = \frac{1}{3} \pi r^3 \cdot \frac{1}{\Delta V} \cdot$$

The unit volume in phase space specified by a set of integers, corresponds to one state or one point ΔV=1. Therefore, the total number of states

$$G = \frac{1}{3} \pi r^3 \cdot$$

Replacing in this expression *r* in accordance with the expression (6.13), we obtain

$$G = \frac{1}{3} \pi \frac{(\sqrt{2mE})^3}{\pi^3 \hbar^3} a_1 a_2 a_3.$$

We further recognize that the product of $a_1 a_2 a_3$ is equal to the well volume *V* , and the $\sqrt{2mE}$ is equal to non-relativistic electron momentum p, we obtain

$$G = 2 \frac{4}{3} \pi p^3 V \frac{1}{(2\pi\hbar)^3}. \qquad (6.2.5)$$

Let us transfer to another phase space of three dimensions, corresponding to the three spatial coordinates *x*, *y*, *z* and to the three momentum projections the p_x, p_y and p_z. In this new phase space the total volume which includes all possible states is equal to the product of V by $\frac{4}{3}\pi p^3$.

Therefore, the expression (6.2.5) can be rewritten in the form corresponding to this phase space with the phase volume

$$V' = V \cdot \frac{4}{3}\pi p^3. \tag{6.2.6}$$

by the following manner

$$G = \frac{2V'}{(2\pi\hbar)^3}. \tag{6.2.7}$$

In the expression (6.2.7) factor 2 corresponds to the number of possible spin projections, and is not associated with the movement of particles in space. If we now apply (6.2.7) to a single state, with a twofold degeneracy of the spin, we find that the minimum volume of a phase including one such state is equal to $(2\pi\hbar)^3$ or h^3. We can this statement to writ in mathematical form as follows:

$$\Delta x \cdot \Delta y \cdot \Delta z \cdot \Delta p_x \cdot \Delta p_y \cdot \Delta p_z = (2\pi\hbar)^3.$$

Further, given that all the directions are equally probable, we get

$$\Delta x \cdot \Delta p_x = 2\pi\hbar. \tag{6.2.8}$$

Thus, each state in the phase space occupies a volume $2\pi\hbar$ *or h for each space coordinate.* This finding is important for understanding the effect of modifying the density of states in low-dimensional systems, which will be discussed after the definition the concept of density of states for arbitrary particle localized in the potential well of arbitrary shape. For this purpose in the completion of this section we generalize expression (6.2.7) for the total number of states for the case of *any* particles as follows:

$$G = J_s \cdot \frac{V'}{(2\pi\hbar)^3}, \tag{6.2.9}$$

where J_s is the number of possible projections of the particle spin, V'- phase volume occupied by the system.

The density of the energy states $D(E)$, by definition, is the number of states accounted for the energy range from E to $E+dE$

$$D(E) = \frac{G(E+dE) - G(E)}{dE} = \frac{dG}{dE}$$

or

$$D(E) = \frac{dG}{dp} \cdot \frac{dp}{dE}. \tag{6.2.10}$$

Taking into account the (6.2.9) we rewrite (6.2.10) as follows:

$$D(E) = J_s \frac{d}{dp}\left(\frac{4}{3} \cdot \frac{\pi p^3 V}{(2\pi\hbar)^3}\right) \cdot \frac{dp}{dE}$$

or

$$D(E) = J_s \cdot \frac{4\pi p^2 V}{(2\pi\hbar)^3} \cdot \frac{dp}{dE}. \qquad (6.2.11)$$

With the help of this relationship we find the density of energy states for electrons and photons.

For non-relativistic electrons ($p = \sqrt{2mE}$, a J_s=2,) we obtain

$$D_e(E) = \frac{\sqrt{2}m^{3/2}}{\pi^2\hbar^3} \cdot V \cdot \sqrt{E}\,. \qquad (6.2.12)$$

For photons, the number of spin projections is two as a result of transverse nature of light waves, $p = E/c$ and

$$D_{ph}(E) = \frac{V}{\pi^2 c^3 \hbar^3} E^2\,. \qquad (6.2.13)$$

Using these expressions, we can obtain the number of particles *N*, in the states with the energy range from *E* to *E* + *dE*. To do this, multiply the value of the density states on the particle distribution function over the states

$$dN = D(E) \cdot f(E) \cdot dE, \qquad (6.2.14)$$

where *f*(*E*) is Fermi-Dirac distribution for electrons and Bose-Einstein distribution for the photons, respectively.

Figure 6.2.2 *a* shows a plot of (6.2.12) for electrons, and Figure 6.2.2, *b* - plot of (6.2.13) for photons from the book S. Gaponenko «Introduction to nanophotonics» [4]

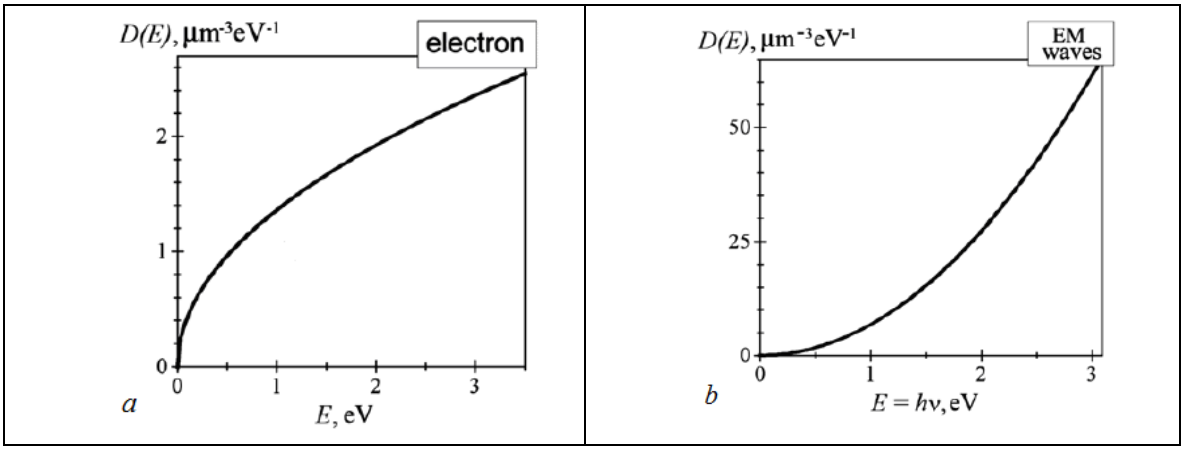


Fig. 6.2.2 – Density of states for electrons (a) and photons (b) in three-dimensional space [4]

Figure 6.2.2 shows the dependence of the electron (fig. 6.2.2, *a*) and photon (fig. 6.2.2, *b*) density of states [4] on the energy for the case of three-dimensional space. From this point forward we will denote the density of states in three-dimensional space, as $D^{(3)}$. Let us consider how this value changes during the

transition to the two-dimensional space for the degenerate two-dimensional electron gas.

Such an electron gas can be considered in four-dimensional phase space (x, y, p_x, p_y). The volume occupied by them with energy in the first (the lowest) subband of sized quantization

$$L_x L_y \pi p^2 = 2\pi L_x L_y m(E - E_1),$$

where L_x and L_y sample sizes in a plane, E_1 – the energy of the lowest quantum state. The minimum volume of the this phase space is $(2\pi\hbar)^2$. Then the total number of states considering the double degeneracy in spin

$$G(E) = \frac{\pi L_x L_y}{\pi\hbar^2} m(E - E_1).$$

Then the density of states for a two-dimensional electron gas $D^{(2)}(E)$ in the first size quantization subband at $E_1 < E < E_2$

$$D^{(2)}(E) = \frac{1}{L_x L_y}\frac{dG}{dE} = \frac{m}{\pi\hbar^2}$$

An increase in energy from E_1 to E_2, new states in the second subband appear and the density of states function undergoes a jump by the amount $\frac{m}{\pi\hbar^2}$.

Such jumps will occur every time if the energy will reach the bottom of another subband. Therefore, the density of states function for a two-dimensional degenerate electron gas can be written as

$$D^{(2)}(E) = \frac{m}{\pi\hbar^2}\sum_n \Theta(E - E_n), \qquad (6.2.15)$$

where $\Theta(x)$ – unit Heaviside function equaled to zero at $x < 0$ and unity at $x > 0$.

For one-dimensional degenerate electron gas in quantum wires with a spectrum of energy states (6.9) we can calculate the number of states in the two-dimensional phase space (x, p_x). The volume occupied by electronic states in two-dimensional phase space

$$2L_x p_x = 2L_x\sqrt{2m(E - E_{11})},$$

where L_x – wire length, and E_{11} - the energy of lowest quantum state. The minimum volume of this phase space is $2\pi\hbar$. Then the total number of states considering the double degeneracy in spin

$$G(E) = \frac{2L_x\sqrt{2m(E - E_{11})}}{\pi\hbar}.$$

The density of states for a one-dimensional electron gas $D^{(1)}(E)$, taking into account the contributions of all subbands per unit length of the wire

$$D^{(1)}(E) = \frac{\sqrt{2m}}{\pi\hbar} \sum_{mn} \frac{\Theta(E - E_{mn})}{\sqrt{E - E_{mn}}} \qquad (6.2.16)$$

This function becomes infinite when $E=E_{mn}$ that is, when the energy reaches a level of size quantization in wire.

For the quasi-zero-dimensional systems of dimension *0D* energy spectrum is of a purely discrete nature, so the number of states is increased by one every time as the energy takes a value equal to the eigenvalue energy. The density of states is described by the Dirac delta function $\delta(x)$, which is not equal to zero only at the point x=0, in the rest region becomes zero and integral $\int \delta(x)dx = 1$ in the vicinity of point x=0. Using this function, we can write that

$$D^{(0)}(E) = \sum_{lmn} \delta(E - E_{lmn}). \qquad (6.2.17)$$

Figure 6.2.3 shows the graphs of the electronic density of states for the bulk (three-dimensional) crystal (Fig. 6.2.3, *a*), the quantum film (Fig. 6.2.3, *b*), quantum wire (Fig. 6.2.3, *c*) and quantum dots (Fig. 6.2.3, *d*).

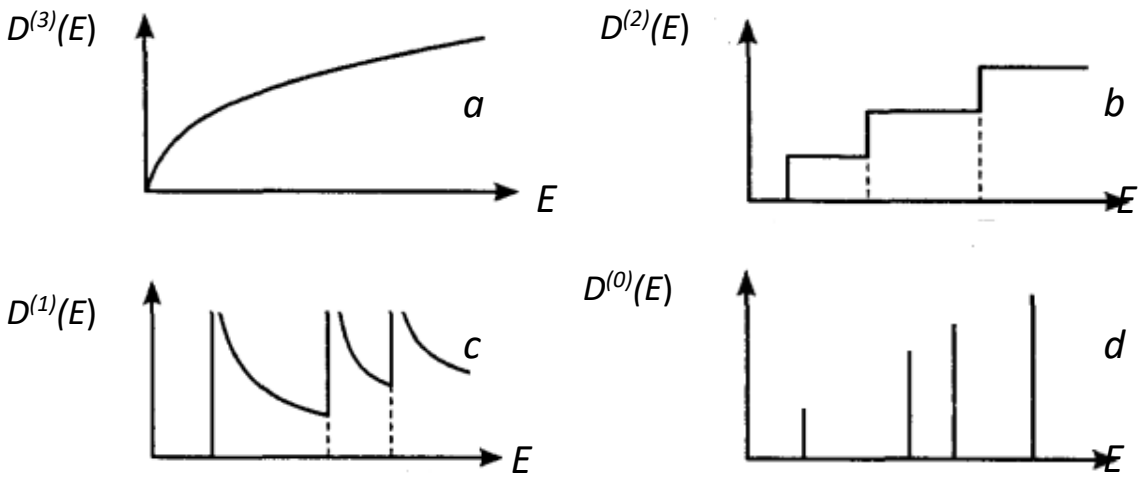


Figure 6.2.3 – Density of states for three-dimensional crystal (a), quantum film (b), quantum wire (c) and quantum dot (d)

The effect of modification of density of states is related to the size quantization is shown in thin films. The energy spectrum of electrons in such films can be represented, as discussed above by expression (6.1.8) as the spectrum of two-dimensional gas with a size-quantization subbands. Graphical representation of the subbands is shown in the graphs of electron energy depending on the pulse (Fig. 6.2.4). Pulse and energy can change in a continuous fashion only in a plane (e.g., *xy*), but not across the film.

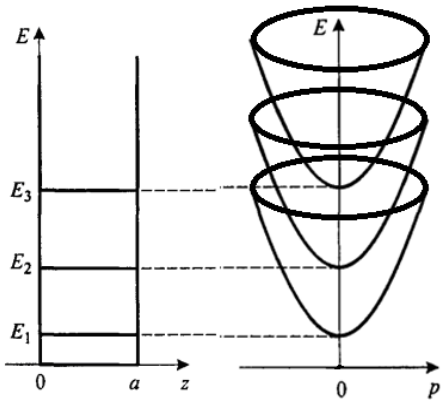


Fig. 6.2.4 – The energy spectrum of charge carriers in quantum-dimensional film with thickness *a* along z axis and momentum in the film plane (*xy*) $p = \sqrt{p_x^2 + p_y^2}$.

One of the quantum size effect results — increasing bandgap crystal thin film while reducing its thickness. It is observed due to the fact that the lower the allowed energy level in the conduction band E_c is raised above the bottom zone, and the upper level in the valence band - falls below the zone roof E_υ. For semiconductor crystals and films into a few nanometers thick the magnitude of this quantum effect can reach several tenths of eV.

In thin quantum wires, this effect is even more pronounced than in the films, as there is the quantization of the energy spectrum of the carriers along two axes across the wire. For example, for silicon quantum wires with a diameter of the order of 1 nm effective bandgap should be increased in comparison with the bulk material on the order of 1 eV. Namely the same shifts of intrinsic absorption bands to higher energy quanta of light are observed experimentally in porous silicon, which is a system of thin quantum wires.

The nonmonotonic dependence of the electrical characteristics of a thin film semiconductor or semimetal on its thickness is a result the modification the density of states. Indeed, the density of states for this film with *N* first occupied subbands per unit of film thickness we obtain by dividing the expression (6.2.15) onto film thickness *d*,

$$D^{(2)}(E)/d = \frac{m}{\pi\hbar^2} \cdot \frac{1}{d} \sum_n \Theta(E - E_n) . \qquad (6.2.18)$$

It follows from (6.29) that density of states decreases with an increase in thickness in the thin film. But simultaneously with increasing of *d* decreases the energy splitting of the size quantization levels. At the same time an increasing number of subbands "plunges" below the Fermi level. Each intersection of the Fermi level of the following subband is accompanied by a jump in density of the

occupied states. As a result a characteristic periodic dependence of the resistivity of the thin film on its thickness (Fig. 6.2.5) is observed. From the period of changes in the electrical properties of the thin film, depending on the value of d can be found p_z at the Fermi level ("Fermi" quasi-momentum), as well as the effective mass m_z^* of carriers [2].

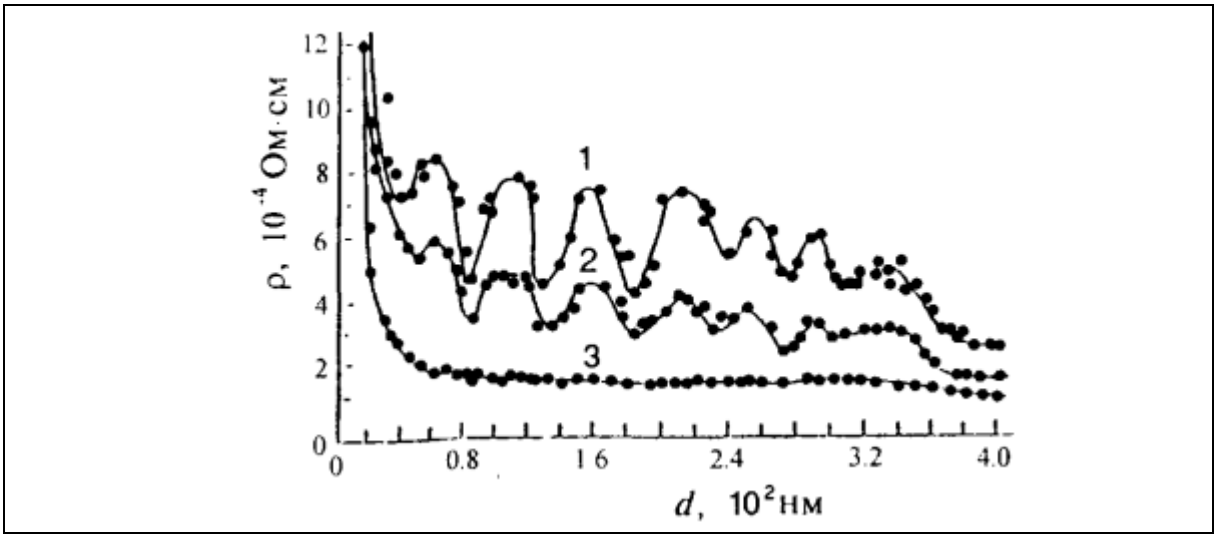


Fig. 6.2.5 –The dependence of the resistivity of the bismuth film on its thickness. The temperature of measurements: 4,2 (1), 78 (2) и 300 K (3) [5])

The positions of the quantum-sized energy levels (subbands borders) in a semimetal or semiconductor film is determined by tunneling spectroscopy by measuring the dependence of the tunneling current magnitude between the quantum film and massive metal electrode on the applied voltage (Fig. 6.2.6, *a*). A thin (<5 nm) oxide of a bulk metal (such as lead oxide) as the tunnel-transparent dielectric spacer between the conductive electrodes usually is used. Under the increasing the voltage applied to bulk electrode more and more electrons from deeper subbands are sequentially joined to the tunneling process . all the deeper subzones voltage sequence includes a massive tunneling electrode. Beginning of "inclusion of" each new subzone is accompanied by a change in the slope of the current-voltage characteristics of the system, that is most clearly seen in the dependences d^2I/dV_g^2 on V_g (fig.6.2.6, *b*).

Similar quantum effects can be observed in thin areas of the space charge on the surface of the massive semiconductor crystals. Quantum effects in thin films, and in the space charge region on the surface of bulk semiconductor crystals appear bright enough only when no more than 1-2 quantum subbands are populated.

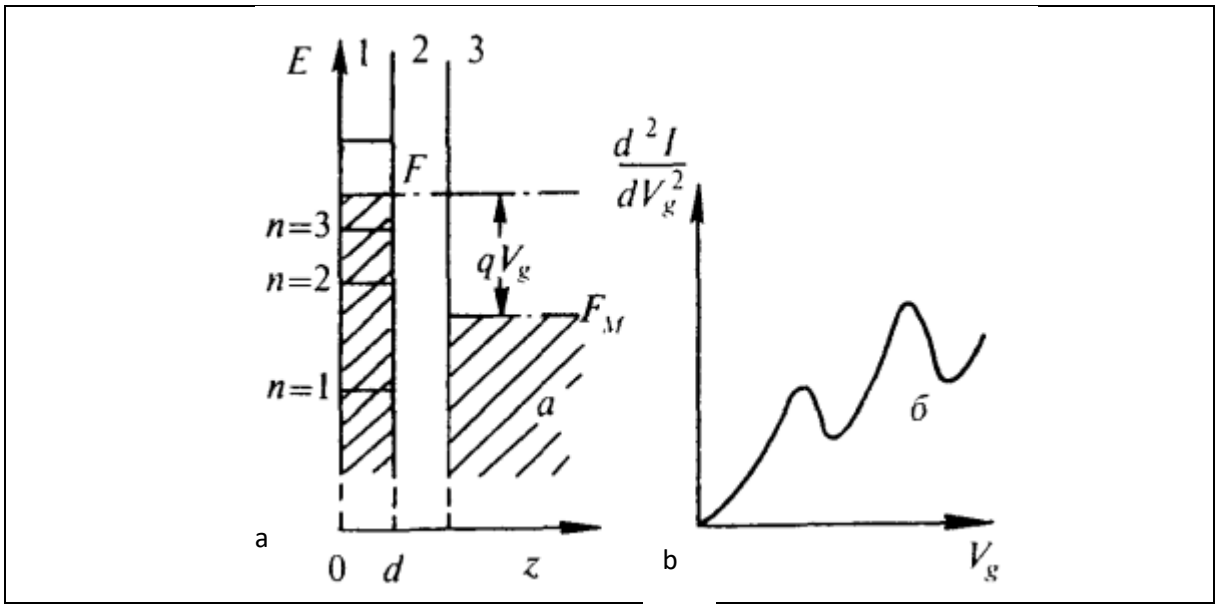


Fig. 6.2.6 a – The energy diagram of the system: quantum film (1), thin dielectric (2), massive metal (3); b — the dependence of the second derivative of the tunneling current on the voltage applied to the structure [5]

As it will be shown below (fig. 6.2.7) absorption spectrum of CdSe / ZnS quantum dots also demonstrate the effect of modification of density of states. The shape of these spectra is also a reflection of the modification of the density of states for quasi-zero-dimensional systems. The minimum energy of the absorbed photon is

$$E_{ph} = E_g + E_1^e + E_1^h, \tag{6.2.19}$$

where E_g – the band gap in the bulk semiconductor, E_1^e – the energy of the electron lowest quantum state, E_1^h – the energy of the hole lowest quantum state. The energy of the absorption edge is greater than in a homogeneous semiconductor, and increases with decreasing the width of the quantum dot diameter. The absorption coefficient should increase stepwise with increasing energy, as it is shown on Fig. 6.6, *d*. At low temperatures, in fact, the absorption spectrum of the quantum dots is similar in form to the dependence shown in Fig. 6.6, *d*. This is explained in the theory of semiconductors that allowed direct optical transitions in absorption spectrum reproduces the function of the density of states. At room temperature the continuous dependence is observed as shown in fig. 6.10 due to the broadening of the bands of the individual transitions. A separate power surges correspond to distinct peaks, which can be distinguished on the background of the continuous spectrum with increasing energy.

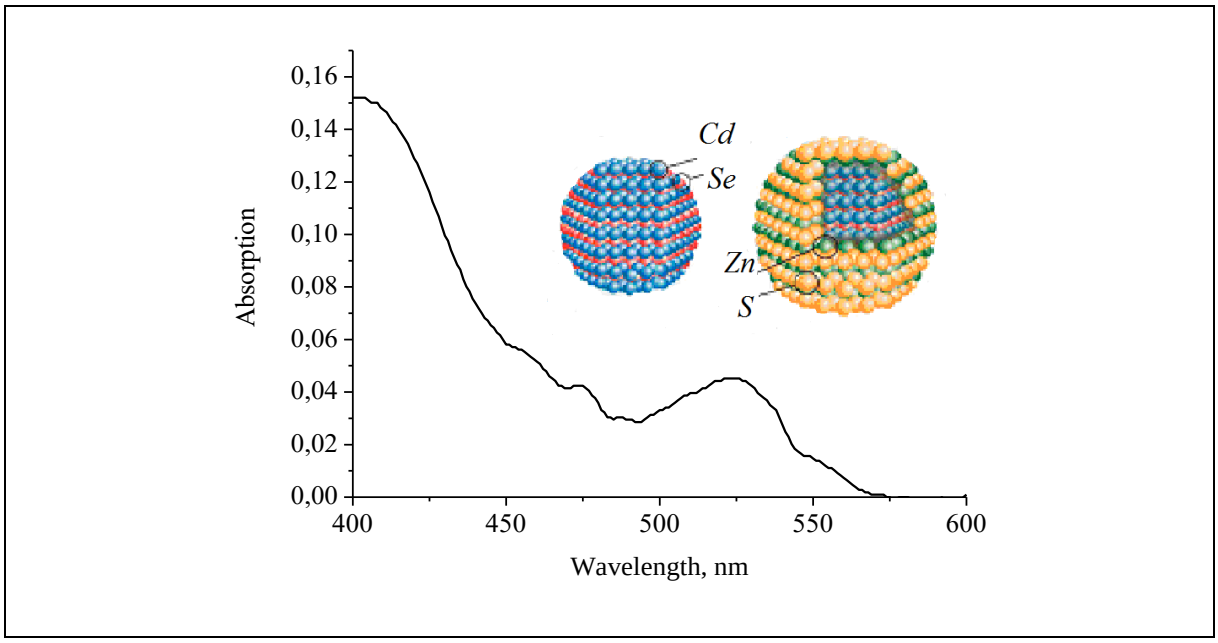


Fig. 6.2.7 – Absorption spectrum of CdSe/Zns quantum dots 2,6 nm in diameter dispersed in toluene at 293K

6.3. Interaction of light with nanostructures

Nanophotonics subject, as mentioned above, is the interaction of light with nanostructures. The dimensions of the nanostructures is much less than the light wavelength, and all the features of this interaction should be considered, using the concept of the near field of the light wave. Near field occurs at all stages of the light waves from the transmitter to the receiver - for radiation stages, distribution, and finally interaction with matter. Firstly, it is the field in the vicinity of the radiating dipole. Secondly, it is a field near the aperture, the size of which is smaller than the wavelength of light. It can also be a field of the light waves near the interface of two materials. As such materials are the two dielectric with total internal reflection on the interface or a metal and a dielectric.

All the effects listed above are caused by so-called *evanescent fields*, i.e. fields that decay at distances shorter than the wavelength of light. Let us consider what determines the properties of the evanescent waves and a what practical interest its represent.

6.3.1 The field of an electric dipole

Let us consider a dipole oscillating along the z-axis, as represented in Fig.6.3.1. The electric polarization of the dipole follows the equation:

$$\vec{P}(\vec{r},t) = p(t)\delta(\vec{r} - \vec{r}_0)\vec{n}\,, \tag{6.3.1}$$

where $p(t)$ is the time-dependent polarizability.

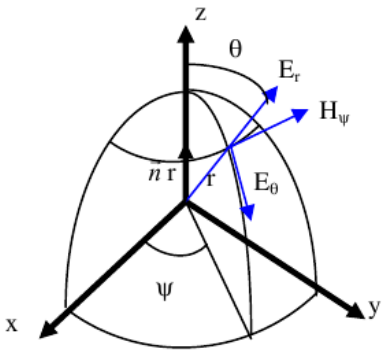


Fig. 6.3.1 – Schematic representation of a dipole with momentum parallel to z-axis

The components of the field generated by the dipole are given by the following equations [6]:

$$E_r = 2\left(\frac{p}{r^3} + \frac{\dot{p}}{cr^2}\right)\cos\theta\,, \tag{6.3.2}$$

$$E_\theta = \left(\frac{p}{r^3} + \frac{\dot{p}}{cr^2} + \frac{\ddot{p}}{c^2 r}\right)\sin\theta\,, \tag{6.3.3}$$

$$H_\psi = \left(\frac{\dot{p}}{cr^2} + \frac{\ddot{p}}{c^2 r}\right)\sin\theta\,, \tag{6.3.4}$$

where p, $\dot{p}$ and $\ddot{p}$ correspond to polarizability and its time derivatives.

The field emitted by the dipole can then be determined from these equations. By calculating the average flux of the Poynting vector through a sphere extending around the dipole, it can easily be demonstrated that only the $1/r$ terms have a non-null contribution. Therefore, the other terms represent the *evanescent waves* associated to the dipole.

In the case of a dipole, the *near field region* corresponds to the region of space where evanescent waves are in the majority. This region can therefore be determined using the previous relations (6.3.2-6.3.4). This leads to an estimation of the near field distance in the order of $\lambda/2\pi$. Everything happens as if the energy associated with evanescent waves is periodically flowing out from and back to the

source without ever being lost by the system. Thus this radiation is evanescent because it could not be detected in the far field. *Far field* is the field separated from the dipole at the distance much more than λ.

For the purpose of demonstrating the existence of evanescent waves associated with the dipole, in 1913 Selenyi conducted the following experiment [7]. Having deposited fluorescent molecules on the plane surface of a semicylindrical prism, he then proceeded to measure their emitted power as a function of the emission angle. Since the signal detected in the prism above the critical angle (fig. 6.3.2) was found to be non-null, Selenyi concluded that the signal was generated by the evanescent waves associated with the fluorescence emission.

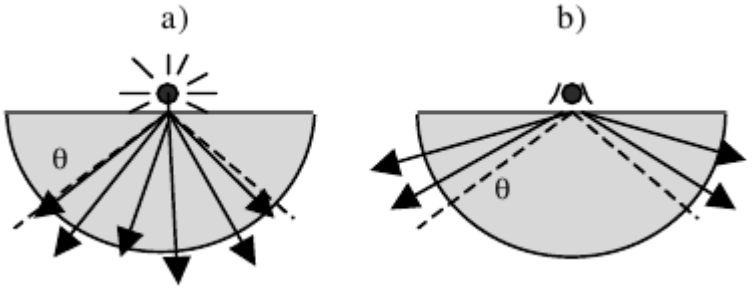


Fig. 6.3.2 – Selenyi's experiment, demonstrating the presence of evanescent waves in the vicinity of a dipole. The delected light comes from (a) waves associated with the dipole, (b) evanescent waves

If the near environment extending around a dipole can transform some of the evanescent waves into propagative waves, this means that the emission rate of the dipole increases. In other terms, the lifetime of the dipolar radiation/emission can be said to decrease. An alternative approach to coupling between the dipole and its environment consists of considering that the dipole acts as a probe in its near field. We shall return to this effect later, considering the so called surface-enhanced phenomena.

6.3.2 Light diffraction by a sub-wavelength aperture

The following fig. 6.3.3 summarizes what happens when the size of an aperture illuminated by a plane wave is reduced. The parameter to be taken into consideration is the relative size of the aperture equaled to 2*a* with respect to the wavelength of the incident wave equaled to λ.

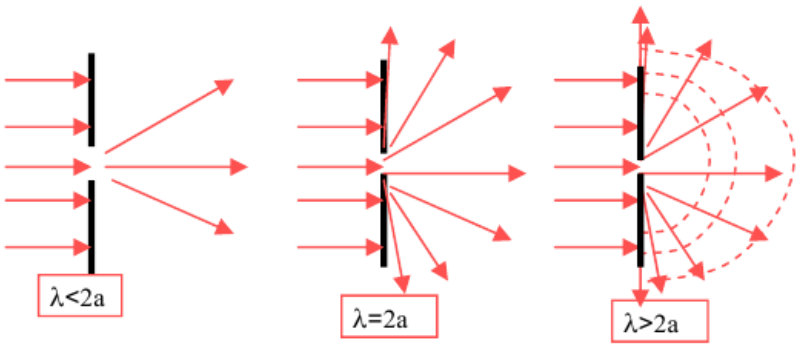


Fig. 6.3.3 – Schematic representation of the transmission of light through an aperture with variable size illuminated by a plane wave. The curves in dotted lines represent evanescent waves [8]

As the size of the aperture decreases, the numerical aperture of the transmitted beam increases up to the point, at $\lambda = 2a$, where it completely fills the half-space. From this value onwards, evanescent waves appear.

The equation for the evanescent field when $\lambda > 2a$ is [9]:

$$E = E_0 \exp(-z/d_p)\exp(i\omega t - k_x x - k_y y), \tag{6.3.5}$$

where d_p is the penetration depth of the evanescent waves, as given by the following equation:

$$d_p = \left(k_x^2 + k_y^2 - \frac{\omega^2}{c^2}\right)^{-1/2}. \tag{6.3.6}$$

The idea of using the field generated by a sub-wavelength aperture for enhancing the resolution power of microscopes was first suggested by Synge in a letter written to Einstein at the beginning of the twentieth century [9]. However, it was only much later that these ideas were actually implemented in optics [10]. Applications to near field microscopy were developed on the basis of the studies conducted on diffraction by a sub-wavelength aperture [11, 12]. The implementation of this idea will be discussed below.

6.3.3 Total internal reflection

Let us consider an ensemble consisting of two semi-infinite media, with refractive indices n_1 and n_2, where $n_1 > n_2$. Depending on the illumination conditions (fig. 6.3.4), either refraction (if $\Theta < \Theta_c$) or total internal reflection (if $\Theta > \Theta_c$) will occur. Here Θ_c is the critical angle of refraction, as given by the equation:

$$\theta_c = \arcsin(n_2/n_1). \tag{6.3.7}$$

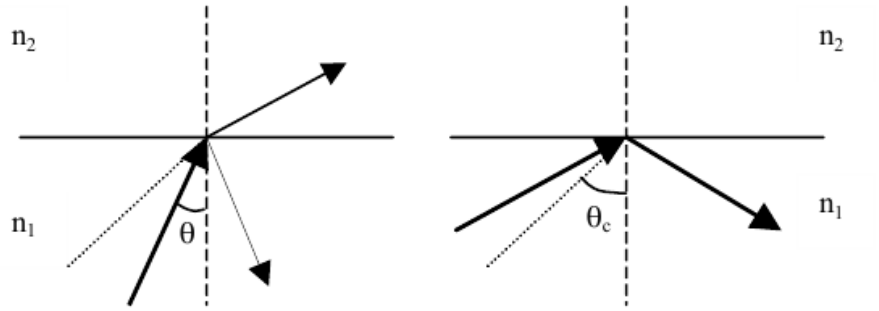


Fig. 6.3.4 – Schematic representation of the refraction phenomenon for $\Theta < \Theta_c$ and $\Theta > \Theta_c$.

Assuming *Oz* to be the normal axis with respect to the interface between the two media, and the incidence plane to be defined *Ox* and *Oz*, then the field in the second medium obeys the following equation:

- in p-polarization:

$$\vec{E}_p(z) = E_p^i \frac{(2\cos\theta)\exp(-z/d_p)}{n^2\cos\theta + j(\sin^2\theta - n^2)^{1/2}}[-j(\sin^2\theta - n^2)^{1/2}\vec{e}_x + \sin\theta\vec{e}_z] \qquad (6.3.8)$$

- in s-polarization:

$$\vec{E}_s(z) = E_s^i \frac{(2\cos\theta)\exp(-z/d_p)}{\cos\theta + j(\sin^2\theta - n^2)^{1/2}}\vec{e}_y . \qquad (6.3.9)$$

The d_p term is the penetration depth of the evanescent field in the second medium.

$$d_p = \frac{\lambda}{2\pi\sqrt{n_1^2\sin^2\theta - n_2^2}} \qquad (6.3.10)$$

d_p depends on the incidence angle, on the refractive indices of the two media, and on the wavelength. It is not polarization-dependent.

For a better insight into the order of magnitude of this parameter, the following table provides a few values for d_p in different configurations.

Let us now return to the criterion that we previously determined for the near field. In the case of a plane surface illuminated under total internal reflection, we can no longer use the same definition of the near field as for a dipole or an aperture. Indeed, the possibility may now exist that Θ is close to Θ_c and that the evanescent wave has such a great spatial extension (with d_p assuming a high value) that only this evanescent wave exists. In such a case, the value for d_p is to be considered as given and defines the spatial extension of the near field. One well-known example of penetration of light through the dielectric separation boundary is Newton's rings.

All three the above-mentioned effects are known a long time and are well described in textbooks on optics. A common characteristic inherent to these effects is the evanescent field. But the evanescent field can not always be regarded as the near field. However, the evanescent field can be determined by the following characteristics. Firstly, the orientation of the dipole moment vector in the medium depends on the phase of the incident radiation. Second, the penetration length depends on the wavelength of incident radiation (expression 6.3.10, table 6.3.1). Finally, in the third, despite the fact that the evanescent fields have limited penetration depth the effects they produce, can be detected in the far field by total internal reflection, for example.

Table 6.3.1. Value of the penetration depth of the evanescent wave for different values of refractive indices of the media, different incidence angles and different wavelengths

λ, nm	n_1	n_2	Θ_c	Θ	d_p, nm
1,300	Glass	Air	43.3	45	825
1,300	Silicon	Air	16.9	45	94
633	Glass	Air	43.3	45	402
633	Glass	Air	43.3	85	96
633	Glass	Water	65.8	85	173
414	Glass	Air	43.3	85	63

There are, however, the effects associated with the optical near field, which are the exclusive prerogative of nanophotonics and therefore, until recently, have not been described in textbooks on optics. They became possible owing to overcoming the diffraction limit and progress in nanotechnology. This is, undoubtedly the near-field microscopy, an idea of which, as mentioned above, has expressed Synge in the early twentieth century. The properties of the optical near-field, due to the interaction with the nanoscale aperture or quantum dots, are fundamentally different from the evanescent fields (table 6.3.2).

Let us consider how to overcome the diffraction limit.

Table 6.3.2. Comparison of an evanescent wave and an optical near field [13]

Properties	Evanescent Wave	Optical Near Field
Alignment of electric dipole moments	Depends on the spatial phase of the incident light	Depends on the size, conformation, and structure of the particle
Decay length	Depends on the wavelength of the incident light	Depends on the size of the particle
Generated propagating light	Reflected light (total reflection)	Scattered light

6.4 Overcoming the diffraction limit and optical near-field microscope.

Let us consider how to overcome the diffraction limit is carried out. According to Heisenberg's uncertainty principle the area of photon localization or the accuracy of determining the location of the light source Δx is related to the photon momentum uncertainty $\Delta p_x = \hbar \Delta k_x$, where p_x and k_x are the components of the photon momentum and wave vector. The uncertainty of momentum increases as decreasing the Δx value. Thinking that the momentum uncertainty should not exceed the value of the momentum we find that $\Delta x \geq \lambda / 2$.

This limit is known more as a Rayleigh criterion $\Delta x = 0.61 \cdot \lambda / \sin \varphi$, where Δx - half-width of the main peak of the diffraction of light distribution, φ - aperture angle. At $\varphi \to \pi / 2$ $\Delta x \to \Delta x_{\min} = 0/61 \cdot \lambda$. Thus, the *diffraction limit* is the minimum possible size of a light source, which can be detected with the device having, as a rule, the aperture diaphragm. As can be seen from the above estimates the value of the diffraction limit depends only on the wavelength of the light, but not the size of the aperture. Therefore, the electron microscope has a spatial resolution of the order of the de Broglie wavelength of electrons and optical microscope - the order of the wavelength of light in the visible region.

A revolutionary breakthrough in the technique of microscopy and in a number of modern technology was the invention of a scanning near-field optical microscope (SNOM), which can rightly be called no longer a microscope but nanoscopy, because the resolution of this microscope comes to the nano level.

Suppose that it is possible to change the components of the photon wave vector in such a way that $k_y = 0$, and $k_z = i\gamma$, where the imaginary components of the k_z value indicates the decay of the field, γ^{-1} value characterizes the depth of

the field decay. In other words, the field strength component of decaying wave described by the expression

$$E_z = E(0)\exp(-\gamma z). \qquad (6.4.1)$$

In this case the wave vector component k_x can be represented as

$$k_x = \sqrt{k^2 - k_z^2} = \sqrt{k^2 + \gamma^2}. \qquad (6.4.2)$$

Equation (6.4.2) shows that can be performed the relation $k_x > k$. As $\gamma \to \infty$ the range of of admissible values k_x increases indefinitely, and the value Δx can be as small as desired [14].Thus to penetrate into the region, where $\Delta x < \lambda$, is necessary to locate probe (fig.6.4.1) within the evanescent field, namely at $z << \lambda$. The value γ^{-1} characterizes the evanescent wave penetration depth and it is an order of magnitude comparable to the size of subwavelength scatterer. In particular [14], for the diaphragm with radius of *a* in a thin conducting screen $\gamma^{-1} \approx 2a$.

Figure 6.4.1 shows a scheme of a near-field optical fiber probe [14].

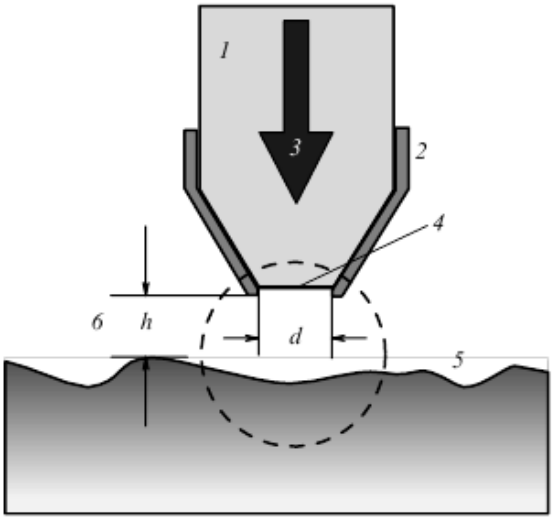


Fig. 6.4.1 – The scheme of near-field optical fiber probe: 1 - pointed optical fiber, 2 - metal coating, 3 - passing through the probe radiation. 4 - probe output aperture ($d << \lambda$), 5 - the test sample surface. 6 - the distance between the sample surface and the probe aperture ($h << \lambda$).The dotted line is delineated the near-field area [14].

Part of the light flux propagating along the fiber, the pointed tip of which is covered with metal, passes through an aperture in a metal screen and reaches the sample located in the near field (NF) source. If the distance *z* to the sample surface and the radius of the aperture *d* satisfy the condition $d, z << \lambda$, where λ - radiation wave length, the size of the light spot on the sample close to the size of the

diaphragm. When the probe is moved along the sample resolution can be realized, not limited by diffraction, or superresolution [14].

Despite the fact that the idea of realization resolution beyond the diffraction was proposed in 1928, its practical realization was made possible only with the development of technology for preparation of probes and probe microscopy techniques. Today, there are about 20 different types of NSOM which differ by optical circuit features and functionality of the probe. The near-field probe can be divided into two main groups depending on the presence or absence of the diaphragm at the end of the probe: the aperture and apertureless. The operating principle of aperture NSOM, constituting the vast majority of modern devices, explains a block diagram of the microscope shown in Fig. 6.4.2 [15].

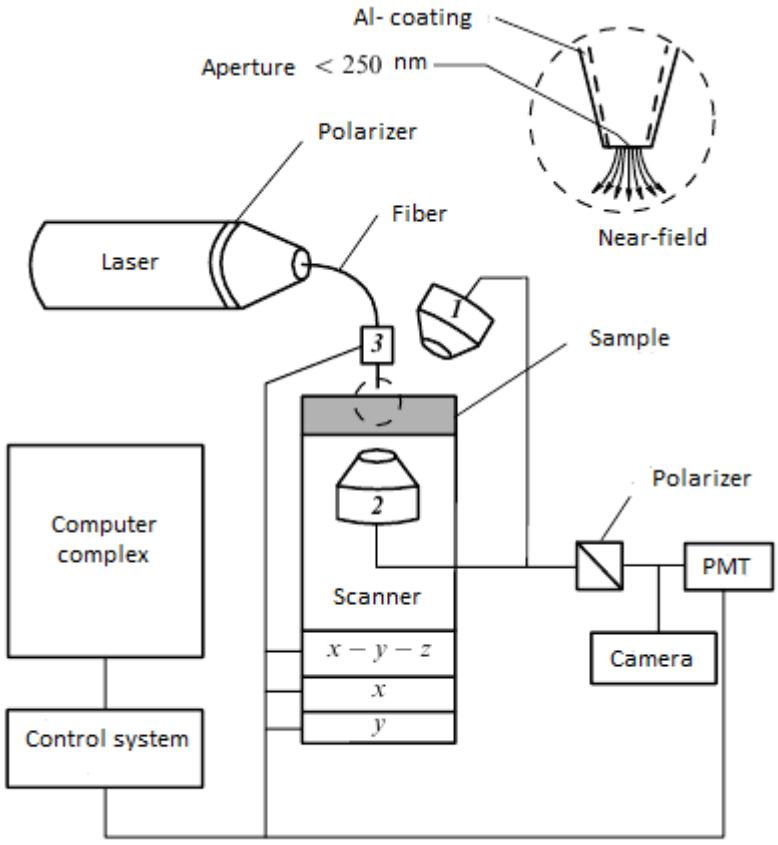


Fig. 6.4.2 – A block diagram of a near-field microscope: 1 - microlens operating in reflected light, 2 - microlens working in transmitted light, 3 - piezoengine to move the probe. The dotted line is outlined near-field contact region [15].

The laser beam through a matching element enters the pointed metallized fiber and narrows down to aperture size in its output. Reciprocal movement of the tip and the sample in three dimensions *x*, *y*, *z* is carried out by means of piezoelectric propulsive device. The photons passed through the sample or

reflected and scattered are captured by one of microscope objectives (2 or 1, correspondingly, fig.6.16), then they are sent photomultiplier. Such a microscope objective usually included in a conventional optical microscope scheme, which allows for selection of the test area and linking it to the wider field. The above scheme applies to devices operating in the *illumination mode*. There are widespread devices operating in the photon *collection mode*, when the probe carries the photons from the sample, lighting, for example, through a microscope objective, to the detector. In combined mode (illumination / collection) the probe performs both functions simultaneously.

To install the tip at the correct height above the sample [15], in all modern probe microscopes used the dependence of the intensity I of the recorded signal on z. In most types of NSOM dependence of $I(z)$ is ambiguous, since in addition to the near-field signal I_1 there is a periodic function of z signal I_2, caused by the interference of the incident and reflected waves. This makes it difficult or fully impossible the reliable control z through the value of $I=I_1 + I_2$ when approaching the edge to the sample. The best solution is to add to NSOM of auxiliary units to enable them to carry out functions as an STM or AFM, in which the definition of z does not cause any significant difficulties. In these combined devices images recording is carried out simultaneously on two channels, one of which reproduces the surface topography and other one - local distribution of refractive index. The ability to distinguish between optical and topographical contrasts greatly simplifies the interpretation of the recordable image. Most widespread z control method based on the change of the tangential component of the force of physical interaction with the sample tip (*shear force*) [16].

6.5 Theoretical approaches to the description of the optical near-field

The problem of optical near-field modeling is the subject of many books on nano-optics [17] and nanophotonics [18]. Nevetheless our knowledge in this subject for a long time remained more or less intuitive. Following to P.N. Prasad [18] the light in the near-field contains a *large fraction of nonpropagating, evanescent field*, which decays exponentially in the far field (far from the aperture or scattering metallic nanostructure). It is clear from the experiment that the enhanced electromagnetic field around the metallic tip in SNOM is strongly confined.

The various theoretical approaches of near-field optics are used in literature. Basic prerequisite of these theories is their ability to assess the evanescent electromagnetic waves The main results of the application of the various practical schemes which all rely on a numerical procedure were discussed in [19]. In general, analytical solutions can provide a good theoretical understanding of simple problems, while a purely numerical approach can be applied to complex structures. A compromise between a purely analytical and a purely numerical approach is the multiple multipole (MMP) model. This method is carefully decribed in [20] and applied to near-field optics in [21].

6.5.1 Multiple multipole method

With the MMP model, the system being simulated is divided into homogeneous domains having well-defined dielectric properties. Within individual domains, enumerated by the index *i*, the electromagnetic field $f^{(i)})(\vec{r},\omega_0)$ is expanded as a linear combination of basis functions

$$f^{(i)}(\vec{r},\omega_0) \approx \sum_j A_j^{(i)} f_j(\vec{r},\omega_0),$$

where the basis functions $f_j(\vec{r},\omega_0)$ are the analytical solutions for the field within a homogeneous domain. These basic functions satisfy the eigenwave equation for the eigenvalue q_j :

$$-\nabla\times\nabla\times f_j(r,\omega_0) + q_j^2 f_j(r,\omega_0) = 0.$$

MMP can use many different sets of basis fields, but fields of multipole character are considered the most useful. The parameters *A* are obtained by numerical matching of the boundary conditions on the interfaces between the domains.

Figure 6.5.1 (left) shows the results of 3D near-field calculations perfomed with MMP method by Nivotny and Pohl [21] for the tip of an aperture SNOM consisting of a cylindrical part and a tapered part. The probe is excited by a waveguide mode of wavelength λ = 488 nm with polarization of $\vec{E}$ vector in plane of the picture (fig. 6.5.1, left) and in orthogonal to the picture plane (fig. 6.5.1, right). There is a factor three between two successive lines on the fig. 6.5.1. Figure 6.5.2 presents contours $|E|^2$ on three perpendicular planes near the aperture of the SNOM probe described in fig. 6.5.1. The arrows indicate the time-averaged Poynting vector. The polarization is in the plane y = 0. The transmission through the probe is increased when a dielectric substrate (ε =2,25) is approached (fig.6.5.2, *b*).

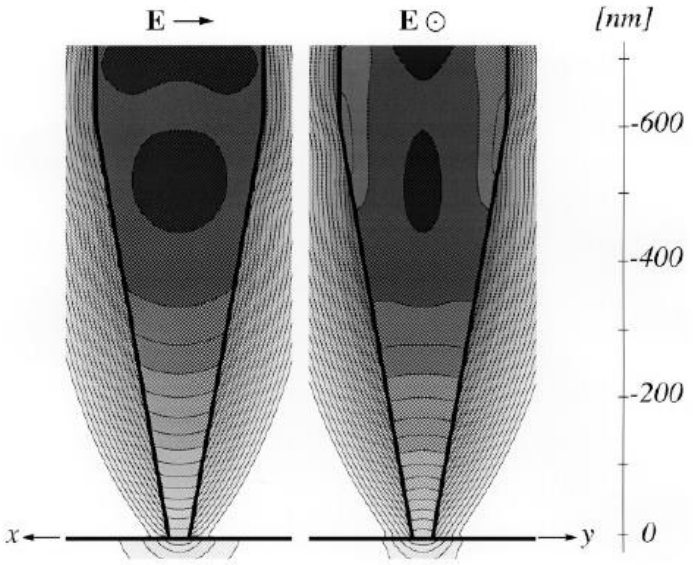


Fig. 6.5.1 – Examples of 3D near-field calculations performed with the MMP method (Research gate www source)

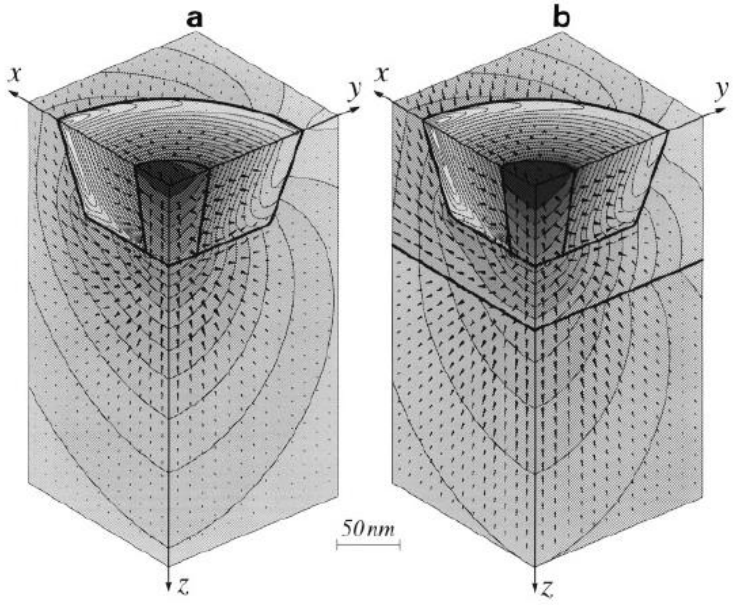


Fig. 6.5.2 – Contours of constant $|E|^2$ on three perpendicular planes near the aperture of the SNOM probe described in fig. 6.5.1 (Research gate www source)

Figure 6.5.2 shows the field in the region of the aperture probe hole in vacuo above the dielectric substrate [22**Error! Bookmark not defined.**]. Coating tapers in the direction of the hole (aperture), and its total thickness is 70 nm. The value of the aperture diameter is selected to be 50 nm. The field enhancement is observed in the polarization plane of the wave (y = 0) at the borders of coating, due to the high value of the field component perpendicular to the border, as well

as the highly curved geometry of the problem (the lightning rod effect). In the plane perpendicular to the polarization plane (x = 0), the electric field is everywhere parallel to the border, so that the level lines are continuous.

Part of the field penetrates through the aperture in the metal border, thereby increasing the effective width of the aperture. When entering the dielectric substrate into the area the probe its rate of energy transfer increases. We can verify this by looking at fig. 6.5.2 and comparing line level probe. Part of the radiated field is dissipated in the space surrounding the probe, transferring energy into external spatial modes propagating back on its shell surface.

External spatial modes can be excited in the forward by the field passed from the probe core to the shell. By analogy with the cylindrical waveguide, they practically do not undergo attenuation [22]. Most of the energy contained in these modes, thus extends in the direction of aperture plane. If the selected shell is too thin, it may happen that the light from the shell will be stronger than the light that came from the core and radiating opening. In this case, the field markedly enhanced by external shell boundaries, resulting in a field distribution shown in fig. 6.5.3 (right). To avoid such undesirable situations, it is necessary to choose a rather thick shell. A reasonable solution to the shell thickness problem may be the use of shells, coming down on the cone.

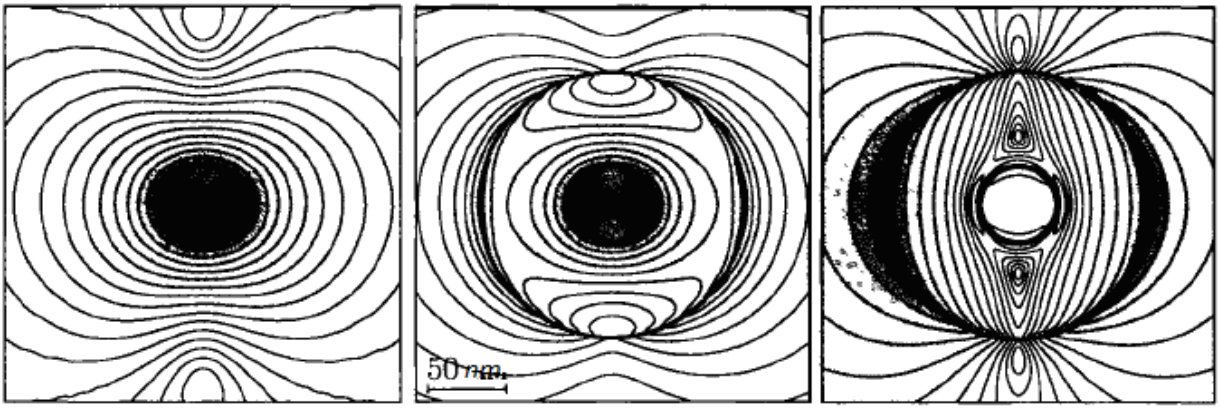


Fig. 6.5.3 – Electrical field strength lines $|E|^2$ (scaling between adjacent lines $3^{1/2}$) in the plane of the aperture opening for the three different probes with different coating thickness. Infinite thickness of the shell shown on the left. In the center: finite thickness; the field radiated by the surface of hole dominates. On the right, finite thickness, the field radiated by the outer surface of the shell dominates.

It must be emphasized that the surface modes cannot be excited by external radiation, since their propagation constant larger than propagation constants of freely propagating light type surface plasmons.

The multipole functions used in the basis set of the MMP method are rather short range so that they affect their close neighbourhood. The method is thus better suited to account for localized geometries than the expansion in plane waves. It is also well designed to describe complicated structures such as cylindrical waveguides coated with a realistic metal whose dielectric function is complex or the scattering inside sharpened cylindrical tips used in near-field microscopy. From a mathematical point of view, the idea of spreading a set of multipole functions and adjusting the coefficients is similar to quantum mechanical techniques used for computing electronic structures such as the linear combination of atomic orbitals (LCAO). The multipoles in electrodynamics thus play a role similar to the atomic orbitals in quantum mechanics. However, the situation is somewhat clearer in electron physics where the atomic orbitals are centred on each nucleus. Whereas a physical meaning is lacking in the mathematical procedure [23] which distributes the coordinates of the multipoles centres in electrodynamics.

6.5.2. Physical picture of near-field interactions

The fundamental difference of evanescent waves from the optical near-field, as shown above in Table 6.3.2, as well as some disadvantages in the near-field notation, which can hardly be called as photons and waves, led the Japanese scientist Motoichi Ohtsu to a very successful model, which he called "*dressed photon*» [24]. This model is very suitable for calculating the interaction between nanostructures and evanescent light waves. One of these effects have been shown in the figure 6.18 (on right), it is the enhancement of the evanescent field arising from the NSOM probe when approaching the dielectric substrate.

For this investigation, Motoichi Ohtsu has been introducing an approach that has never been described in conventional textbooks on optics. The main scope of conventional optics textbooks has been a comparison between the classical and quantum features of light.

Dressed photon (DP) science and technology exploits the electromagnetic field localized in nanometric space i.e. non-propagating near-field wave. The principles and concepts of DP science and technology are quite different from those of conventional wave-optical technology encompassing photonic crystals, plasmonics, metamaterials, silicon photonics, and quantum-dot photonic devices. This is because these devices use propagating light even though the materials or particles used may be nanometer-sized.

The theoretical picture of DP has been proposed by M. Ohtsu to describe the electromagnetic interactions between nanometric particles located in close proximity to each other. Follow to M. Ohtsu, the DP is a virtual cloud of photons that always exists around an illuminated nanometric particle. Its energy fluctuation, δE, and duration of the fluctuation, τ, are related by the Heisenberg uncertainty relation, $\tau \cdot \delta E \cong \hbar$. From this relation, the linear dimension of the virtual cloud of photons (the virtual photons for short) is given by $r \cong c\tau \cong hc/\delta E$, where c is the speed of light. In the case of visible light illumination (photon energy ~2 eV), r is estimated to be about 100 nm. This means that the effect of the virtual photons at the surface of the illuminated particle is important if the particle is smaller than 100 nm. In other words, the optical properties of sub-micron-sized matter are not free from the effects of virtual photons.

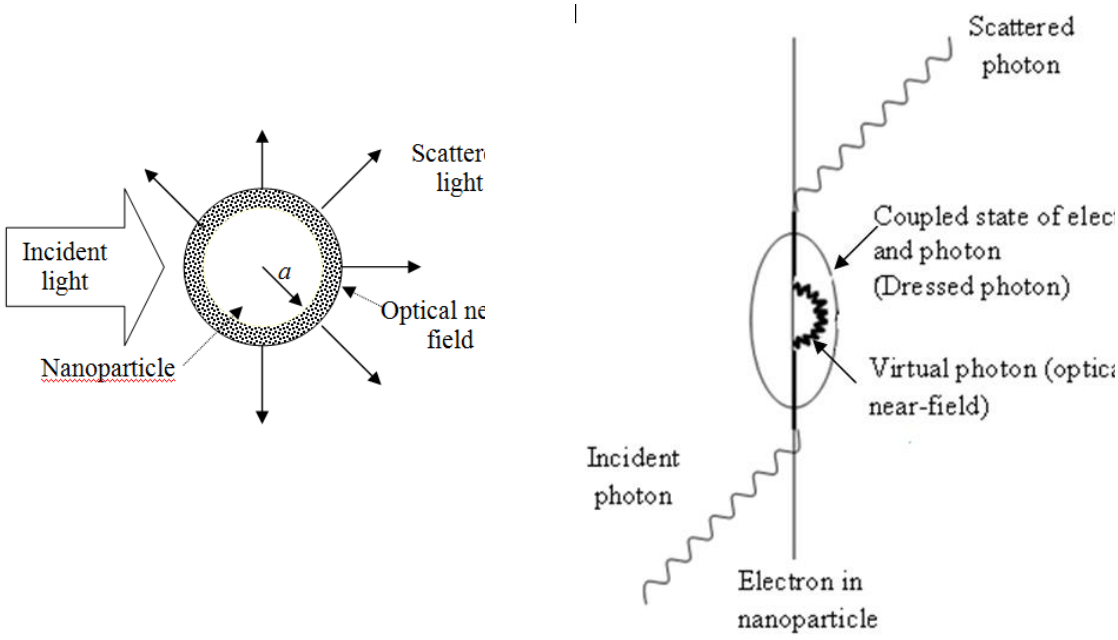


Fig. 6.5.4 – Feynman diagram (right) representing the generation of optical near-fields near the nanoparticle (left) [25]

This feature of the optical near-field, i.e., the virtual photons, can be most appropriately described by using a Feynman diagram (Fig. 6.5.4), popularly employed for elementary particle physics. In this figure, a photon is emitted from an electron in the illuminated nanometric particle and can be re-absorbed within a short duration. This photon is nothing more than a virtual photon, and its energy is localized at the surface of the nanometric particle. Independently of this virtual photon, a real photon (also called a free photon) can also be emitted from the electron. This photon is conventional propagating scattered light. Since the virtual photon remains in the proximity of the electron, it can couple with the electron in

a unique manner. This coupled state, called a dressed photon, is a quasi-particle from the standpoint of photon energy transfer and has applications to novel nanophotonic devices and fabrication technologies. It is the dressed photon, not the free photon that carries the material excitation energy. Therefore, the energy of the dressed photon, $h\nu_{DP}$, is larger than that of the free photon, $h\nu_{free}$, due to contribution of the material excitation energy.

Independently of the real photon (i.e., conventional propagating scattered light), a virtual photon is emitted from the electron, and this photon can be re-absorbed within a short duration (fig. 6.5.4).

To detect the dressed photon, a second nanometric particle should be placed in close proximity to the first particle to disturb the dressed photon on the first particle. This disturbance generates a free photon, which is propagating scattered light that can be detected by a conventional photodetector installed in the far field. This detection scheme suggests that the dressed photon energy is exchanged between the two particles. That is, after a dressed photon is generated in the first particle, it is transferred to the other particle. It can be transferred back to the first particle again, which means that the dressed photon can be exchanged between the two particles. The detectable free photon is generated in the process of this exchange.

The DP has been theoretically described by assuming a multipolar quantum electrodynamic Hamiltonian in a Coulomb gauge in a finite nano-system [25]. The creation and annihilation operators of the DP are expressed as the sum of the operators of the real photon and an electron-hole pair. After a unitary transformation and some simple calculations, its annihilation and creation operators are respectively expressed, in the lowest order, as

$$\tilde{a}_{\vec{k}\lambda} = a_{\vec{k}\lambda} - iN_{\vec{k}} \sum_{\substack{\alpha>F \\ \beta<F}} \{\rho^*_{\beta\alpha\lambda}(\vec{k})A^*_{\alpha\beta} + \rho^*_{\alpha\beta\lambda}(\vec{k})A_{\alpha\beta}\}, \tag{6.5.1}$$

$$\tilde{a}^*_{\vec{k}\lambda} = a^*_{\vec{k}\lambda} + iN_{\vec{k}} \sum_{\substack{\alpha>F \\ \beta<F}} \{\rho_{\beta\alpha\lambda}(\vec{k})A^*_{\alpha\beta} + \rho_{\alpha\beta\lambda}(\vec{k})A_{\alpha\beta}\}, \tag{6.5.2}$$

where $\vec{k}$ is the wave-number of the free photon, λ is the polarization of a free photon, $N_{\vec{k}}$ is the normalization constant, α and β represent the electronic energy states above and below the Fermi energy level F, ($a_{\vec{k}\lambda}$, $\tilde{a}^*_{\vec{k}\lambda}$) and ($A_{\alpha\beta}$, $A^*_{\alpha\beta}$) are the annihilation and creation operators of the free photon and an electron-hole pair, respectively, and $\rho_{\beta\alpha\lambda}(\vec{k})$ is the Fourier transform of the spatial distribution of the transition dipole, $\rho_{\beta\alpha\lambda}(\vec{r})$ – of the nanometric particle. Based on this dressed

photon picture, interactions between the nanometric particles can be simply described by emission, absorption, and scattering of dressed photons, which provides a physically intuitive picture of the optical near-field interaction between the two particles. The real system is more complicated because the nanometric subsystem (composed of the two nanometric particles and the dressed photons) is buried in a macroscopic subsystem composed of the macroscopic substrate material and the macroscopic incident and scattered light fields. A novel theory was developed to avoid describing all of the complicated behaviors of these subsystems rigorously, since we are interested only in the behavior of the nanometric subsystem. In this theory, the macroscopic subsystem is expressed as an exciton-polariton, which is a mixed state of material excitation and electromagnetic fields. Since the nanometric subsystem is excited by an electromagnetic interaction with the macroscopic subsystem, the projection operator method is effective for describing the quantum mechanical states of these systems [26]. As a result of this projection, the nanometric subsystem can be treated as being isolated from the macroscopic subsystem, where the magnitudes of effective interaction energy between the elements of the nanometric subsystem are influenced by the macroscopic subsystem. This local electromagnetic interaction can take place within a sufficiently short duration in which the uncertainty relation allows the exchange of dressed photons nonresonantly, as well as the exchange of a free photon resonantly. The interaction due to the non-resonant process is expressed by a screened potential using a Yukawa function $\exp(-r/a)\cdot r^{-1}$, which represents the localization of the optical near-field around the nanometric particles. Its decay length *a* is equivalent to the particle size [26], which means that the extent of localization of the optical near-field is equivalent to the particle size, as was described above. On the other hand, the interaction due to the resonant process is expressed by a conventional spherical wave function $\exp(-ir/a)\cdot r^{-1}$, which is mediated by a free photon (a conventional propagating field).

Because the extent of localization of the dressed photon is equivalent to the nanometric particle size, the long-wavelength approximation, which has always been employed for conventional light-matter interaction theory, is not valid. This means that an electric dipole-forbidden state in the nanometric particle can be excited as a result of the dressed photon exchange between closely placed nanometric particles, which enables the operation of novel nanophotonic devices. A real nanometric material is composed not only of electrons but also of a crystal lattice. In this case, after a dressed photon is generated on an illuminated

nanometric particle, its energy can be exchanged with the crystal lattice, as shown by the Feynman diagram of fig. 6.5.5, *a*. By this exchange, the crystal lattice can excite the vibration mode coherently, creating a coherent phonon state. As a result, the dressed photon and the coherent phonon can form a coupled state, as is schematically explained by fig. 6.5.5, *b*. The creation operator $\widehat{a}_i^*$ - of this novel form of elementary excitation is expressed as

$$\widehat{a}_i^* = \tilde{a}_i^* \exp\left\{-\sum_{p=1}^{N} \frac{\chi_{ip}}{\Omega_p}(b_p^* - b_p)\right\}, \tag{6.5.3}$$

where $\tilde{a}_i^*$ is the creation operator of the dressed photon (refer to (6.5.2)) localized on the *i*- th site of the crystal lattice, *N* is the number of sites, χ_{ip} is the phonon-photon coupling in mode *p* at site *i*, and Ω_p is the eigen-frequency of the phonon mode *p*. The exponential function in this equation is called a displacement operator, composed of the phonon creation and annihilation operators ($b_p^* - b_p$).

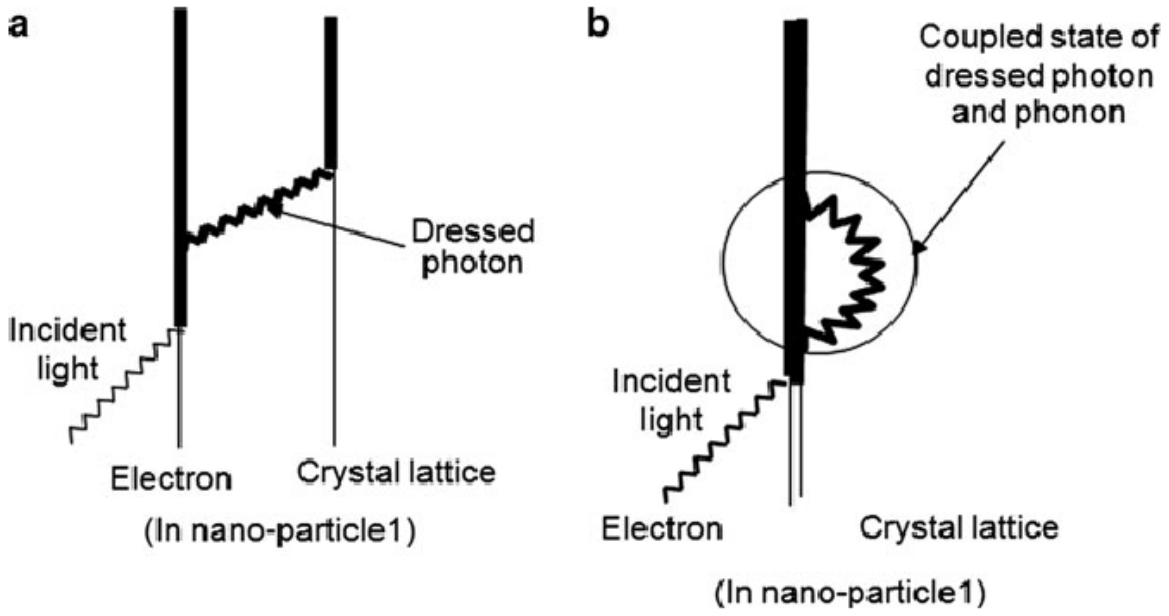


Fig. 6.5.5 – Feynman diagrams representing the coupling of a dressed photon with phonons. (a) Generation of a dressed photon and exchange with the crystal lattice, (b) A coupled state of a dressed photon and a coherent phonon

This is an operator representing the creation of the coherent phonon state [27, 28]. This coupled state (the dressed photon carrying the coherent phonon energy (DP-CP)) is a quasi-particle and is generated only when the particle size is small enough to excite the crystal lattice vibration coherently. If not, the vibration is incoherent, and thus, its energy is dissipated, heating the particle. It is easily understood that the energy of the DP-CP, $h\nu_{DP-CP}$ is higher than that of the dressed

photon. It is also higher than the free photon energy, $h\nu_{FP}$, incident on the nanometric particle. The relation between these energies is represented by

$$h\nu_{FP} < h\nu_{DP} < h\nu_{DP-CP}\,. \qquad (6.5.4)$$

Three examples for application to energy conversion were published by Ohtsu group: (1) Up-conversion of optical energy [29]. (2) Conversion from optical to electrical energy [30]. (3) Conversion from electrical to optical energy by taking a LED as an example. Novel LED [31] and a laser [32] will be demonstrated by using indirect transition-type semiconductors (Si, SiC, and GaP), which are fabricated and operated by dressed photon − coherent phonon novel technology.

In conclusion, it should be noted that a strict solution in quantum field theory was obtained for the case of a static source by the authors of [33].

6.6 Molecular electronics and photonics devices

We express our sincere gratitude to the Research Supervisor of the Southern Federal University (Rostov-on-Don, Russia), Academician of the Russian Academy of Sciences Vladimir Isaakovich Minkin, and Valery Alexandrovich Barachevsky, Head of the Laboratory of the Photochemistry Center of the Russian Academy of Sciences (Moscow, Russia) for the information provided.

6.6.1 Basic concepts of molecular electronics and spintronics

For decades, molecular and supramolecular systems with their discrete energy levels and the ability to switch the molecular system from one state to another have been the prototype of the ideal element base of computing devices. The complex and faultless electronic processes, including photosynthesis and the transmission of neural signals, occur on the molecular level, and the choice of nature is optimal [34].

Estimates based on the principles of the theory of molecular structure and confirmed by experimental studies of a large number of molecules, show that in comparison with the semiconductor element base the molecular elements could provide [35]:

- a higher degree of integration,
- significantly lower switching energy,
- higher circuitry stability with respect to ionizing radiation.

This molecular element base could lead to the following fundamentally new capabilities:

- full identity of the molecular elements, that are not subject to scatter due to the inevitable technological errors,
- noise-free one-electron processes,
- specific molecular signal transmission processes that may enable to create a more complex logic elements.

Only the possibility of using molecular systems in the computing and information-logical devices will be the subject of further consideration.

Broadly, molecular electronics can be defined as the scope of the molecules and molecular materials capable to receive, store and process information (Figure 6.6.1) [34, 36].

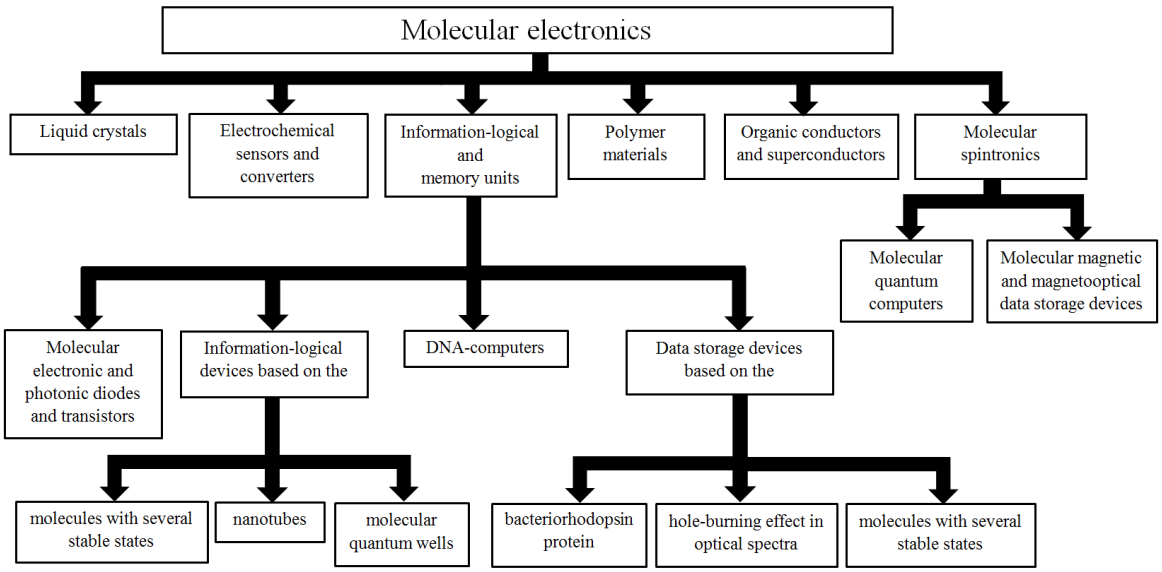


Fig. 6.6.1. Basic fields of molecular electronics and photonics

Creation of a wide range of switches and nanowires has provided the possibility of formation of molecular logic devices on their basis. Molecular photonics is studying a complex of photophysical and photochemical processes in electronically excited molecular systems.

Molecular materials have a number of important advantages in comparison with traditional makromaterials. First of all, the molecular structures have a wide range of optical, electrical and magnetic properties, and also possibilities of optimum “adjustment” of these properties. Secondly, the molecular self-assembly and molecular recognition mechanisms can be used to designing molecular devices.

Molecular electronics is closely related with the rapidly developing field of molecular spintronics - electronics, in which the data carrier is not the electron charge but its spin [37].

The architecture of a molecular computer is expected to be similar to the architecture of classical computers based on semiconductor technology. The main components are the same: switches, memory, communication lines (wires). However, in the molecular computer all these components are based on functionalized molecular structures (Figure 6.6. 2).

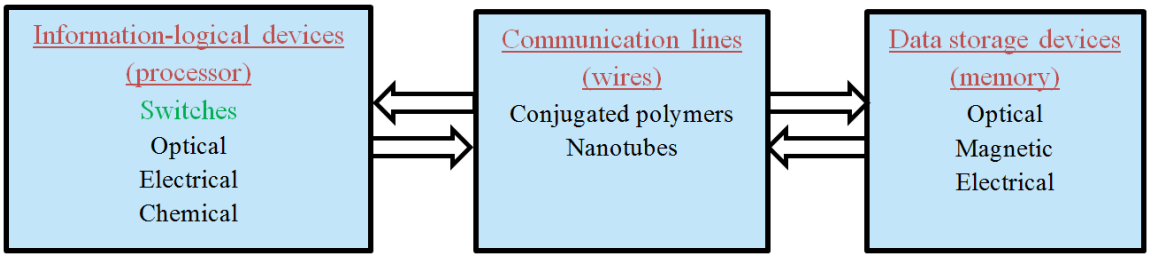


Fig. 6.6.2. The main components of molecular computer.

"Smart" bistable molecular and supramolecular structures will be the basis of the elemental base molecular computers. The bistability (polystability) is the possible of existence of two (or more) thermodynamically stable states, which correspond to the local minima on the potential energy surface (Figure 6.6.3.). In terms of computer science concepts such structures correspond to a logical zero (0) and one (1), and the reconfiguration of the bistable structures corresponds to information transfers between zero and one. Switching between the stable states of bistable molecular structures and logical elements on their base is performed through a variety of external actions (light and electrical impulses or electrochemical reactions) and is accompanied by changes of the physical properties (Figure 6.6.4). Table 6.6.1 shows the main parameters of modern semiconductor (silicon) computers and the corresponding operating capabilities of the hypothetical molecular computers [38-41].

Table 6.6.1. Comparison of characteristics of modern semiconductor and hypothetical molecular computers [34].

Parameter	Computer	
	semiconductor	molecular
Transistor size, нм	45 nm	1-10
Number of transistors per cm^2	10^9	~10^{13}
Response time, s	<10^{-15}	<10^{-12}
Performance	1	10^{11}

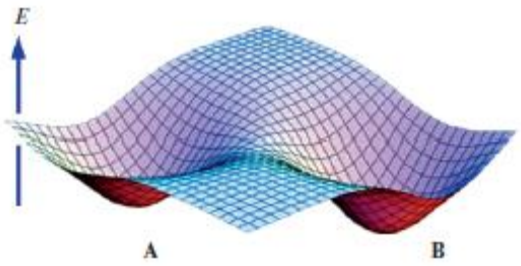


Fig. 6.6.3. The potential energy surface of bistable molecular structure that implements reversible reconfiguration between isomeric forms **A** and **B** by external factors [34].

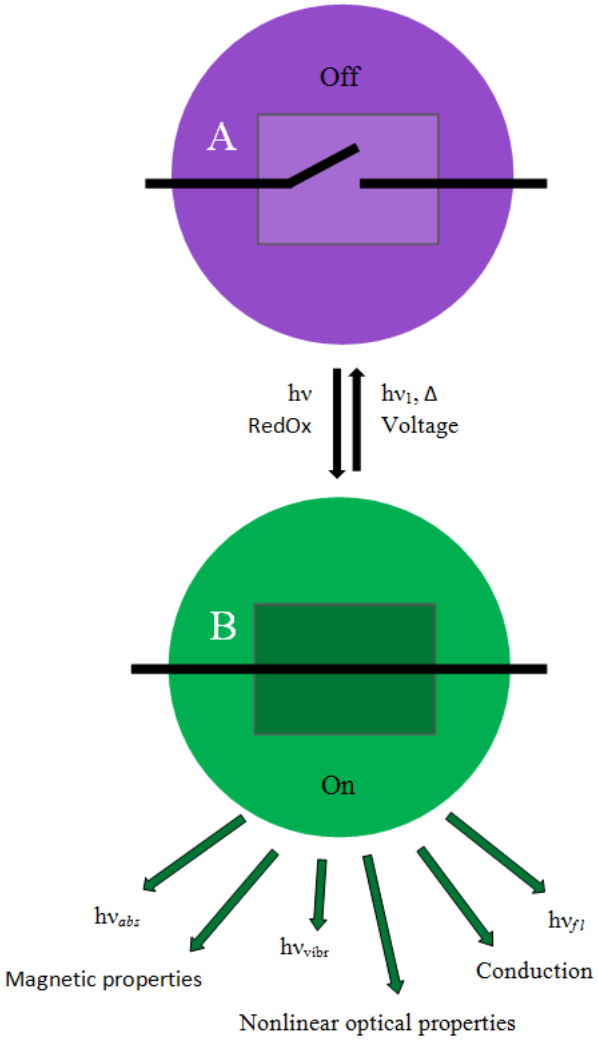


Fig. 6.6.4. Switchable properties of bistable structures

6.6.2 Introduction to the theory of bistable molecular systems

Most often switching of bistable molecular systems is carried out by a light signal, and bistable molecular systems have photochromic properties. The molecules and supramolecular formation undergoing reversible reconfigurations under light exposure are the part of photochromic systems [42, 43]. The photochromism is usually interpreted as reversible transformation of substance from one state to another, which occurs in at least one direction under the light exposure and accompanied by changes in optical and other physical and chemical characteristics of the substance [44].

The photochromic cycle in generally looks like a transition of a substance under the exposure of light quantum $h\nu_1$ from the state A with the absorbance at a wave length λ_1 to the state B with the absorbance at λ_2:

$$A(\lambda_1) \underset{kT,\ h\nu_2}{\overset{h\nu_1}{\rightleftarrows}} B(\lambda_2) \qquad (6.6.1)$$

The substance returns into its original state either spontaneously (due to the thermal energy kT), or under the light exposure $h\nu_2$.

Reversible phototransformations of substances accompanied by a change of the absorption spectra in the visible region are based on the phenomenon of "physical" or "chemical" photochromism. Physical photochromism is based on the transition of atoms and molecules during the absorption of light at the end time (lifetime) in the electronically excited states characterized by new absorption spectra. Chemical photochromism is connected with deep intramolecular restructurings of substance under the influence of light that lead to the temporary formation of new thermodynamically unstable forms of chemical compounds.

Unlike chemical reactions in which the molecules are activated by the redistribution of energy in the collision, the molecules in photoprocesses obtain necessary energy from an external source. Therefore the main characteristics of phototransformations of photochromic systems derive from the laws of photochemistry.

A necessary condition for the implementation of any photoprocess is the absorbtion of light by the molecule. According to the law of quantum equivalence each absorbed light quantum $h\nu$ causes a physical or chemical change of a single molecule. The number of modified molecules in unit time is proportional to the number of quanta of monochromatic radiation absorbed by molecules in the same time. Therefore the transformation rate of the system is determined by the photochromic absorption rate.

Since the excited molecules can lose their energy and return to its original unexcited state until the attaining the expected result, the absorption of light does not always lead to a transition of molecules into the expected state. Quantitative evaluation of the phototransformations effectiveness is determined by quantum yield λ_φ, under which is defined as the ratio of the generated molecules to the number of absorbed quanta in unit time:

$$\varphi(\lambda) = \frac{dn/dt}{dN/dt} \tag{6.6.2}$$

where *n* is the number of generated molecules; *N* is the number of absorbed light quanta; *t* is the time. Since the number of photons is determined by the ratio of the absorbed energy to the energy of a one quantum, the rate of formation of photoinduced molecules D in under the exposure of light λ_i absorbed the initial form A:

$$\frac{dn_B}{dt} = I_0(\lambda_i) \cdot \varphi_{AB}(\lambda_i) \cdot (1 - e^{-k_A(\lambda_i) \cdot l}) \tag{6.6.3}$$

where $I_0(\lambda_i)$ is the light quanta flux density entering at a photochromic system; $k_A(\lambda_i)$ is the absorption coefficient of the original form A; *l is* the absorption layer thickness. If the absorption bands of the original and photoinduced forms do not overlap, then the rate of return of the optical density of the photochromic system in the range of absorption band of photoinduced form (λ_j) when excited by radiation absorbed by the original form A:

$$\frac{dD_B(\lambda_j)}{dt} = CI_0(\lambda_i) \cdot \varphi_{AB}(\lambda_i) \cdot \varepsilon_B(\lambda_j)\left[1 - 10^{(-\varepsilon_A(\lambda_i) \cdot C_A \cdot l)}\right] \tag{6.6.4}$$

where $D_B(\lambda_j)$ is the optical density of the photoinduced form; *C* is a constant; $\varepsilon_B(\lambda_j)$ and $\varepsilon_A(\lambda_i)$ are molar coefficient of photoinduced and the initial forms respectively; C_A is the concentration of the original shape of the molecules.

The peculiarity of photoprocesses that form the basis of photochromism phenomenon is that under the light exposure in photochromic system occurs the energy storage, which is spent on return of photochromic material to its initial state after termination of irradiation [44]. At the physical photochromism energy is stored in the form of energy of electronically excited states of molecules involved in the act of light absorption. At the chemical photochromism the reaction products are formed, which are in the ground electronic state, but have a greater enthalpy than the original condition.

Let us define the conditions for the existence of the physical photochromism under which photochromic effects by creating a high occupation density of electronically excited molecule levels A can be observed. Imagine the Scheme (Figure 6.6.5) of the abstract closed cycle that is carried out during the transition

of substance under the light exposure ($h\nu_1$) from state A in a state B characterized by a new absorption spectrum ($h\nu_2$) and the constant of spontaneous transition rate from B to A equal to k'_T.

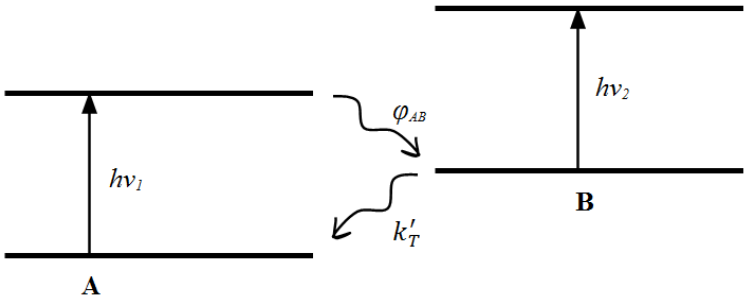


Fig. 6.6.5. Scheme of energy levels and transitions for the physical photochromism model [44].

The number of molecules (atoms) in the states A and B per unit volume is respectively n_A and n_B. Since in a system the substance is only in a state A or B, $n_A+n_B=n_0$ at any time, where n_0 is a constant. Then the rate of population of the state B is determined by the product of $I_0 \cdot \sigma_A \cdot \varphi_{AB} \cdot (n_0-n_B)$ (where σ_A is the molecular absorption coefficient, $\sigma_A=3{,}8 \cdot 10^{-21} \varepsilon_A$; φ_{AB} is the quantum yield of molecules B) . The extinction rate of the state B will be equal to $k'_T \cdot n_B$. In case of equality of formation and extinction rates of the state B, i.e. in condition of photostationary state, it can written

$$\frac{n_B}{n_0 - n_B} = \frac{\sigma_A \cdot I_0 \cdot \varphi_{AB}}{k'_T} \qquad (6.6.5)$$

Physical photochromism can be observed in conditions of a powerful light emissions in the range of photon flux densities 10^{20}-10^{29} $s^{-1} \cdot m^{-2}$. Photochromic processes caused by photochemical restructions in organic and inorganic photochromic compounds occur at much lower excitation densities.

It is known that chemical photochromism is connected with the formation of thermodynamically unstable ground electronic state of photoreaction product [44]. Photochemical processes develop in the background of thermal reactions, and photochromic cycle is complemented by spontaneous thermal process of formation of form B with a rate constant k_T. There is a thermal equilibrium between the forms A and B in the initial state in the photochromic system. The equilibrium constant K is proportional to the ratio of the rate constants of the thermodynamic processes k_T and k'_T:

$$K = \frac{[B]_{t\to\infty}}{[A]_{t\to\infty}} = \frac{k_T}{k'_T} \qquad (6.6.6)$$

where $[A]_{t\to\infty}$, $[B]_{t\to\infty}$ are the equilibrium concentrations of forms A and B established at long stay of photochromic system in the dark; k_T and k'_T are the rate constants of spontaneous thermal transitions of the form A to B and inversely. The reaction rate constant determined by the formula

$$k_T = \chi \frac{R \cdot T}{N_A \cdot h} \cdot \exp\left(\frac{\Delta S^{\neq}}{R}\right) \exp\left(-\frac{\Delta H^{\neq}}{R \cdot T}\right) \qquad (6.6.7)$$

where χ is the probability that the molecule after reaching the transition state will form reaction products; N_A is the Avogadro number; h is the Planck's constant; $\Delta S^{\neq}$ is the activation entropy, $\Delta H^{\neq}$ is the activation enthalpy.

After the experimental determination of the activation energy, it is possible to calculate the activation enthalpy, and the activation entropy can be calculated by a known absolute value of the rate constant of process k_T in the assumption χ=1. k_0 and $E_{акт}$, as well as $\Delta S^{\neq}$ and $\Delta H^{\neq}$ are equally frequently mentioned in the literature on photochromism as the parameters characterizing the thermal processes.

The correlation between forms A (thermodynamically stable) and B (thermodynamically unstable) in an unexcited photochromic system is determined by the correlation of the rate constants of the thermal processes k_T and k'_T. Thermal balance becomes disturbed, when the photochromic system is efected by the light absorbed forms A or B and transforming molecules in electronically excited states.

There are two possible ways of photochemical reactions development [44]. If during the initial form A excitation the photoinduced form B is formed in the electronically excited state (Fig. 6.6.6a), there is the adiabatic photoprocess with the smooth movement of the point representing the state of the system along the potential energy surface. Enrichment of the system with the electronically excited molecules of the form B during the lifetime of the electronically excited molecules of the form A is caused by difference between system potential energy surfaces in the electronically excited state and in the ground state.

More often there are non-adiabatic processes (Figure 6.6.6b) in photoreactions, in which the ground state of the photoinduced form B is formed from the electronically excited state of the original form A.

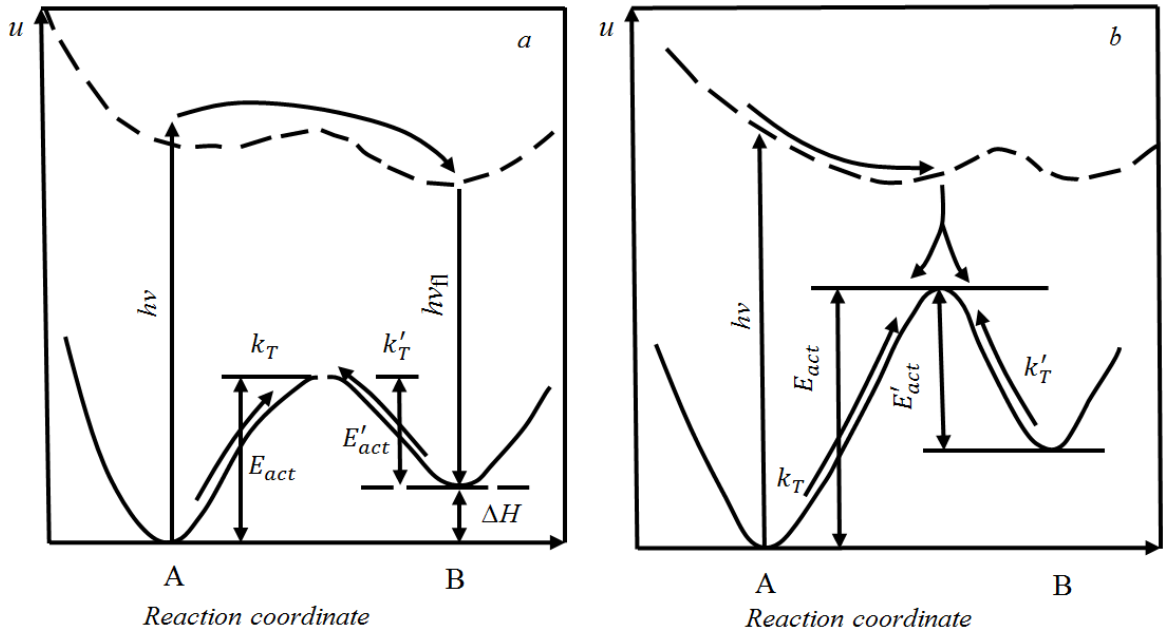


Fig. 6.6.6. Dependence of the internal energy of the photochromic system from coordinate of the reaction in the ground (solid line), electronically excited (dotted line) states and the ways of reaction for the adiabatic (a) and non-adiabatic (b) photoprocesses [44].

Thanks to the large thermodynamic potential, in the electronically excited state the original form A appears in the position of minimum potential surface of the excited state with molecular configuration that is intermediate for forms A and B. From this position, there is a rapid non-radiative deactivation of electronically excited state of form A either into the initial state of the form A, or into the ground state of the form B. The ratio of the probability of these forms occurrence depends on the relative position of the minimum of the potential surface of the electronically excited state and the maximum of the potential surface of the transition complex of the ground state. Thereby there overcome a significant activation barrier that exists in the ground state. Thus there is the overcoming of the significant activation barrier that exists in the ground state.

Thus the most important characteristics of photochromic systems that determine the effectiveness of their phototransformations are the absorption spectra of the original and the photoinduced forms of matter, quantum yields of the direct φ_{AB} and reversed φ_{BA} phototransformations, thermochromic transition rate constants [44].

These characteristics determine a specific area of practical use of photochromic materials and systems, as well as mechanisms of photochromic transformations.

6.6.3 Components for molecular electronics and photonics

Molecular switches and logical gates

Let us consider molecular switches activated with the help of light. As external switch source the light has a number of advantages: the photons have zero mass and move with the maximum high speed without release of heat; light impulses control requires less energy than the use of electrical signals and it is free from interference from electromagnetic fields. Due to the sensitivity of fluorescence detection methods, molecular switches, which one or both isomeric forms are fluoresce, are preferable.

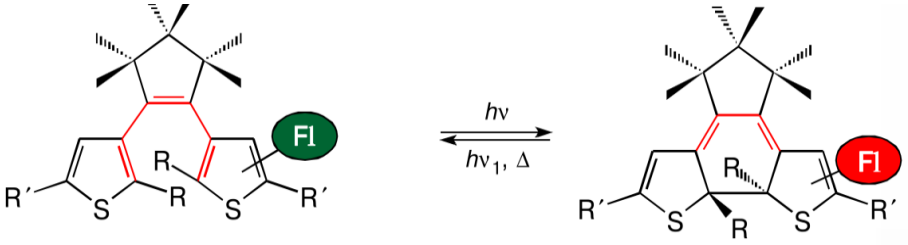
The most effective molecular switches are based on the photochromic compounds which becomes isomerized during the transition into the higher excited electronic state. This may be the process of cis–trans isomerism, pericyclic transformations, proton phototransfer. Electronic configuration of the system radically rebuilt after switching.

Spiropyrans and spirooxazines have especially high photosensitivity (high quantum yields of photoreaction, including record high values of the two-photon absorption cross-sections – more than 10^{-48} $cm^4 \bullet s \bullet photon^{-1}$) [45], the disadvantage of these compounds is relatively low thermal stability, i.e. fast reverse dark reaction that leads to uncontrolled erase of information recorded by the use of photo switches. This can be avoided with the use of photochromic systems, where adjustment between the isomers in both directions are carried out only as a photoreaction.

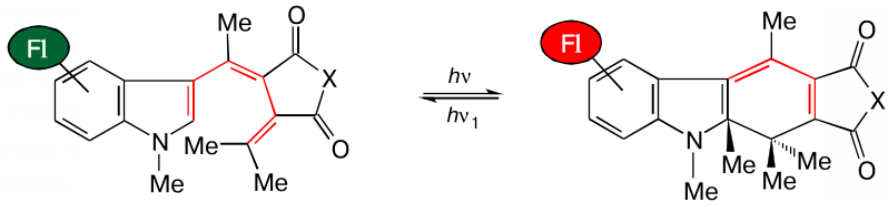
At present time there are about 20 different mechanisms of photochromic transformations, but only two basic types of photoregrouping provide high thermal barriers of direct and reverse reactions [34]. Those are, first of all, the pericyclic reactions of conrotatory circuit of triene system in diaryl - or heterylethene (Scheme 1) [46, 47] and fulgides (Scheme 2) and, secondly, initiated by light transformations of supramolecular aggregates formed by thermally stable topological structures (e.g. pseudorotaksane). Dithienylethene and indolylfulgide molecules are functionalized with a substituent (fluorophore group Fl) to read the information by fluorescence (see. Schemes 1,2).

High thermal stability ia typical for the bistable topological structures, which isomers interconversions do not include the energy-intensive acts cleavage formation of covalent chemical bonds. Electrically switchable catenane system [48] is shown in Scheme 3.

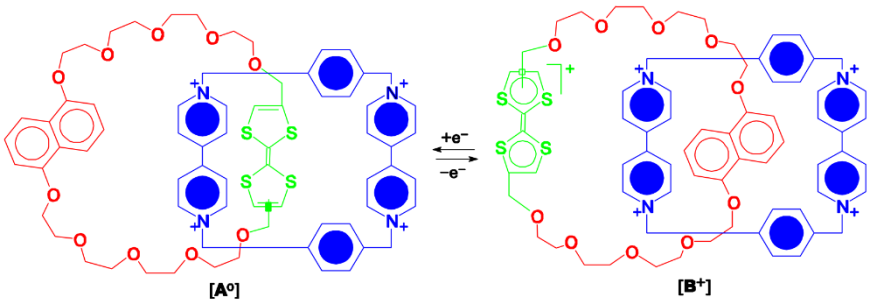
Scheme 1



Scheme 2



Scheme 3



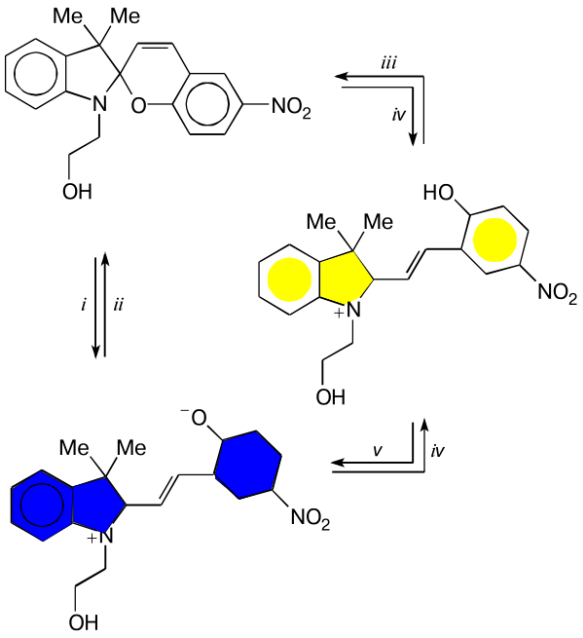
A monolayer of catenane molecules is placed between the metal and silicon electrodes. After electrochemical oxidation of a supramolecule, on the one of its parts an additional positive charge appeares. As in the original form of this part is close of the same charge, after the oxidation the pluces repel and molecules becomes reconfigured. There is a formation of a second stable form and the change of electrical resistance. The reaction is completely reversible. The state «on» occurs at the voltage of +2 V; state «off» occurs at the voltage of -2 V; readour is carried out by the resistance change at potential (~ 0.1V) superposition.

Photochromic regrouping of diarylethene (DAE) and spirocyclic (SPS) photochromic systems lead to drastic changes not only of spectral, but also of others properties of the isomers such as the oxidation-reduction. This feature can be used to create molecular optically controlled switches of electrical signal.

Molecular switches can be used to construct various logic elements (logic gates), which refers to the possibility of constructing a computing system based on the molecular elements [34]. All digital computing systems are described with

mathematical apparatus of the Boolean algebra (propositional algebra). The simplest logical elements model the Boolean algebra operations: AND, OR, NOT. More complex elements are described with logic functions for examples NAND, NOR, EXCLUSIVE OR (XOR), and others. Currently created molecular logic devices include multiplexers and demultiplexers [49-51]. Scheme 4 and Figures 6.6.7 and 6.6.8 show how to form a simple logic gates NOT and NOR on the basis of the photochromic spiropyran system [49, 52]. The input signal is UV radiation incoming to the photochromic cell with a spiropyran molecules, which leads to the opening of *2H*-pyran cycle and the formation of the colored form (Scheme 4). The output signal is the radiation from visible light source passing through the photochromic cell. When the UV source is switched off (Figure 6.6.8, *a*), the visible light passes through the cell with an unpainted (Figure 6.6.7, curve 1) spiropyran form hardly absorbed. This corresponds to the first row of the truth table (Figure 6.6.8, *a*.): input - "0" outputs - "1". When the UV source (input - "1") is switched on, the photoisomerization occurs and the dyed (Figure 6.6.7, curve 2). merocyanine form (there is ~ 3% of undyed form in the photostationary state) appears. Visible light hardly passes through the cell. This corresponds to the second row of the truth table (input - "1", output - "0"). Thus, the truth table corresponds to a logic operation "NOT".Figure 6.6.8, *b* shows the molecular logic gate with two UV radiation sources (two inputs). Its truth table corresponds to the logic gate "NOR".

Scheme 4



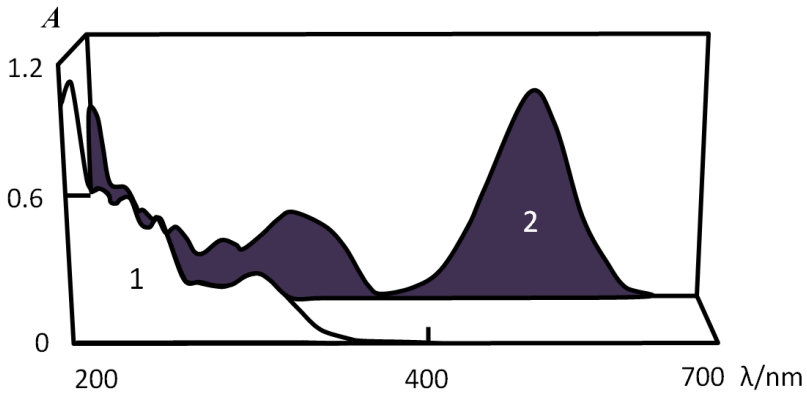
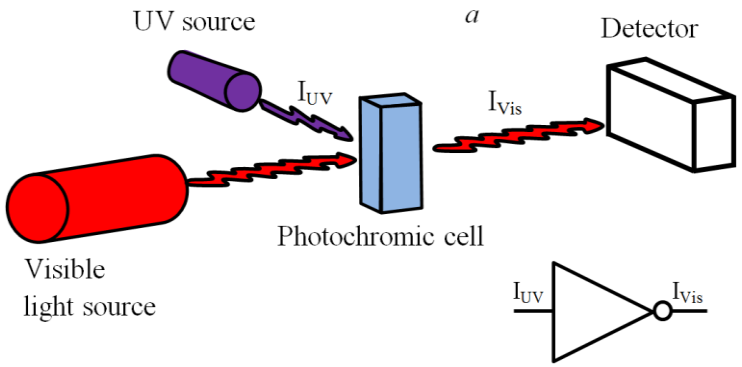
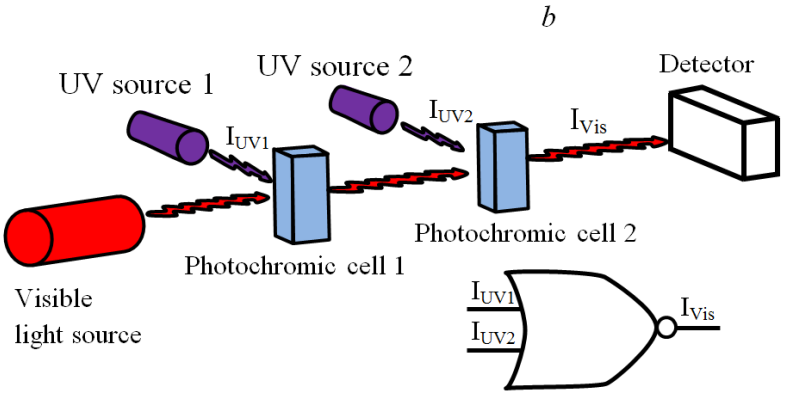


Fig. 6.6.7. The absorption spectra of the cyclic (1) and merocyanine (2) forms of spiropyran [34].



UV source	I_{Vis} (%)	Truth table	
		I_{UV}	I_{Vis}
Off	100	0	1
On	3	1	0



UV source		I_{Vis} (%)	Truth table		
1	2		I_{UV1}	I_{UV2}	I_{Vis}
Off	Off	100	0	0	1
Off	On	3	0	1	0
On	Off	4	1	0	0
On	On	0	1	1	0

Fig. 6.6.8. The photochromic systems are the analogs of logic gates with one (*a*) and two *(b)* inputs.

Optical molecular memory

Principles of constructing of molecular memory are the same as for the switches, and its basis is bistable molecular structures and their transformations. Different memory types require different characteristics of these transformations, and in order to provide a long storage of the recorded information, the systems with a long lifetime of isomer B are required (Figure 6.6.3).

The existing storage devices record the information on the active medium surface (two-dimensional memory). The principle of three-dimensional optical memory [53] is shown in Figure 6.6.9. Bistable photochromic compounds are used as the active medium, and the method of two-photon absorption is used for recording. Necessary for the implementation of the photochemical regrouping hν energy is delivered to a certain point the volume by two quanta, which total energy corresponds to the excitation energy of $h\nu = h\nu_1 + h\nu_2$. After absorbtion of two photons A molecule regroups to a colored form B. The reading of the recorded information occurs during the detection of fluorescence, which must be present at one of the isomers of the photochromic system.

In case of use of the laser with the radiation of 532 nm (see Figure 6.6.9), the density of recorded information is about 1013 bit•cm–3, and in case of use of UV lasers, this value can be increased by an order [34].

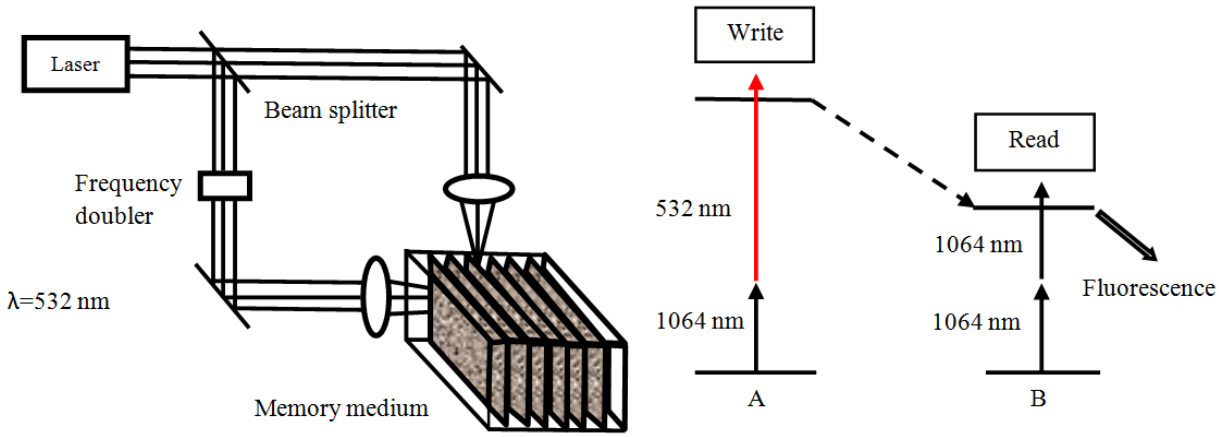


Fig. 6.6.9. The principle of three-dimensional optical memory with two-photon recording and fluorescence reading [34].

The diarylethenes and indole series fulgides are among the best photochromic systems for memory.

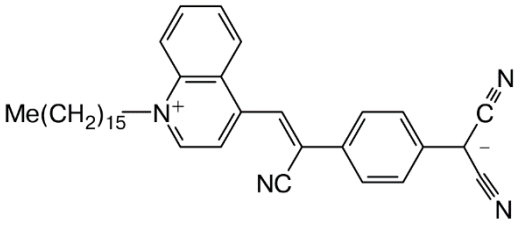
Molecular diodes and transistors

Optical molecular switches and storage devices have reached the maximum possible indexes at the molecular level - the speed characteristic of elementary reaction rate and the recording density of one bit - one molecule. However, the most convenient and technically implemented way of logical devices management, based on the molecular mechanisms, is the transmission of electrical signals [34]. This involves the effective molecular rectifiers and molecular conductors of electricity.

The properties of a molecular diode, i.e. rectifier, can be a characteristic of molecular structures **DσA** (**D** is an electron-donor group with sufficiently low ionization potential, **A** is an electron-acceptor fragment with high electron affinity, and σ is a spacer). Molecule **DσA** must have close enough, in terms of energy, polar zwitterionic form $\mathbf{D}^{+}\sigma\mathbf{A}^{-}$, in which the reverse intramolecular electron tunneling can be implemented. The introduction of spacer fragment is necessary for the separation of conjugated systems of donor and acceptor molecule fragments and obstacles substantial overlap of orbitals. At the same time, the spacer length should not exceed a certain limit, beyond which the probability of an electron tunneling is considered negligible. Figure 6.6.10 explains the mechanism of action of molecular rectifier [34].

Molecule or monomolecular layer of **DσA** molecules is placed between the metallic electrodes M1 and M2. When voltage is supplied on the M1 electrode, lowering the Fermi level to the level of the highest occupied MO (HOMO) localized on the donor fragment, the electron transfers from the contained among molecules electrodes to the M1 electrode. This process triggers the resonant electron transfer from M2 electrode on the lower unoccupied MO (LUMO) localized on the acceptor, and the formation of a metastable zwitterionic form. At the final stage, the tunnel intramolecular electron transfer is carried out. The total result is the electron transfer from M2 electrode to M1 electrode [34]. The first and perhaps the most strongly characterized example of the molecular diode at the present - rectifier of electric current - is a zwitter-ion, which structure is shown below

Scheme 5.



Two conditions are particularly important for the effective work of molecular rectifier: the localization of the HOMO in the donor molecule fragment and narrow to the maximum the energy gap among the boundary of MO.

The molecular structures with a narrow energy gap is of great interest not only for the creation of molecular diodes, but also for other branches of molecular electronics - organic semiconductors, transistors, light-emitting diodes, infrared electrochromic displays, etc. [54].

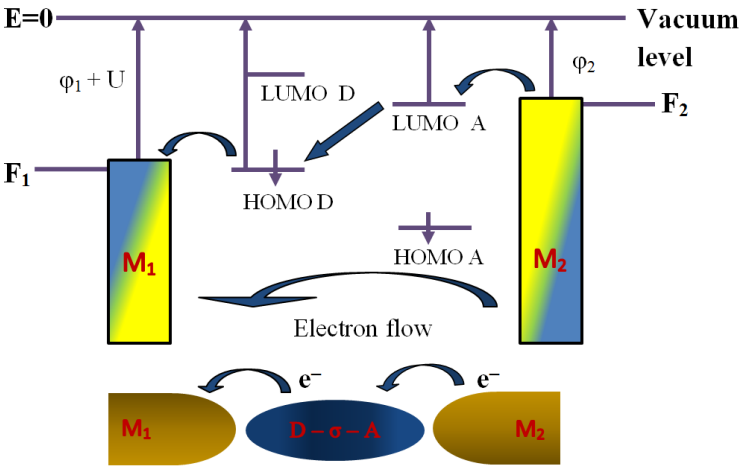


Figure 6.6.10. The action mechanism of molecular diode (F –are the Fermi levels of the metal electrodes M1 and M2, φ - electron work function, U - applied electric potential). [34]

The example of organic field-effect transistor based on DAE using polymethyl methacrylate PMMA layer [55] is shown in Figure 6.6.11. This structure provides the photoconversion of transistor due to photo-induced conductivity changes in the photochromic layers.

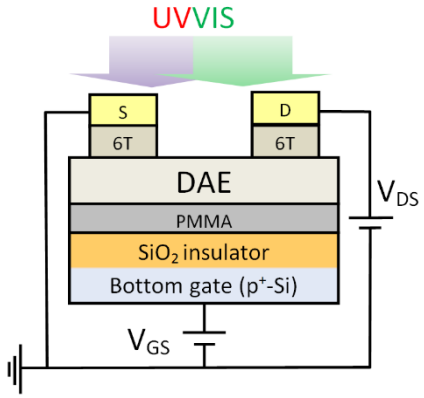


Fig. 6.6.11. The structure of organic field-effect transistor using diarylethene [55].

In previous examples, all functions of computer components were provided from the movement of electrons in complex molecular ensembles. Meanwhile,

these functions may be performed by photons as well. There are different variants of photonic devices, such as molecular photonic transistor. The molecular fragment of photonic transistor that absorbs a quantum of light (dipyrrhilborondiftoride) plays the role of drain electrode, the following molecule (zinc porphyrin) - conductor, and the last emitting porphyrin fragment of the molecule corresponds to a source electrode. Magnesium porphyrin acts as a control electrode - gate. If we oxidize the gate, after absorption of light the energy will be transferred not to the zinc porphyrin, but to non-radiating magnesium. The operation control and energy conversion in the devices on such transistors will be carried out using light quanta.

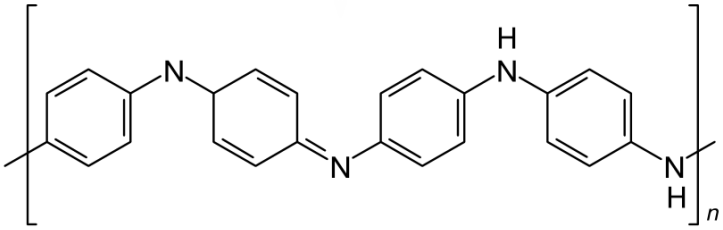
Molecular wires

The elements of molecular computer must be connected with conductors having cross-sections of size corresponding to the size of molecular switches and transistors.

Molecular wires may be created on a different basis. The first type - conducting polymers (doped polyacetylene, polythiophene, polyaniline, etc.). The second type of conductors is nanotubes. Nanotubes with monolayer or multilayer walls are obtained by passing an electric discharge between two graphite electrodes. Carbonic or boron-nitrogen nanotubes can be filled with metals and, in the result, we can obtain one-dimensional conductors consisting of chains of metal atoms.

The third type - conjugated oligomeric structures having a cross-section of about 0.3 nm and the length of 1 to 100 nm. They represent long conjugated molecules in which an electron is transferred through the chain π-bonds. If we attach metal-containing groups to the ends of such conjugated chain, the oxidation or reduction of one of them will provide sufficient conductivity across the chain. Combining doped (conductive) and undoped (with properties of semiconductors or insulators) sections of polymers, we can obtain electric circuits with required properties. Organic polymers may have electrical conductivity that is very close to this property of metals. Protonated emeraldine (oxidized form of polyaniline) is characterized by especially high conductivity, similar to that of metals, which is explained by the cationic radical type of resonance structure of this form (Scheme 6). [56] Emeraldine is used in rechargeable cells and electrochromic devices.

Scheme 6



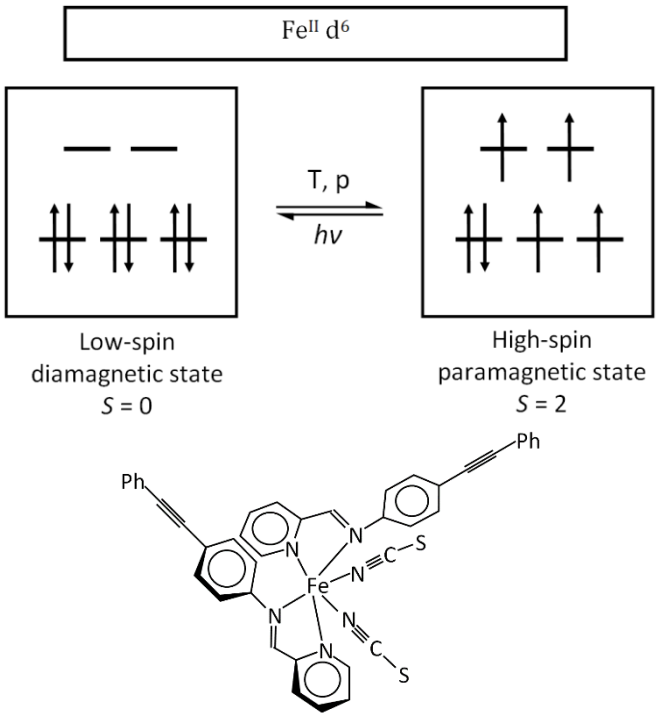
Molecular spintronics. Molecular magnetic memory. Molecular quantum computers.

Modern computers comprise a magnetic memory device in which information is recorded in the form of magnetization of the ferromagnetic disc. Information storage capacity is determined by the size of the magnetic domain in which the attached field (or lack of it) provides the parallel orientation of atomic spins. The recording density is increased by reducing the area of the domain; but when its size reaches 10-20 nm because of thermal motion the uncontrolled spontaneous disordering of spin system occurs due to so-called superparamagnetic effect. Therefore, there is a physical limit of the recording density of magnetic hard disk memory.

The transfer to molecular materials is necessary to overcome it. Thus, complexes Fe^{2+} [57] exhibit magnetic bistability - the ability to exist in the low-spin (diamagnetic) and high-spin (paramagnetic) states. The transitions between these states (crossover effect) have intramolecular nature and in the condition of the lack of interaction with the environment are carried out in steps. In case of strong intermolecular interactions, these transitions become cooperative in nature and can be characterized by a hysteresis loop due to the inclusion of intermolecular interactions in the process of changing the properties of the system in response to the external stimulus (temperature change or the magnitude of the applied field). [34] The curve of magnetization structure dependence $\chi(T)$, which illustrates the crossover effect and obtained by gradually increasing the temperature, and the curve obtained by cooling the sample vary. In the area of the hysteresis loop of the temperature dependence of magnetization structure $\chi(T)$ the system "remembers" the action (heating or cooling) in the result of which it occurred in this state, whether the transition from the high-spin to low-spin structure was carried out or vice versa. Thus, in the temperature region where the hysteresis loop is shown, the material can store information, i.e. it has the memory. As an illustration, the Scheme 7 shows the structure and spin transitions;

and Figure 6.6.12 shows the temperature hysteresis loop for a complex of divalent iron with Schiff bases and isothiocyanate ligands.

Scheme 7



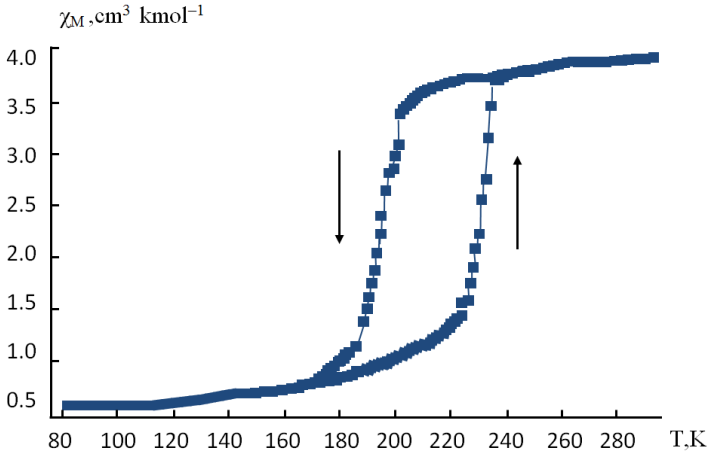


Fig. 6.6.12. The temperature dependence of magnetization for complex 7 [34].

Molecular magnetic materials possess a number of fundamentally new properties [34]: the high value of magnetic susceptibility and magnetization; high coercive fields (till 1-2 T) that is higher than for chromium oxide used in hard disk drives; low magnetic anisotropy; low specific weight and ease of machining; compatibility with polymeric materials for composites and solubility; low electrical conductivity that is typical for semiconductors and insulators; transparency, and possibility of control spin transitions using light; opportunities to modulate properties by structural modifications. The main potential advantage

of molecular magnetic materials is the principle possibility to achieve on their basis the maximal high data density, that is one bit per molecule. On the basis of molecular magnetic memory it is planned to develop the devices of nonvolatile magnetic random access memory MRAM, which unlike DRAM dynamic memory devices do not require ongoing support of renewable electrical pulses, thus avoiding the necessity in the constant heat dissipation.

Organic compounds containing unpaired electrons (organic radicals) in a crystalline state may exhibit paramagnetic and even ferromagnetic properties. Photoswitchable magnetic materials are no less interesting. Thus, in spiropyran cations with paramagnetic layered polymeric trioxalate anions [58, 59] we can observe the implemented modulation of magnetic properties of the material as a result of photoinduced structural changes of a bound photochromic fragment. The opportunities of molecular spintronics can be implemented in molecular quantum computer as well.

A quantum computer is a computing device that uses the phenomenon of quantum superposition and quantum entanglement for data transmission and processing [60]. The need for a quantum computer arises in cases when it is necessary to examine complex many-body systems, such as biological, using the methods of physics. The space of quantum states of such systems increases as the exponent from the number *n* of the components of their real particles, making it impossible to model their behavior on classical computers already for *n=10*.

The quantum computer uses for calculation not usual (classical) algorithms, but the processes of quantum nature, so-called quantum algorithms, using quantum mechanical effects - such as quantum parallelism and quantum entanglement. If the classical processor at each moment can be in exactly one of the states $|0\rangle,|1\rangle,...,|N-1\rangle$ (Dirac notation), the quantum processor at each moment is simultaneously in all of these basic states, upon that with its complex amplitude λ_j in each state $|j\rangle$. This quantum state is called "quantum superposition" of given classical states of and is referred to as

$$|\psi\rangle = \sum_{j=0}^{N-1} \lambda_j |j\rangle \qquad (6.6.8)$$

The basic conditions may also have a more complicated form. Quantum computing is the sequence of unitary operations of the simple form (over one, two or three qubits) controlled with the help of classic administrative computer. At the end of calculation, the state of quantum processor is measured, and that gives the desired result of the calculation.

The idea of quantum computation is that the quantum system of two-level quantum L elements (quantum bits, qubits) has 2L linearly independent states, and therefore, due to the principle of quantum superposition, the state space of the quantum register is 2^L -dimensional Hilbert space. The operation in quantum computation corresponds to a rotation of the vector of register condition in this space. Thus, quantum computing device of *L*-qubit size actually involves simultaneously 2^L classical states. That is, the computation takes place simultaneously in 2^L states, whereas in a conventional computer 2^L successive operations are necessary.

Physical systems that implement the qubits can be any objects having two quantum states: polarization states of photons, electron states of isolated atoms or ions, spin states of the nuclei of atoms, etc.

One classical bit can be in one and only one of the states $|0\rangle$ or $|1\rangle$. The quantum bit is in the state $|\psi\rangle = a|0\rangle + b|1\rangle$ such that $|a|^2$ and $|b|^2$ - probabilities to get 0 or 1 respectively in the measurement of this condition; $a, b \in C$; $|a|^2 + |b|^2 = 1$. Immediately after the measurement, the qubit comes into the basic quantum state corresponding to the classical result.

At present, the following systems are considered as prospective for quantum algorithms:

- the chain of ions in the trap
- superconductor structures (rings)
- nuclear spins in organic molecules
- semiconductors doped with spin atoms.

In cases when the role of qubits is played by nuclear spins connected with spin-spin interactions, NMR spectrometer can be used as a quantum computer. Then, by means of various pulse sequences we can ask any correlations among qubits.

In 2001, scientists developed the system consisting of 7 qubits on the basis of the molecule of (pentafluorobutadiene-2-yl) iron cyclopentadienyldycarbonyl [34] (5 fluorine atoms and 2 atoms 13C, Figure 6.6.13), and first implemented Shor's algorithm - number 15 was decomposed into prime factors of 5 and 3.

In future it will be possible to create quantum computers based on quantum dots.

In addition, DNA computing is promising. If an ordinary computer manipulates the combinations of values "0" and "1", there are four basic states (A, G, T, C) in the DNA, respectively, the number of combinations increases many times. Information density of DNA computers is 10^{27} bit/nm^3, while for a classic

modern digital computer it is 10^{12} bit/m^3. DNA computer is capable to calculate 10^{19} operations per second, while a classic modern computer provides no more than 10^{13} operations per second. [34]

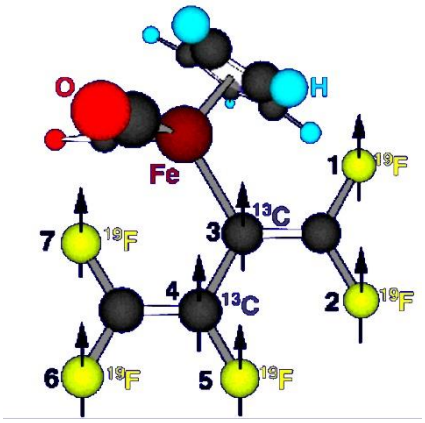


Fig. 6.6.13. 7-qubit molecular quantum computer

There is a significant success in the development of the three components of a molecular computer - molecular switches, memory and conductors. However, the task of synthesizing a computer from these components has not yet been solved. While this problem is not solved, it is planned to create hybrid devices that combine the advantages of molecular approaches with the most successful semiconductor technologies. Hybrid devices can be fabricated, for example using high affinity of some atoms in organic molecules to heavy metals (for example, sulfur to gold, and oxygen to silver). This creates contacts among metal electrodes and molecular conductors.

6.6.4 Nanophotochromism

In recent years, a new nanotechnology direction is forming - nanophotochromism [61] associated with the development of photochromic hybrid (composite) systems (Figure 6.6.14) comprising molecules of photochromic organic compounds, in particular from the DAE classes, and nanoparticles of noble metals, namely gold and silver, as well as inorganic semiconductors. In such systems, molecules of photochromic DAE may be physically (due to the Coulomb interaction) or covalently (due to chemical interaction, because of the presence of substituent) associated with nanoparticles of noble metals.

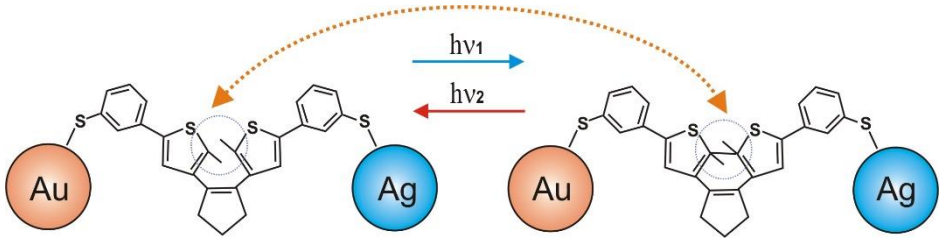


Fig. 6.6.14. The scheme of photochromic transformations in a nanocomposite organometallic system based on DAE.

The presence of chemical bonds of DAE molecules with a metal nanoparticle while maintaining the photochromic properties opens great prospects for the use of nanocomposites in molecular electronics and photonics devices. With the help of metal nanoparticles the electrical contact may be carried out in case of the application of these nanocomposites as molecular switches with the photo-induced change of electrical conductivity. In addition, by varying the characteristics of metal nanoparticles (material and dimensions) and, thus changing the resonant frequency of localized plasmons $\omega_0 \approx \sqrt{\frac{2\varepsilon_F}{mR^2}}$, we may control the parameters of photochromic processes in nanocomposites due to the plasmonic effects.

When molecules are adsorbed on the metal surface with nanoscale roughness (Figure 6.6.15) along with the enhancement of the Raman scattering, we can observe enhanced luminescence, as well as increased efficiency of photochemical processes [62].

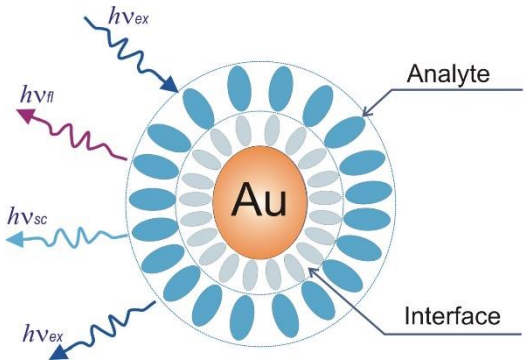


Fig. 6.6.15. The scheme of optical processes in a nanocomposite organometallic system

Photochromism of molecules in the solid phase composite metal-organic nanostructured systems consisting of Ag and Au nanoparticles with a shell, for example, from molecules of photochromic diarylethene or spirocyclic

compounds, manifests in photo-induced changes of electronic (Figure 6.6.16) [63] and vibrational (Figure 6.6.17) [64] structures of nanocomposites and their electrical conductivity (Figure 6.6.18) [65].

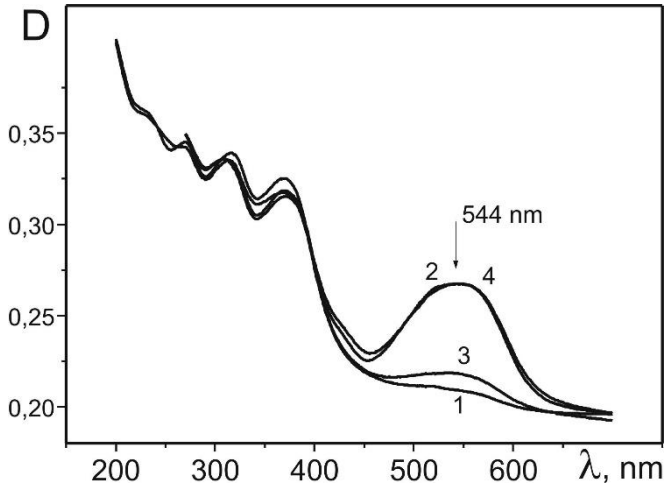


Fig. 6.6.16. Photo-induced changes of electronic structure manifested in changes of absorption spectra of solid-phase composite nanostructured systems "metal nanoparticle-photochromic molecule" (1,2) based on DAE and silver sols on quartz and solid-phase DAE films on quartz (without sols) (3,4). Before irradiation (1,3) and after UV irradiation (2,4) [63].

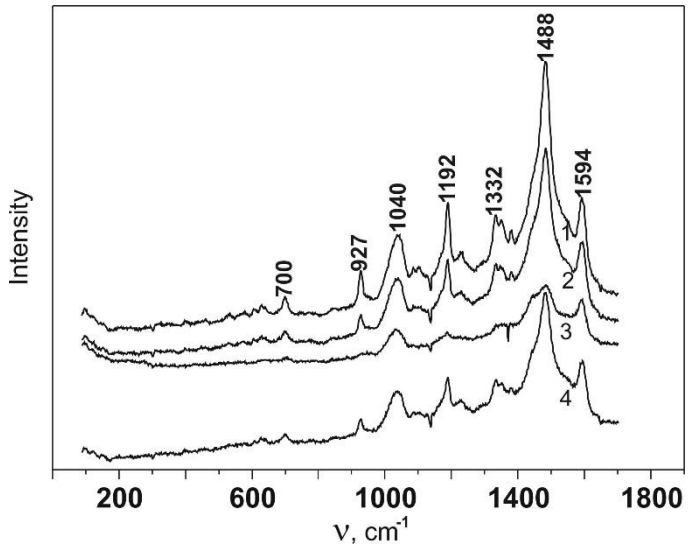


Fig. 6.6.17. Photo-induced changes of vibrational structure manifested in changes of SERS spectra of solid-phase composite nanostructures "metal nanoparticle-photochrome" (based on DAE and silver sol on a quartz substrate) λ_{exc} = 633 nm (1 - before irradiation, 2 - after UV irradiation, 3 - after subsequent laser radiation λ = 633 nm; 4 - after subsequent UV irradiation) [64].

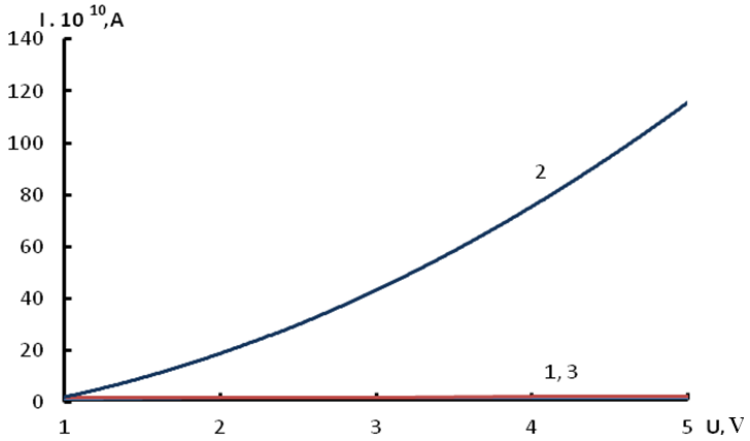


Fig. 6.6.18. Photo-induced changes of volt-ampere characteristics of the solid-phase film containing DAE, oxazine 17, and Ag nanoparticles (before (1), after UV irradiation (2) and subsequent visible light irradiation (3)) [65].

In the case when quantum dots are used as nanoparticles, we can observe the reversible fluorescence modulation (Fig. 6.6.19), consistent with the diarylethene photochromic transformations and determined by fluorescence resonance energy transfer [66].

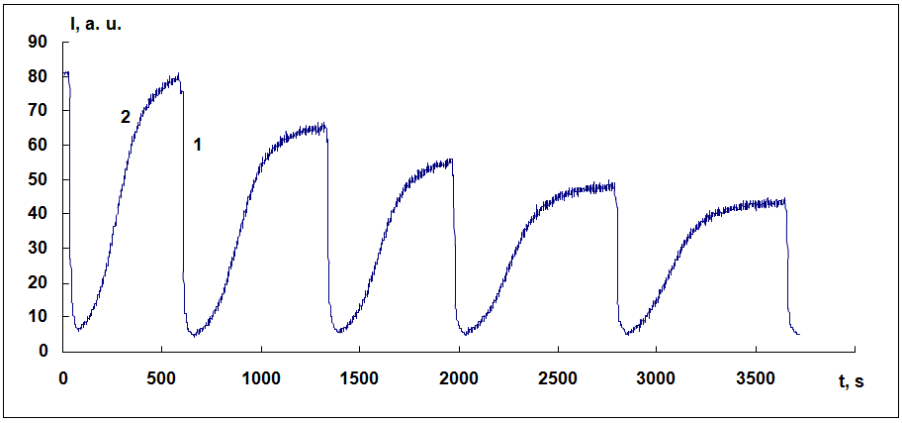


Fig. 6.6.19. Photo-induced cyclic changes of fluorescence intensity in the maximum of fluorescence band of quantum dots 580 nm (when fluorescence is excited by light with a wavelength of 500 nm) for a solid-phase film based on the photochromic DAE and quantum dots contained in the ratio of 2:1 (under sequential UV (1) and visible (2) irradiations) [66].

Such composite nanosystems can be used in the development of electro-optical photoswitchable elements (allowing to provide a reversible photo-induced modulation of the absorption and fluorescence with a simultaneous varying the photoconductivity) for ultra capacity memory devices and photocontrolled molecular switches.

The above information shows that in the near future it is possible to create the necessary components for the molecular computer. Table 6.6.2 presents the

main achievements in this area and the problems that need to be explored and resolved.

Table 6.6.2. Major achievements and problems of the modern stage of the development of the components for molecular computer [34].

Component	Achievements	Problems
Switches	Optical	
	Response time—up to 10^{-12} s «on» / «off»—contrast	Photodegradation Optical – to - electrical signal conversion
	Electrical	
	High cyclicity Compatibility with modern architecture	Slow response Limited number of contact electrodes
Memory	Optical	
	Three-dimensionality Non-volatility FMD/WORM-Terabyte disks Data density 1 bit/1 molecule	Nondestructive reading
	Redox	
	Data density 10^{11} bit·cm^{-2} Pitch 33 nm Memory cell size —0,001 μm^2	Diffraction limit Defects in the self-assembling monolayers
	Magnetic	
	MRAM, non-volatility Density 1 bit/1 molecule	Low operating temperature
Wires	Metallic	
	Thickness 6-10 atoms Pitch to 16nm	Atomic assembling
	Molecular	
	Conductivity—up to 10^4cm·m^{-1} Current density up to 10^{12}cm^{-1}·nm^{-2}	Short length Low electric current

6.7. Metamaterials

6.7.1. Introduction

Metamaterials are synthetic materials which have unique properties that conventional materials known to man don't possess [67]. This section describes electromagnetic properties of metamaterials, i.e. their properties in electromagnetic field in various wavelength bands, e.g. microwave, terahertz and optical bands.

The properties of metamaterials usually depend heavily on their structure, i.e. their shape, size and elements position. The material, these elements are made of, usually have less influence on its properties. The prefix μετα takes its origin from the Greek word "beyond", so it allows us to study materials as objects with properties that go beyond the properties of their elements. The position of elements is repeated through space, therefore metamaterials are considered to have discrete and periodic structure.

Metamaterials have two main differences to photonic bandgap crystals. First, metamaterials usually have rather dense periodic structure with a step, considerably smaller wavelengths of an electromagnetic field. At the same time in photonic crystals, which also have periodic structure, the components are positioned more rarely in space. Second, metamaterials generally consist of metallic elements, and photonic crystals contain dielectric (non-conducting) elements.

As the period of elements position in metamaterials is much larger than wavelength, the diffraction of electromagnetic waves doesn't occur in these elements. Therefore, waves are transmitted (propagated) in the material as in a uniform medium, which is characterized by particular average parameters. Hence, it is possible to provide metamaterials with effective values of dielectric and magnetic inductivity, which depend on the cell geometry and cell components.

6.7.2. Two and three-dimensional metamaterials

Bose's swirling bundle (fig. 6.7.1)can be assumed as the first metamaterial, investigated at the end of the 19^{th} century [68, 69].

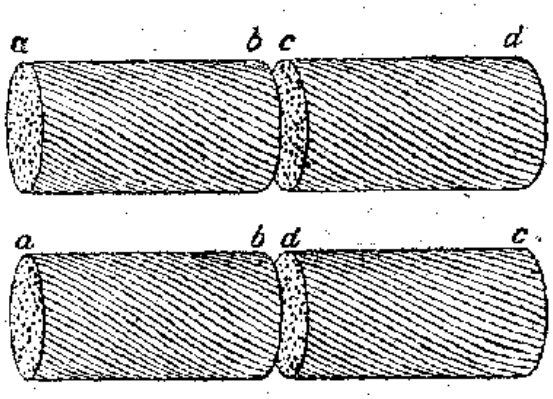


Fig. 6.7.1 –Artificial spiral “molecules”, made by Bose from jute [63]

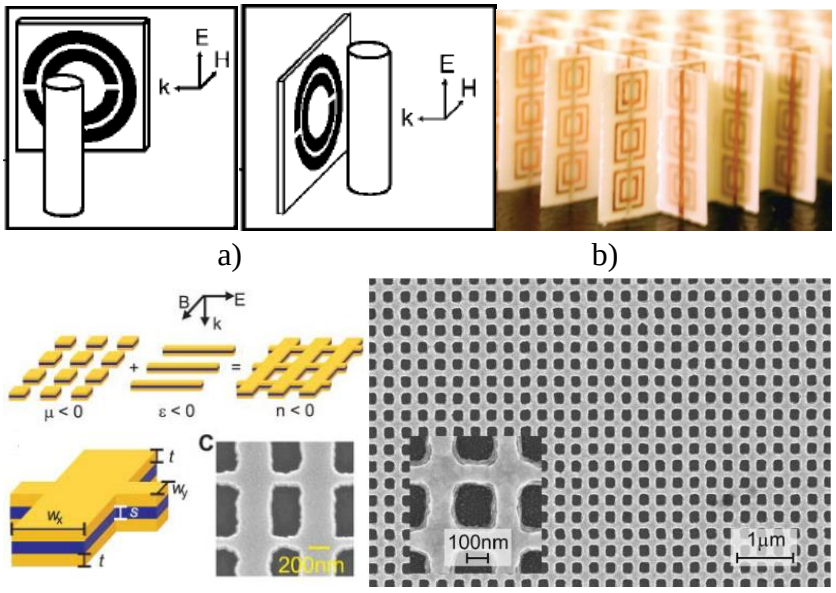
In 1914 Lindman studied the artificial environments representing a set of randomly focused small wires twisted in a helix and enclosed in the environment that was fixing them [70]. Among the first structures, satisfying the definitions of metamaterials given above, there were two and three-dimensional arrays of metal rod stock which behaved like dielectrics with the index of refraction less than unity [71, 72] and the negative dielectric permittivity below plasma frequency. Then frequency selective surfaces, the majority of which can be considered as flat metamaterials, were investigated.

In 2000 John Pendry et alias [73] suggested using split-ring resonators (discussed earlier in [74]) for creation of artificial environments with a possible negative magnetic response. Later Smith et alias [75] showed split-ring resonators and the waveguide conducting metamaterials with simultaneously negative dielectric permittivity (the waveguide conductors) and magnetic permeability (split-ring conductors) (fig. 6.7.2).

In a year’s time Smith’s team designed the metamaterial that was first used to demonstrate the negative index of refraction [76]; in 2006 another group of researchers showed the negative index of refraction in optical band received by means of layer structure [77]; in 2007 the same group of researchers designed structures with the negative index of refraction at a wavelength of 780 mil [78].

The materials with both negative dielectric and negative magnetic conductivity at the same time (and, therefore, negative refraction index) were predicted theoretically in 1967 by V. Veselago [79]. It should be noted [80] that this case was previously discussed by Sivukhin [81] and then by P. Pafomov in his articles mainly for Cherenkov’s effect [82-84], the connection between negative refraction and negative group velocity was also discussed in scientific papers [85-88]. K. McDonald in his articles [89] provided an overview on the background of negative group velocity. This information was also mentioned in

Lamb's [90] and Laue's [91] early studies. Later, V. Veselago [92] classified Schuster's [93] and Pocklington's [94] scientific papers as early studies in this area.



(a) –the first suggested metamaterial with symalteniously negative dielectric permittivity and magnetic permeability [9]; (b) –metamaterial that was first used to show negative refraction [76]; (c) –layer structure for obtaining negative refraction in optical band [77]; (d) –SEM photograph of a structure with negative refraction index at the wavelength of 780 mil [78].

Fig. 6.7.2–Metamaterials with negative refraction index

In 2000 John Pendry [95] concluded that the plate with a single negative index of refraction is a perfect objective, not limited by diffraction, and therefore it can focus on arbitrary small point. Such structure came to be called a superlens. The potentialities of superlens application for visual representation, data storage and lithographic processing have been one of the main driving motives of metamaterials research ever since. After negative refraction index [96] and superlens [97], which were demonstrated in a microwave band, scientists' attention was focused on getting negative refraction index for a visible wavelength band. However, high metal losses at optical frequencies make magnetic response of split-ring resonators worse [98], that is why structures of different kind are required. The design of a metamaterial with negative refraction index evolved from layer-like double conductive lines [99] to structures like "fishing net", suggested theoretically first in [100] and created by experiment first in [101], and carried out for the wavelength of 780 mil in [77, 78]. Until recently all material samples with negative refraction index have had considerable losses, which are rather high for most practical applications. The difficulties in the production of

such small-scale materials limit the attempts of achieving the effect of even less wavelength light.

6.7.3. Metamaterial as a medium with simultaneously negative values of dielectric permittivity and magnetic permeability

The scheme below indicates the classification and comparison of different materials depending on the signs of dielectric and magnetic permeability.

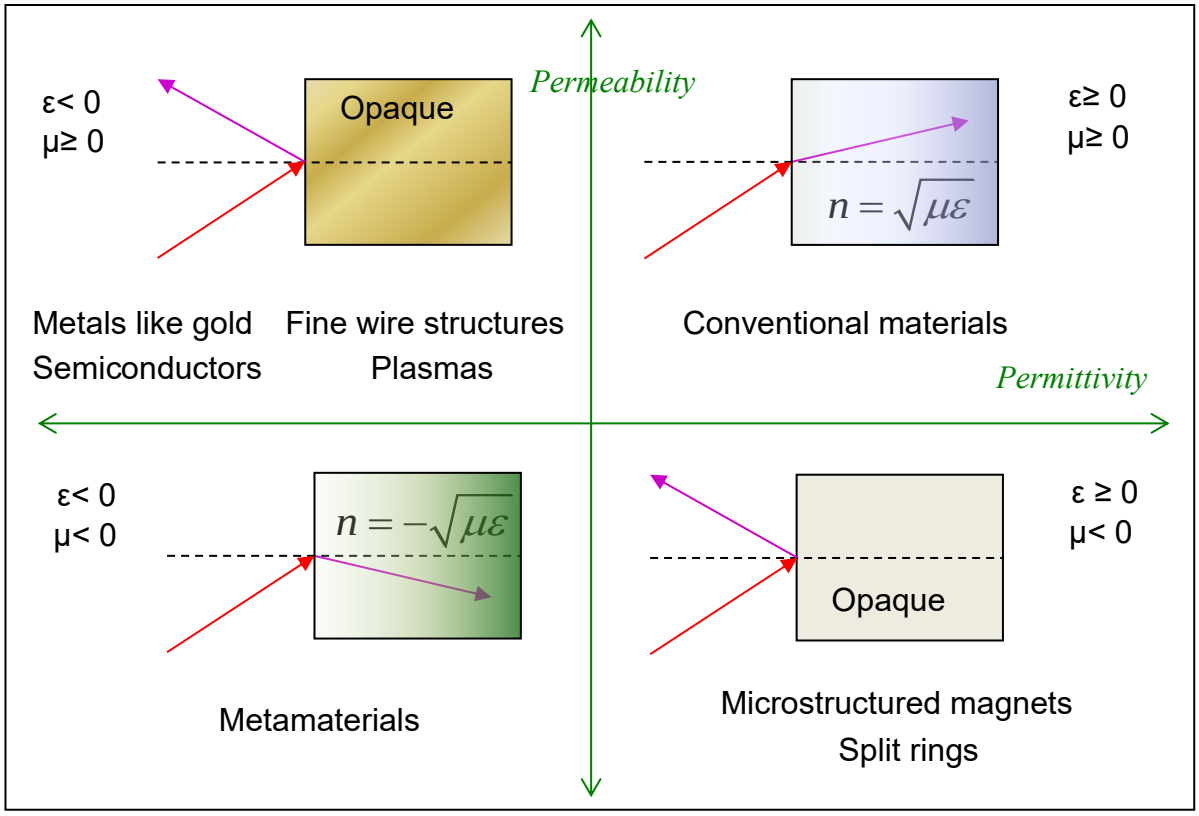


Fig. 6.7.3 –the comparison of different materials properties depending on the signs of dielectric permittivity and magnetic permeability

The use of chiral elements drew scientists' attention due to the fundamental problem of the "left-hand medium", which is a medium with simultaneously negative dielectric permittivity and magnetic permeability.

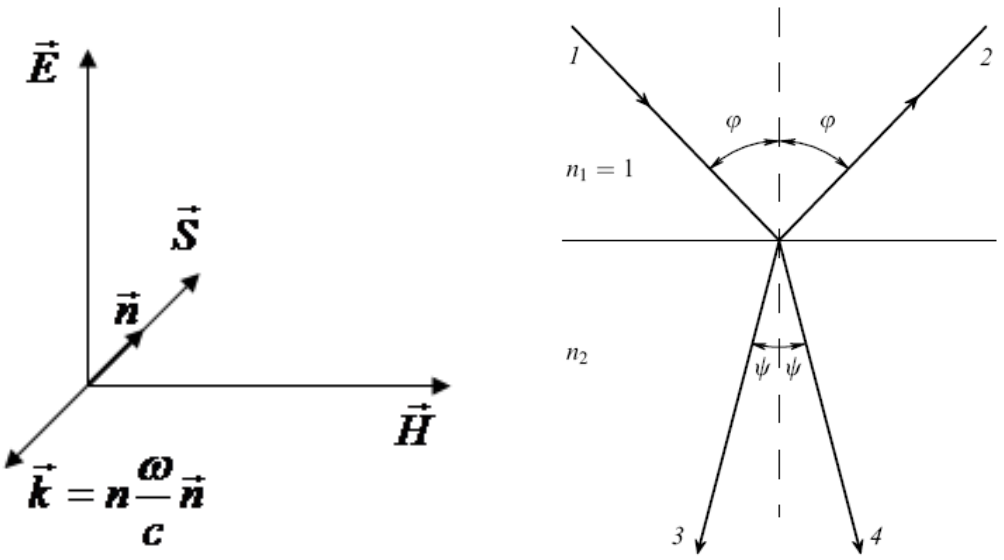
In 1967 V. Veselago [79] introduced the idea that it would be possible to create media with negative refraction index. That idea was summarized as follows. If constitutive equations are written in the form of $\vec{B}=\mu_0\mu_r\vec{H}$, $\vec{D}=\varepsilon_0\varepsilon_r\vec{E}$ (here ε_r is a relative dielectric permittivity, μ_r is a relative magnetic permeability), for monochromical electric and magnetic fields, time dependence of which is

described by means of the function $\sim e^{-i\omega t+ikz}$ and wave vector of a plane Hertzian (electromagnetic) wave $\vec{k}=\frac{\omega}{c}n\vec{n}$, where $\vec{n}$ is a unit vector along the axis z, from Maxwell solutions the expression for refractive index is known $n^2=\varepsilon_r\mu_r$.

$$rot\vec{E}=i\omega\vec{B},\quad i\left[\vec{k}\vec{E}\right]=i\omega\vec{B},\quad i\frac{\omega}{c}n\left[\vec{n}\vec{E}\right]=i\omega\mu_0\mu_r\vec{H},$$
$$\sqrt{\varepsilon_0\mu_0}\,n\left[\vec{n}\vec{E}\right]=\mu_0\mu_r\vec{H},\quad n\left[\vec{n}\vec{E}\right]=\sqrt{\frac{\mu_0}{\varepsilon_0}}\mu_r\vec{H} \qquad (6.7.1)$$

$$rot\vec{H}=-i\omega\vec{D},\quad i\left[\vec{k}\vec{H}\right]=-i\omega\vec{D},\quad i\frac{\omega}{c}n\left[\vec{n}\vec{H}\right]=-i\omega\varepsilon_0\varepsilon_r\vec{E},$$
$$\sqrt{\varepsilon_0\mu_0}\,n\left[\vec{n}\vec{H}\right]=-\varepsilon_0\varepsilon_r\vec{E},\quad n\left[\vec{n}\vec{H}\right]=-\sqrt{\frac{\varepsilon_0}{\mu_0}}\varepsilon_r\vec{E} \qquad (6.7.2)$$

If $\varepsilon_r<0$ and $\mu_r<0$ at the same time, you need to choose $n=-\sqrt{\varepsilon_r\mu_r}$ to make the equations (6.7.1) and (6.7.2) solved. Then vectors $\vec{E},\vec{H}$ and $\vec{k}$ form left-hand system (see fig. 6.7.4a).



a) b)

Rays 1–4 are usual refraction, rays 1–3 are negative refraction, rays 1–2 are reflection.

Fig. 6.7.4 – Mutual direction of electric and magnetic-field vectors and the wave vector of a plane Hertzian (electromagnetic) wave in a left-hand medium (a); light refraction and reflection at the interface if two media (b)

In this medium a plane wave is a backward wave phase and group velocities of which are directed oppositely. This results in an unusual form of Snel's law (fig. 6.7.4),

$$\frac{\sin\psi}{\sin\varphi} = \frac{n_1}{n_2} \tag{6.7.3}$$

when there is a wave refraction in the medium at an obtuse angle. Only now it is possible to find ways to solve a problem posed by Veselago due to the progress in the sphere of artificial composite media.

The model of such medium based on the cylinders with the conductivity along helical lines was offered in the research paper [102]. Such medium is known to be simulated by a set of straight-line conductors and split rings [103]. It is reported in the paper [104] that "left-hand medium" may consist of wave leading structures, which are beyond the limits and responsible for negative dielectric permittivity, and one-dimension chiral elements in the form of multifilar helix, responsible for negative magnetic permeability. The clearing of the below-cutoff waveguide was identified while putting chiral samples in it, and this owes to the "formation" of the medium with simultaneously negative dielectric permittivity and magnetic permeability. The paper [105] considers the task of simple circular cylinder (consisting of "left-hand medium") diffraction.

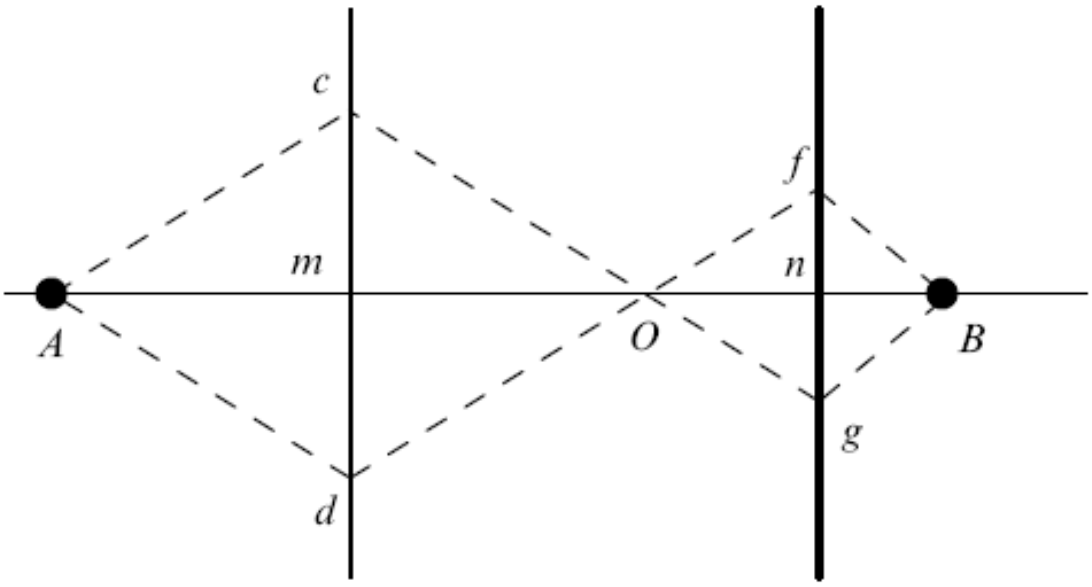


Fig. 6.7.5 - «Veselago Lens»

The paper [106] presents the numerical investigation of the optical properties of square nanohelices, made of gold, in the terahertz band depending on their geometrical parameters. The magnetic resonance was noted, the range and the position of which changed with the increase of the number of nanohelices turns. While increasing the number of nanohelices turns by more than three, the actual part of magnetic permeability possessed negative value on the frequencies which are more than 400 THz.

The paper [107] considers the focusing properties of bilayered plate (so-called "2Dlens") which consists of isotropic chiral electromagnetic media with negative values of dielectric permittivity and magnetic permeability.

This device is a *parallel-sided plate*, made of material with negative refraction n = -1. With the help of "Veselago lens" you can get pictures of the objects, which are at a distance less than lens thickness, but you can't get pictures of a source at a larger distance.

The paper [108] demonstrates experimentally and theoretically dielectric metamaterials of adjusting band with negative effective magnetic permeability in terahertz band (0,2–0,36 THz). These structures consist of an array of nonmagnetic rods made of ferroelectric SrTiO. Magnetic response and its adapting can be obtained by means of temperature effect on dielectric permittivity. They are defined by resonance state of electromagnetic field inside the rods.

The paper [109] presents frequency-response characteristics of the devices based on the chiral metamaterials considering terminal conditions. Chiral metamaterials were made on the basis of micro-scaled Y-structures based on Al using the lithographic technique on dielectric substrates of Mylar type. The results of computer-based simulation and experimental data are also presented.

The paper [110] studies reflective properties of electromagnetic waves of optical band in isotropical absorbing and chiral nonabsorbent media. It also says that unusual negative reflection occurs at boundary surface of chiral medium and perfectly conducting plane at a higher chilarity parameter. This results in focalizing property of such conducting plane in strong chiral medium.

The paper [111] studies the achievability of negative refraction in a chiral composite, consisting of chiral and dipole particles mixture. The negative refraction can be observed in the neighborhood of tuned frequency of chiral particles. Resonance chiral composites can be applied for realization of negative refraction and creation of a superlens in optical band.

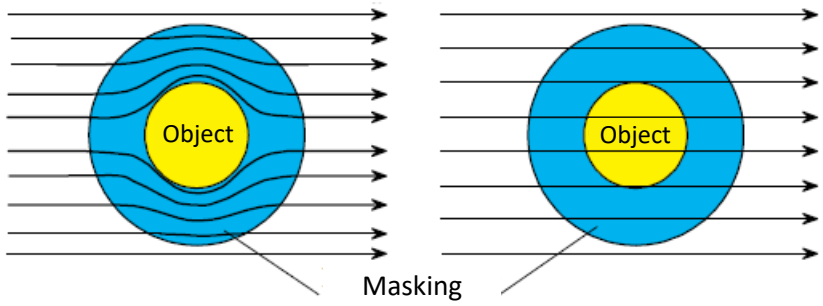
The paper [112] presents a new computational method of effective constitutive parameters of 2D arrays. The object of research was an array made from resonance chiral cylindrical diffusers in the form of quill cylinders with surface conductance along helical lines which have opposite chiral signs. Such structure was found to have “left-hand medium” properties (Veselago medium) in a narrow band.

6.7.4. The application of metamaterials for objects camouflage using the wave flotation method

After the first paper about the metamaterials with negative refraction index it became clear that the potential of metamaterials is much larger. As a result, the possibility to control electrical and magnetic properties of the material gave the rise to a new research area of transformation [113]. Its most remarkable application is the design of camouflage coating.

Nowadays among the variety of camouflage concepts there are two most developed principles [114]:

- the camouflage based on the wave flotation phenomenon (fig. 6.7.6a);
- the camouflage based on the dispersion compensation (fig. 6.7.6 b).



a) b)

Fig. 6.7.6 – The camouflage based on the wave flotation phenomenon (a); based on the dispersion compensation (b) [48]

The invisibility, achieved by means of camouflage based on the wave flotation phenomenon, is realized due to the fact that the object is placed inside the shell with the help of which electromagnetic waves envelop the object. After that they renovate their wave front and intensity distribution. Such shell works regardless of the properties of a hidden object, because electromagnetic waves don’t reach the object in this case. A specific of the camouflage based on the

dispersion compensation is that the dispersion from the object and the dispersion from the special coating compensate each other. As a result, the radiation is transmitted through the object without noticing it. In this case the properties of the developing coating must consider the properties of the hidden object. When the object absorbs the radiation it is necessary to use active materials or inner sources in the coating for complete dispersion compensation.

The basis of the transformation optics is the invariance of Maxwell's equations concerning coordinate transformations on the condition that electromagnetic material parameters of the media (tensors of dielectic ε and magnetic μ permittivity, as a rule) also transform as required. The condition of object's invisibility in the uniformity medium (without coating) is its vanishingly small size, as any object of finite size is sure to dissipate the field falling on it. While making the coordinate transformation, turning the point-like object into the object of finite sizes, it is possible to keep the invisibility condition by creating the necessary distribution of ε and μ around it [114].

To realize the necessary distribution of ε and μ of coating thickness, calculating with the help of mathematical apparatus of the transformation optics it is impossible to apply the natural materials because they don't let us get the required value set of material parameters. We need the metamaterials the macroscopic properties of which can be obtained by means of the synthesis of the arrays of building blocks (particles). The building blocks of the metamaterial, as a rule, are resonant metallic elements.

Depending on the chosen working band of the device it can be split-ring resonators, canonical helices, the pairs of plasma particles, plasma nanowires, etc. In the radio frequency band metaatoms, which are the unit elements of the metamaterial, can be formed by a non-conductor with large dielectric permeability [114].

First experimental camouflage devices based on the transformation optics were realized for the microwave frequency band [115]. Then some attempts were made to realize the devices in infrared and visible bands. The priority in the cloaking idea and simulation by means of wave flotation method belongs to John Pendry et alias [116].

To create the camouflage coatings on the basis of the transformation optics we need extremely high-anisotropy and rather non-uniform metamaterials with low optical losses, which is very difficult to realize. That is why the designers have to make the compromise between the ideal distribution of ε and $\mu$and the possibility of its realization. This compromise can be reached due to the moving

from the condition of complete invisibility to the condition of partial invisibility, e. g. within the framework of 2D optical camouflage, refusal of some other requirements concerning invisibility, frequency bandwidth narrowing in particular, the reduction in size of the objects to be camouflaged, etc. [114].

6.7.5. The optimum shape of the helix as a metamaterial element: the equality of dielectric, magnetic and chiral susceptibility

One of the main helix characteristics is its specific torsion *q*, related to the helix pitch distance *h* by the formula:

$$h = \frac{2\pi}{|q|} \tag{6.7.4}$$

The value sign q indicates the direction of the helix torsion in space. If q > 0, the helix forms a right-hand screw. It is possible to get the following formula for the x-component of the electric dipole helix momentum (x-axis is directed along the helix axis):

$$p_x = \frac{i}{\omega}\int_{x_1}^{x_2} I(x)dx\,. \tag{6.7.5}$$

Here I is current rate, ω is current cyclic frequency, i is unit imaginary number, x_1 and x_2 are the helix ends coordinates.

This is a quite well-known formula [117 - 119].

Considering the helix geometric parameters let's determine the *x*-component of a magnetic momentum:

$$m_x = \frac{1}{2}r^2 q\int_{x_1}^{x_2} I(x)dx \tag{6.7.6}$$

where r is a helix radius. The formulae (6.7.5) and (6.7.6) are followed by the formula that shows momenta projection onto an axis [120]

$$p_x = \frac{2i}{\omega r^2 q}m_x \tag{6.7.7}$$

This is a universal formula because it doesn't depend on the current distribution in the helix. It is *x*-components of the helix momenta that are very

important as they play a key role in the electromagnetic wave radiation in the direction perpendicular to the helix axis.

It is possible to obtain the condition of circularly polarized wave radiation in the direction orthogonal to the helix axis:

$$|p_x| = \frac{1}{c}|m_x|. \tag{6.7.8}$$

where c is the speed of light in vacuum. This condition shows that the helix has optimal parameters. It might also be called balanced as it has equally valued momenta: electric, dipole and magnetic.

The condition of the main frequency resonance for the electric current in the conductor L length is the formula

$$\frac{\lambda}{2} = L, \tag{6.7.9}$$

λ is the wave length of the falling electromagnetic radiation. While satisfying this condition the helices become extremely active and the metamaterial may have simultaneously negative permittivity: dielectric and magnetic.

To characterize the helix we may introduce the rise angle, i.e. the angle between the line, tangent to the helix in any point, and the plane perpendicular to the helix axis.

The optimal rise angle of the helix, which is calculated under the conditions of the main frequency resonance, can be written as [120, 121]

$$\alpha = \arcsin\left(-2N_в + \sqrt{4N_в^2 + 1}\right) \tag{6.7.10}$$

where $N_в$ is the number of helix turns. If the helix has such an optimal rise angle the formula (6.7.8) is realized, i.e. it is balanced out.

Every helix possesses dielectric, magnetic and chiral properties at the same time. Hence, its behavior in the electromagnetic field can be described by means of coupling equations (equations of constraints) [122]

$$\vec{p} = \varepsilon_0 \alpha_{ee} \vec{E} + i\sqrt{\varepsilon_0 \mu_0}\, \alpha_{em} \vec{H} \tag{6.7.11}$$

$$\vec{m} = \alpha_{mm} \vec{H} - i\sqrt{\frac{\varepsilon_0}{\mu_0}}\, \alpha_{me} \vec{E} \tag{6.7.12}$$

here α_{ee} and α_{mm} are tensors of dielectric and magnetic susceptibility, α_{em} and α_{me} are pseudotensors characterizing chiral properties of the helix, ε_0 and μ_0 are

permittivity and permeability of vacuum respectively. The term «pseudotensor» means that its components act in a different way at the axis coordinates inversion in comparison with tensor components. From the principle of reflection (symmetry) of kinetic coefficients it follows that the next formula is satisfied:

$$\alpha_{em} = \alpha_{me}^{T} \tag{6.7.13}$$

where symbol T is the tensor conjugating. The pseudotensor α_{em} has only actual components for the nonabsorbent helix. Besides, the pseudotensor components α_{em} have the dimension $м^{-3}$, and the chirality parameter, characterizing the metamaterial as a whole, is non-dimensional.

The use of the formula (6.7.8) results in [123]

$$\alpha_{ee}^{(11)} = \alpha_{mm}^{(11)} \tag{6.7.14}$$

$$\alpha_{ee}^{(11)} = \pm\alpha_{em}^{(11)} \tag{6.7.15}$$

where $\alpha^{(ik)}$ is the components of the tensors and pseudotensors discussed; the «+» sign corresponds to the helix with right-hand winding, the «-» sign corresponds to the left-hand helix.

The formulae (6.7.14) and (6.7.15) show that helices with obtained optimal parameters display dielectric as well as magnetic and chiral properties. Inotherwords, theheliceswithoptimalparametersarecharacterizedby three equal susceptibilities for the fields directedalong the helix axis: dielectric, magnetic and chiral.

The equality of all three susceptibilities for optimal helices is confirmed by the experimental data, namely by the optimal helix radiation circularly to the polarized wave in the direction perpendicular to the helix axis.

As we show below, the optimal helices can be widely used, for example, to create reflectionless coatings and metamaterials with negative refraction of electromagnetic waves. The helices under study show optimal properties while being activated by both electric and magnetic fields, i.e. at any orientation of the polarization plane of the falling wave. And that is the advantage of the optimal helices over other possible metamaterial elements, e.g. linear oscillators and ring-shaped resonators.

The chiral medium properties can be described by means of the following equations [122, 124]

$$\vec{D} = \varepsilon_0\varepsilon_r\vec{E} + i\sqrt{\varepsilon_0\mu_0}\kappa\vec{H} \tag{6.7.16}$$

$$\vec{B} = \mu_0\mu_r\vec{H} - i\sqrt{\varepsilon_0\mu_0}\kappa\vec{E} \tag{6.7.17}$$

where ε_r is a relative dielectric permittivity, μ_r is a relative magnetic permeability, κ is a parameter of the chirality structure.

These formulae are also valid for the metamaterial formed by helices. For the isotropic medium with low inclusions concentration we can neglect the interaction between structure elements and define effective parameters as

$$\varepsilon_r = 1 + N_h \alpha_{ee} \tag{6.7.18}$$

$$\mu_r = 1 + N_h \alpha_{mm} \tag{6.7.19}$$

$$\kappa = N_h \alpha_{em} \tag{6.7.20}$$

where N_h is inclusions concentration. For the "optimal" form helices satisfying the formula

$$\frac{\omega}{c}|q|r^2 = 2 \tag{6.7.21}$$

we have $\varepsilon_r = \mu_r = 1 \pm \kappa$ (the upper sign corresponds to the right helix). Thus, there is an "optimal" relation between the helix radius and helix pitch distance, (6.7.21), at which all three susceptibilities are equal on a particular frequency: dielectric, magnetic and chiral (magnetoelectric).

On some frequency near the main helices resonance actual parts of the permittivity and permeability become zero. That's why the following formulae are satisfied [125, 126]

$$\mathrm{Re}\{\varepsilon_r\} = \mathrm{Re}\{\mu_r\} = 0, \quad \mathrm{Re}\{\kappa\} = \mp 1 \tag{6.7.22}$$

Refraction indices of the two circularly polarized natural modesfor $q > 0$ have the form

$$n_+ = 1 + i(\sqrt{\varepsilon_r'' \mu_r''} - \kappa'') \tag{6.7.23}$$

$$n_- = -1 + i(\sqrt{\varepsilon_r'' \mu_r''} + \kappa'') \tag{6.7.24}$$

where $\kappa'' > 0$. Here, imaginary parts of complex values are marked with two primes. We see that one natural mode has a single refraction index and very low losses because for the optimal helix $\varepsilon_r'' \approx \mu_r'' \approx \kappa''$. The same is true for left helices.

6.7.6 Metamaterials for a microwave band on the basis of chiral elements

The photos of several experimental metamaterial samples for a microwave band, manufactured at F. Skorina Gomel State University on the basis of optimal helices, are shown in Figure 7. The sample a) is used as a converter of a linearly polarized wave during its off-normal incidence into reflected circularly polarized wave [120, 127, 128]. The sample b) shows weak reflection properties during normal wave incidence [129, 130]. The samples c) and d) are frequency-selective microwave-absorbing material close to the ideal and without reflective wave in a wide band [131]. Such absorbing material is double-sided as it doesn't have metallic base and can be considered as a metasurface.

Doubled mutually identical with right-hand and left-hand torsion helices are used in the samples b), c) and d). Such helices are contained in metamaterials in equal concentrations. Hence, the metamaterial doesn't possess chiral properties. They are compensated, so in the formulae (6.7.16), (6.7.17) it should be written as $\kappa = 0$.

In this case, while using the conditions of vectors $\vec{E}$ and $\vec{H}$ continuity at metamaterial and air edges it is possible to define the formula for the reflected wave amplitude:

$$E_0^r = \frac{\left(\sqrt{\frac{\varepsilon_r}{\mu_r}} - \sqrt{\frac{\mu_r}{\varepsilon_r}}\right)\left(e^{-ikL} - e^{ikL}\right)E_0^i}{\left(1-\sqrt{\frac{\mu_r}{\varepsilon_r}}\right)\left(1-\sqrt{\frac{\varepsilon_r}{\mu_r}}\right)e^{-ikL} + \left(1+\sqrt{\frac{\mu_r}{\varepsilon_r}}\right)\left(1+\sqrt{\frac{\varepsilon_r}{\mu_r}}\right)e^{ikL}}, \qquad (6.7.25)$$

where E_0^i is incident wave amplitude. In the formula (6.7.25) the equation $k = \frac{\omega}{c}\sqrt{\varepsilon_r \mu_r}$ for wave number is used, which is a generally complex quantity.

If for some critical frequency the following equation is satisfied

$$\varepsilon_r = \mu_r, \qquad (6.7.26)$$

according to the formula (4.48) the reflection coefficient becomes zero.

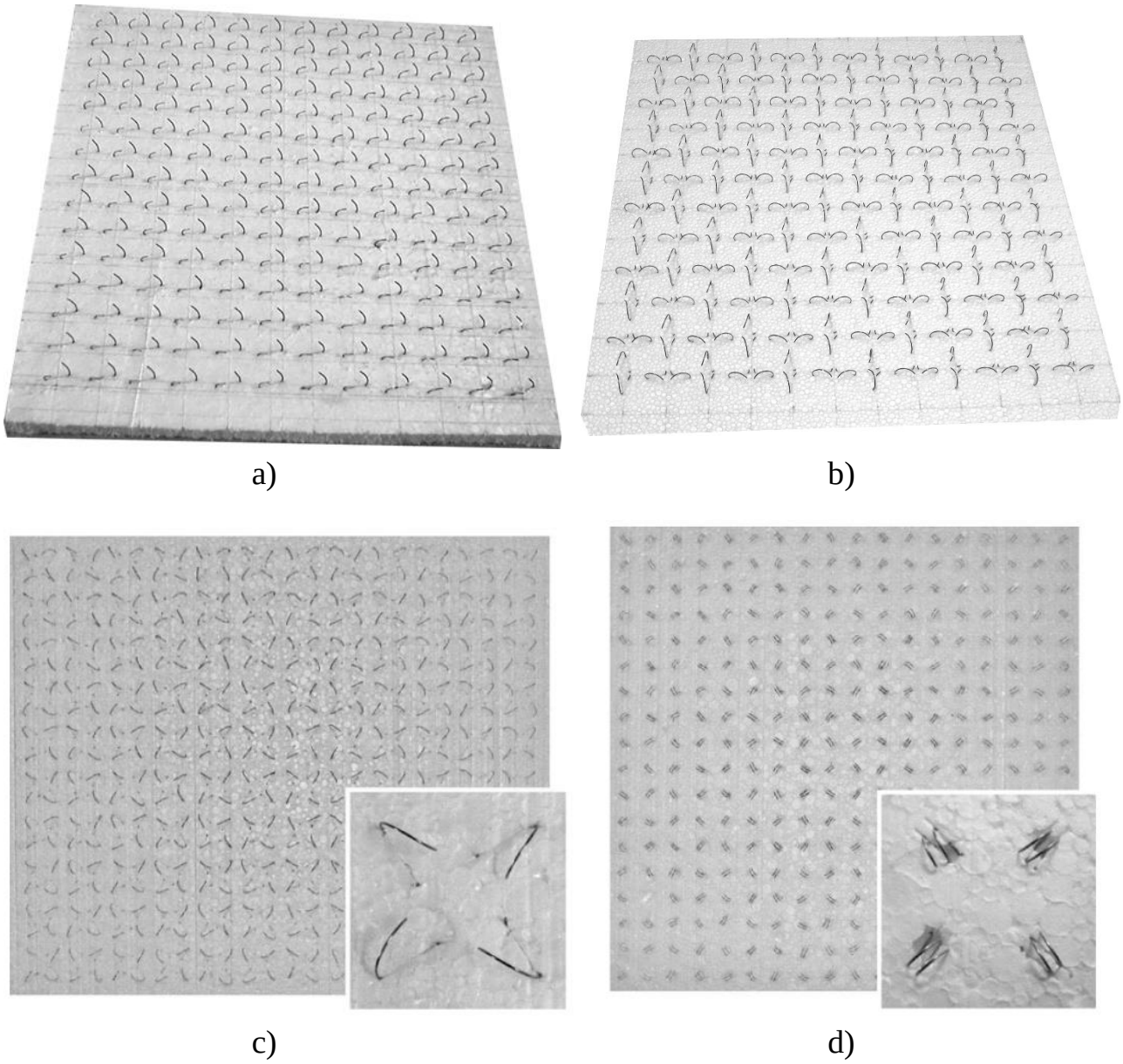


Fig. 6.7.7 – Photos of produced experimental metamaterial samples consisting of single-turn (a)-(c) and double-turn (d) optimal helices

Figure 6.7.8 displays the diagrams of the frequency dependence of the optimal helix susceptibilities. Each of the three susceptibilities (dielectric, magnetic and magnetoelectric) has actual and imaginary parts. The diagrams show that for these susceptibilities the formulae (6.7.14), (6.7.15) are satisfied, i.e. the helix is balanced out.

Figure 6.7.9 displays the frequency dependence of reflection, transmission and absorption coefficients of the double-sided absorbing material on the basis of optimal helices. The reflection coefficient is close to zero in a wide band. But at the same time the absorption is frequency-selective and increases greatly near the resonance frequency.

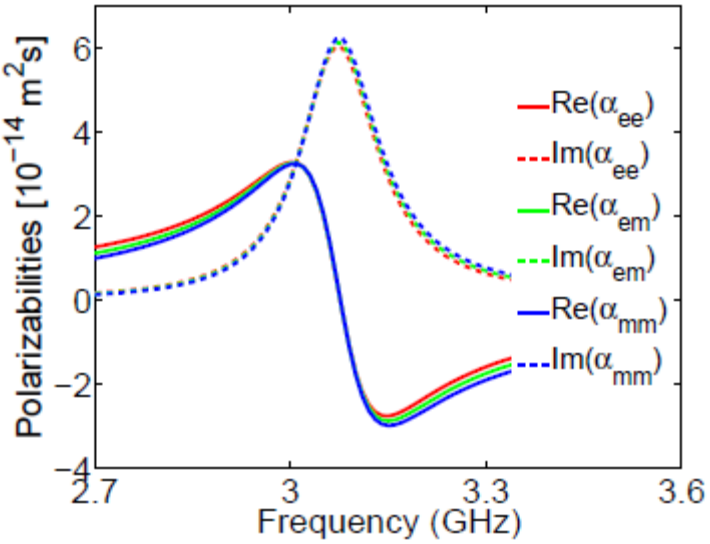


Fig. 6.7.8 –The susceptibilities of optimal helix dependence of the frequency [65]

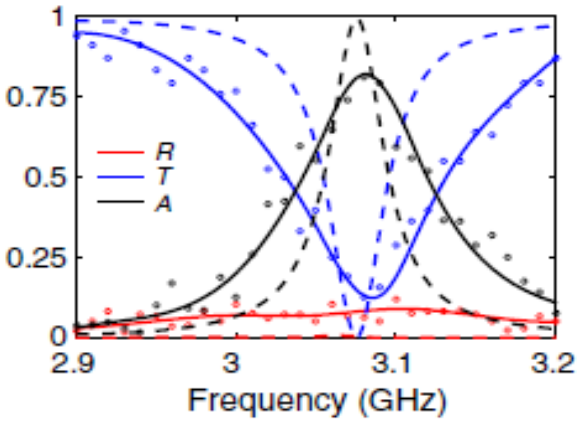


Fig. 6.7.9 -The frequency dependence of reflection, transmission and absorption coefficients of the double-sided absorbing material on the basis of optimal helices [65]

6.7.7 Chiral metamaterials for the terahertz band on the basis of helix elements

The precise 3D nano-design (Prinz-technology) method, developed at the Institute of Semiconductors Physics, Siberian Branch of the Russian Academy of Sciences, is used to make the samples of helix-structured metamaterials with the parameters optimal for THz band [132–134]. This method of three-dimensional micro and nanostructures is based on strained semiconductor film-stripping from substrate and its turning to 3D objects. The technology was called after V. Ya. Prinz who offered this method in 1995 while working at the Institute of Semiconductor Physics SB RAS. It is in practice in research laboratories of all

developed countries (the USA, Japan, Germany, etc.), but the application of this technology for developing helix-structured arrays of electromagnetic resonators and metamaterials takes place only at ISPhSB RAS and F. Skorina GSU.

The samples, having the form of square lattice made of helices which are fixed on the substrate by means of resist matrix, were produced at the Institute of Semiconductor Physics SB RAS. The helices have central contact with the substrate and the resist, the rest of the helix is in the air (fig. 6.7.10).

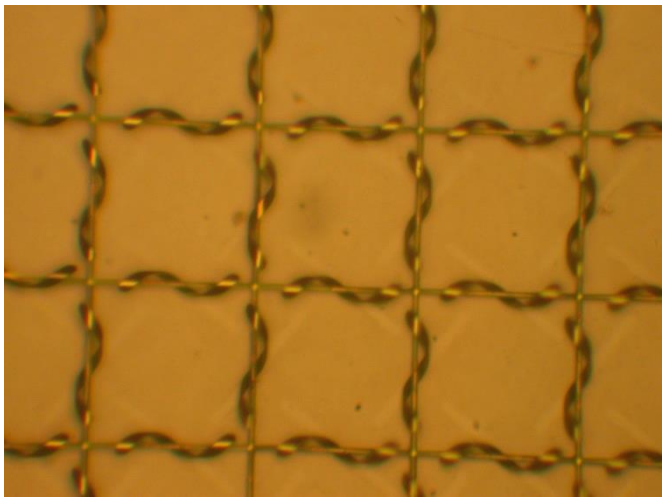


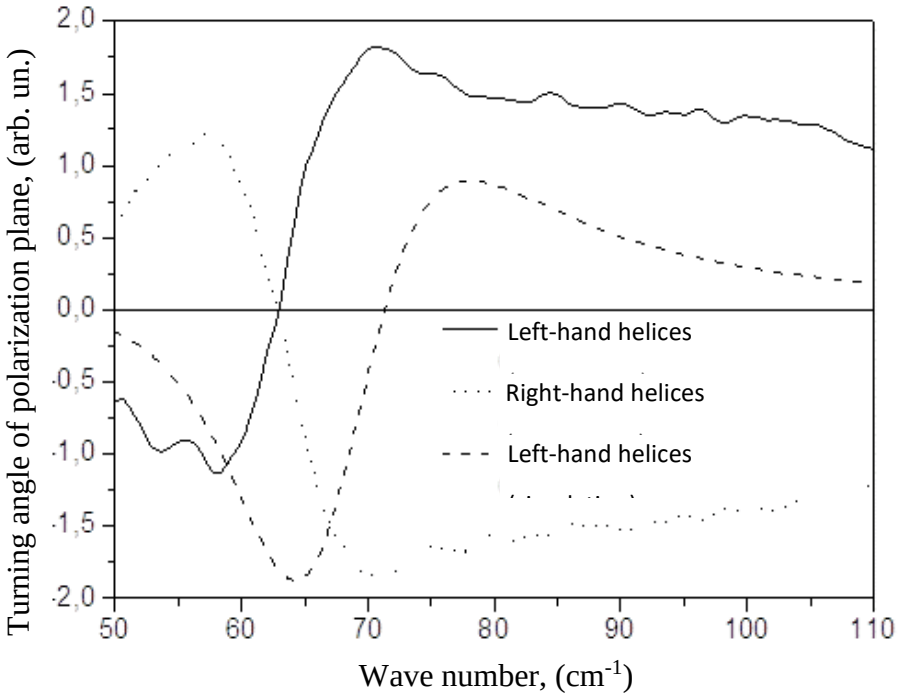
Fig. 6.7.10 –The photo of a helix array (the square net in the photo is a negative photoresist made of polymer material 1 μm thick)

The parameters of the strip when unrolled are the following: the length is 77 μm, the width is 6 μm. The strips are made of a film $In_{0.2}Ga_{0.8}As/GaAs/Ti/Au$ (16/40/3/65nm), in the middle the helix faces the substrate with the side of $In_{0.2}Ga_{0.8}As$. The rise angle of the helix is 52-53 degrees, the diameteris 11μm. The period of structure is 84 μm.The rise angle of the helix, which is equal to 52-53º, is optimal for obtaining the samples with maximal gyrotropic properties, as shown in [127, 128].

The samples are of different sizes maximum of which is 2 cm to 3 cm. The substrate GaAs is undoped; the thickness of the substrate is 400 μm.

The experimental research of the samples' properties was carried out at the Institute of Semiconductor Physics SB RAS. The results are given on fig. 6.7.11, 6.7.12.

These figures also show the results of computer-based simulation of the artificial anisotropic structure properties [135]. The parameters of this structure were chosen according to the experimental samples.



Simulation results for the array of left-hand helices (a broken line). 5˚ corresponds to1 by Y-axis. The observer is watching toward the wave, positive angle reading is clockwise.

Fig. 6.7.11 – The turning angle of polarization plane of the transmitted radiation for the array of left-hand (solid line) and right-hand (broken line) helices [135]

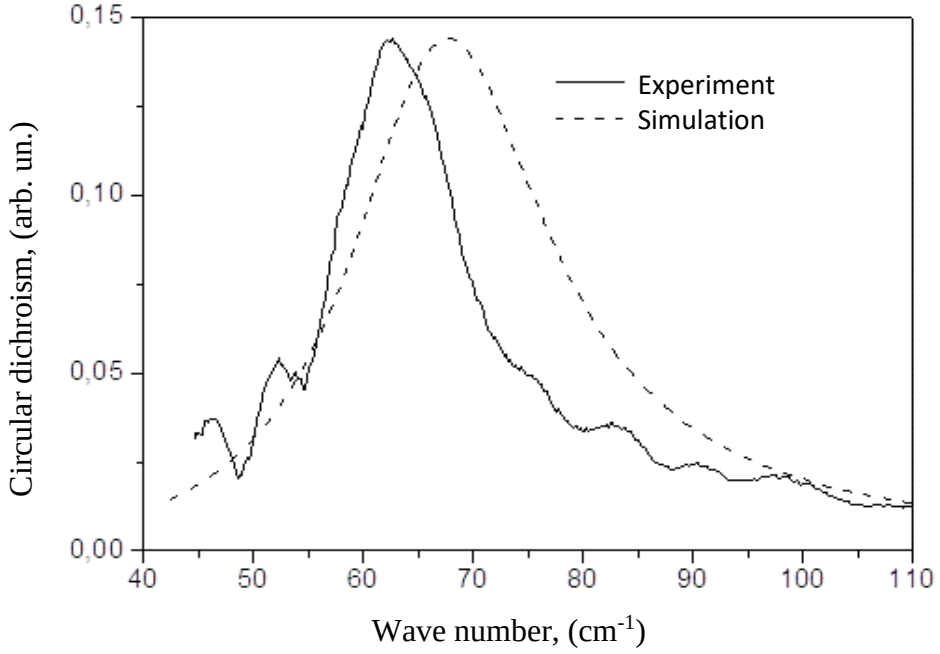


Fig. 6.7.12 − Circular dichroism of the array of left-hand helices. Based on the experiment (a solid line) and simulation results (a broken line) [135]

On comparing the experimental diagrams and simulation results we may conclude that a suggested model gives a reasonable description of the properties of the metamaterial with high chirality. The maxima of faraday rotation angle of the wave and circular dichroism, based on the suggested model, correspond to the ones observed in the experiment. The frequency dependence of calculated values characterizing chiral properties of the medium is in qualitative agreement with experimental data near a resonance.

6.7.8 Low-reflection metamaterials with compensated chirality for terahertz band

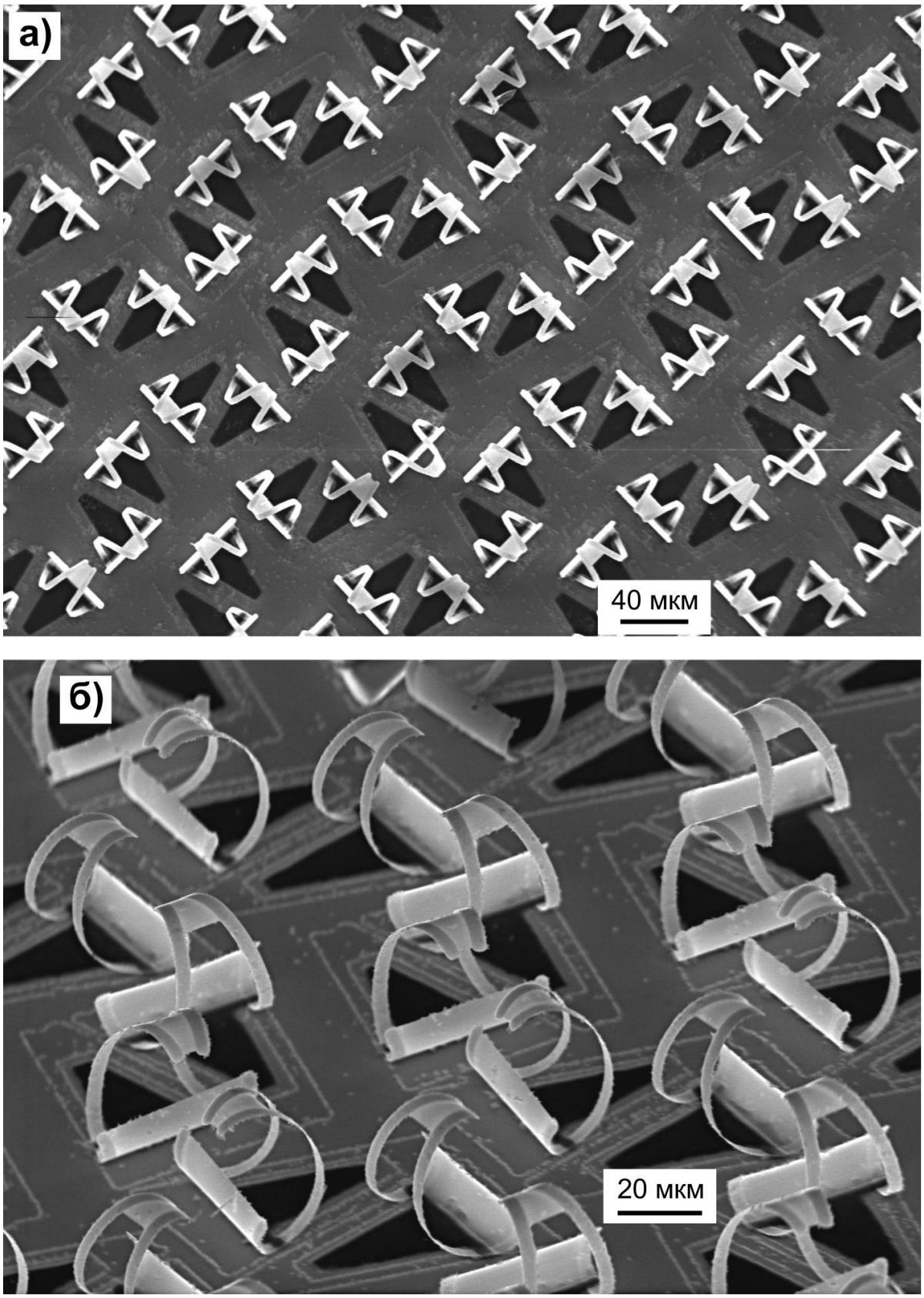
The metamaterial samples which don't possess chiral properties, though made on the basis of helical elements, were produced at the Institute of semiconductor physics SB RAS [136]. As every right helix has a pair – left helix, chiral properties of the sample are compensated. These metamaterials represent a lattice of right and left metallic helices on semiconductor scaffolds. Helices pairs are disposed horizontally and vertically at a sample plane (fig. 6.7.13). The picture was obtained with the help of scanning electron microscopy method.

The parameters of the strip when unrolled are the following: the length is 65 μm, the width is 3 μm. The strips are made of a film $In_{0.2}Ga_{0.8}As/GaAs/Ti/Au$ (16/40/3/65 nm). The rise angle of the helix is 13,5 degrees, the radius is 12,4 μm. The helices concentration in the arrow is $2,3 \cdot 10^{13} m^{-3}$. This rise angle of the helix, which is equal to 13,5º, is optimal for the equality of electric and magnetic helix polarizability, as shown in [120, 121, 123].

Experimental research of the properties of produced metamaterials was carried out at the Institute of semiconductor physics SB RAS.

We used computerized simulation of the metamaterials properties and also made the comparison with experimentaly obtained reflection indices and E-M radiation transmission in THz band.

The metamaterial obtained on the basis of double helix array show similar significant dielectric and magnetic propertiesdue to the optimum helix shape. At the same time chiral properties of artificial structure are compensated because double optimal helices with right and left directions of swirl are used. As a result, the obtained metamaterial has wave impedance in terahertz band which is similar to the free space impedance. Waves reflection coefficient of such metamaterial is small.



a) top view; b) at an angle

Fig. 6.7.13 – SEM-photograph of an array of formed InGaAs/GaAs/Ti/Au single-turn helices [136]

Figures 6.7.14 –6.7.15 show the simulation data and the comparison with experiment of normalized reflection and transmission coefficients. The structure parameters for simulation have been chosen according to the experimental samples.

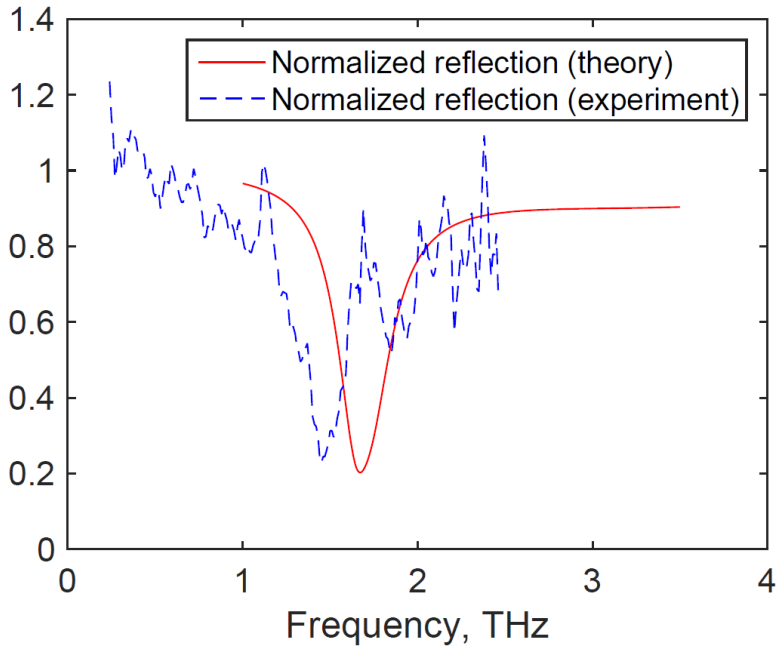


Fig. 6.7.14 – The frequency dependence of its wave reflection of the “metamaterial-substrate” structure

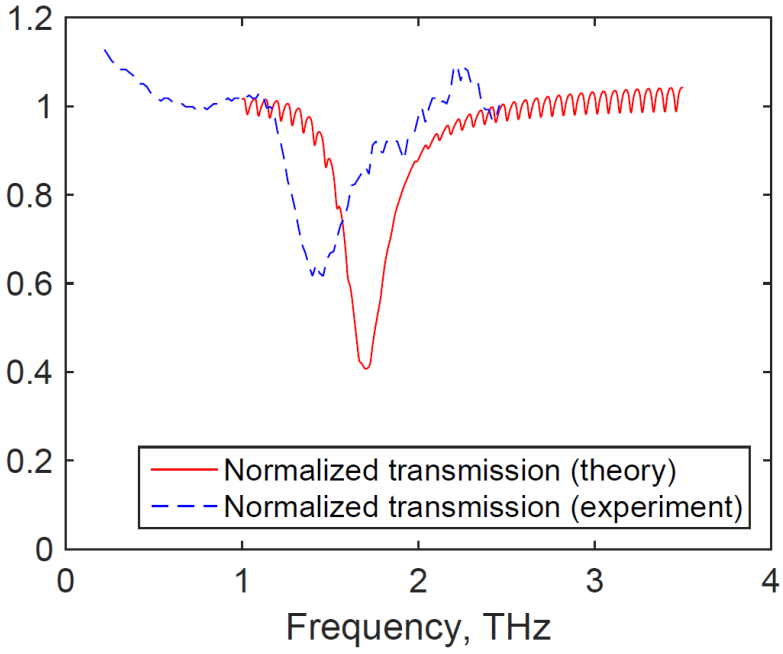


Fig. 6.7.15 – The frequency dependence of its wave progress through the structure “metamaterial-substrate”

The simulation data prove almost complete agreement of the results obtained on the basis of theory and experiments.

References

1. Maskevich S.A., 2010. Atomic physics. Workshop on problem solving: a tutorial, Minsk: High Sc., 455 p. (in Russian).

2. Kiselev V.F., Kozlov S.N., Zoteev A.V., 1999. Fundamentals of Physics of the Surface of a Solid Body, MSU, 387 p. (in Russian).

3. Martin B. R., 2009. Nuclear and Particle Physics : An Introduction 2-nd edition, WILEY.

4. Gaponenko, S.V. 2010. Introduction to Nanophonics, New York : Camridge University Press. 465 p.

5. Ogrin Yu.F., Lutsky VN., Sheftal P.M., Arifova M.U., Elinson MI. 1967 Radio engineering and electronics 12 (4): p.748.

6. Born M. and Wolf E. 1959 Principles of Optics, Pergamon Press, London.

7. Selenyi P. 1913 Comp. Rend., 157: 1408.

8. Rigneault H., Lourtioz J. M., Delalande C., Levenson A., 2005. Nanophotonics. *ISTE Ltd*, 324p.

9. Synge E. H., 1928. Phil. Mag. 6: 356.

10. Leviatan Y. 1986. J. Appl. Phys. 60: 5.

11. Bethe H. A. 1944. Phys. Rev. 66(8): 163–182.

12. Bouwkamp C. J. 1950. Philips Recs. Rep. 5: 401–422.

13. Ohtsu M., 2008 Principles of nanophotonics. (Series in optics and optoelectronics) by Taylor & Francis Group, LLC.

14. Zhdanov G.S., Libenson M.N., Martsinovsky G.A. 1998. UFN 168(7): 801-804.

15. Cline J A., Isaacson M., 1995.Appl Opt. 34: 4869.

16. Betzig E., Finn P. L., Weiner J. S. 1992 Appl. Phys. Lett. 60: 2484.

17. Nano-Optics S. Kawata, M. Ohtsu, M. Irie, (Eds.) Springer Series in Optical SCIENCES Springer-Verlag Berlin Heidelberg 2002. 321 P.

18. Paras N. Prasad, 2004. Nanophotonics A JOHN WILEY & SONS, INC., PUBLICATION. 415 P.

19. Girardy C. Dereux A., 1996. Rep. Prog. Phys. 59: 657-699.

20. Novotny L., Hecht B. 2006. Principles of Nano-optics, Cambridge.

21. Novotny L., Pohl D. W. 1995. Light propagation in scanning near-field optical microscopy Photons and Local Probes (NATO ASI Series E) ed O Marti and R Moller (Dordrecht: Kluwer).

22. Novotny L., Hafner C. 1994. Light propagation in a cylindrical waveguide with a complex metallic, dielectric function. Phys. Rev E 50: 4094-4106.

23. Regli P, 1992 Automatische Wahl der sphaerischen ntwicklungsfunktionen fuËr die 3D-MMP Methode PhD Thesis ETH ZuËrich, Switzerland.

24. Ohtsu M. 2014. Dressed Photons Concepts of Light–Matter Fusion Technology, Springer-Verlag Berlin Heidelberg.

25. Kobayashi K., Sangu S., Ito H., Ohtsu M., 2001, Near-field optical potential for a neutral atom. Phys. Rev. A 63, 013806.

26. Ohtsu M., Kobayashi K., 2004 Optical Near Fields, Springer, Berlin, pp. 109-120.

27. Tanaka Y., Kobayashi K., 2008. J. Microsc. 229: 228.

28. Sato A., Tanaka Y., Minami F.. Kobayashi K., 2009. J. Luminescence 129: 1718.

29. Kawazoe T., Fujiwara H., Kobayashi K., Ohtsu M., 2009. J. of Selected Topics in Quantum Electron. 15: 1380.

30. Yukutake S., Kawazoe T., Yatsui T., Nomura W., Kitamura K.. Ohtsu M., 2010. Appl. Phys. B 99: 415.

31. Kawazoe T., Mueed A., Ohtsu M., 2011. Appl. Phys. B 104: 747.

32. Kawazoe T., Ohtsu M., Akahane K., Yamamoto N., 2012. Appl. Phys. B 107: 659.

33. Henley E., Therring W., 1962 Elementary quantum field theory. MEGRAW-HIIL BOOK COMPANY, Ney York-San Francisco-Toronto-London.

34. Minkin V. I., 2008. Russ. Chem. Bull., Int. Ed., 57 (4), 687-717.

35. Rambidi N. G., 2007. Nanotechnology and molecular computers. Moscow: Phizmatlit.

36. Joachim C., Gimzewski J. K., Aviram A., 2000. Nature, 408, 541-548.

37. Denning P. J., Metcalfe R. M., 1997. Beyond calculation: the next fifty years of computing. New York: Springer.

38. Rouvray D., 2000. Chemistry in Britain, 36 (7), 26-29.

39. Ellenbogen J. C., Love J. C., 2000. Proc. IEEE, 88 (3), 386-426.

40. Kuekes P. J., Stewart D. R., Williams R. S., 2005. J. Appl. Phys., 97, 034301-1-034301-5.

41. Ivanov V. A., Aminov T. G., Novotvortsev V. M., Kalinnikov V. T., 2004. Russ. Chem. Bull., Int. Ed., 53 (11), 2357-2405.

42. Photochromism: Memories and Switches, Chem. Rev. (Spec. Issue), 2000, 100, 1683—1890.

43. Guglielmetti R., Crano J. C., 1999, 2000. Organic Photochromic and Thermochromic Compounds, Vol. 1, 2, Main Photochromic Families, Plenum Press.

44. Barachevskiy V. A., Lshkov G.I., Tsehomskiy V.A., 1977. Photochromism and its directions. Moscow: Khimiya.

45. Akimov D. A., Fedotov A. V., Koroteev N. I., Levich E. V., Magnitskii S. A., Naumov A. N., Sidorov-Biryukov D. A., Sokolyuk N. T., Zheltikov A. M., 1997. Opt. Mem. Neural Networks, 6, 31-48.

46. Irie M., 2000. Chem. Rev., 100 (5), 1685-1716.

47. Matsuda K., Irie M., 2004. J. Photochem. Photobiol. C, 5 (2), 169-182.

48. Stoddart J. F., Heath J. R., 2000. Science, 289 (5482), 1172-1175.

49. de Silva A. P., McClenaghan N. D., 2004. Chem. Eur. J., 10 (3), 574-586.

50. Credi A., 2007. Angew. Chem. Int. Ed., 46 (29), 5472-5475.

51. de Silva A.P., Uchiyama S., Vance T. P., Wannalerse B., 2007. Coord. Chem. Rev., 251 (13-14), 1623-1632.

52. Raymo F. M., Giordani S., 2002. Proc. Natl. Acad. Sci. USA, 99 (8), 4941-4944.

53. Parthenopoulos D. A., Rentzepis P. M., 1989. Science, 245 (4920), 843-845.

54. Klauk H., 2006. Organic Electronics. Weinheim: Wiley VCH.

55. Hayakawa R., Petit M., Higashiguchi K., Matsuda K., Chikyow T., Wakayama Y., 2015. Organic Electronics, 21,149–154.

56. MacDiarmid A. G., 1993. Conjugated Polymers and Related Materials, Ch. 7. Oxford: Oxford Univ. Press.

57. Kahn O., 1999. Chemistry in Britain, 35, 24.

58. Benard S., Riviere E., Yu P., 2001. Chem. Mater., 13 (1), 159-162.

59. Aldoshin S. M., Sanina N. A., Minkin V. I., Voloshin N. A., Ikorskii V.N., Ovcharenko V. I., 2007. J. Mol. Struct., 826 (2-3), 69-74.

60. Quantum computing, 2001. Wikipedia, The Free Encyclopedia. Available from: https://en.wikipedia.org/wiki/ [accessed 19 January 2017].

61. Barachevsky V. A., 2015. Org. Photon. Photovolt., 3 (1), 8-41.

62. Vasilyuk G. T., Maskevich S. A., German A. E., Sveklo I. F., Lukíyanov B. S., and Ageev L. A., 2009. High Energy Chemistry, 43 (7), 521–526.

63. Vasilyuk G.T., Maskevich S.A., Askirka V.F., German A.E., Yaroshevich A.A., Yasinsky V.M., Sveklo I.F., Barachevsky V.A., Ait A.O., Kobeleva O.I., Valova T.M., Yarovenko V.N., Krayushkin M.M., 2017. Journal of Applied Spectroscopy, 84 (4), 570–577.

64. Vasilyuk G.T., Maskevich S.A., Askirka V.F., Lavysh A.V., Kurguzenkov C.A., Yasinsky V.M., Kobeleva O.I., Valova T.M., Ait A.O., Barachevsky V.A., Yarovenko V.N., Krayushkin M.M., 2017. Journal of Applied Spectroscopy, 84 (5), 710–719.

65. Barachevsky V.A., Kobeleva O.I., Ayt A.O., Gorelik A.M., Venidiktova O.V., Krayushkin M.M., Tameev A.R., Sigeikin G.I., Saveliev M.A., Vasilyuk G.T., 2015. Proc. IEEE., 2802, 124–129.

66. Barachevsky V.A., Kobeleva O.I., Ayt A.O., Gorelik A.M., Valova T.M., Krayushkin M.M., Yarovenko V.N., Levchenko K.S., Kiyko V.V., Vasilyuk G.T., 2013. Optical Materials, 35, 1805–1809.

67. Metamaterial [Electronic resource]: [article] from Wikipedia, the free encyclopedia // Available from: https://en.wikipedia.org/wiki/Metamaterial - [accessed 12.10.2012].

68. Bose, J. C. On the rotation of plane of polarization of electric waves by a twisted structure / J. C. Bose // Royal Society of London: proceedings, 1898. – Vol. 63. – P. 146–152.

69. Emerson, D. T. The Work of Jagadis Chandra Bose: 100 Years of Millimeter-Wave Research / D. T. Emerson // IEEE Transactions on Microwave Theory and Techniques. – 1997. – Vol.45, № 12. – P. 2267–2273.

70. Lindman, K. F. Öfversigt af Finska Vetenskaps-Societetens förhandlingar / K. F. Lindman // A. Matematik och naturvetenskaper. – 1914. – Vol. LVII, № 3. – P. 1.

71. Brown, J. Artificial dielectrics having refractive indices less than unity / J. Brown // Institution of Electrical Engineers : proceedings, London. – 1953. – № 100. – P. 51–62.

72. Rotman, W. Plasma simulation by artificial dielectrics and parallel-plate media / W. Rotman // Antennas and Propagation IRE Transactions. – 1962. – Vol.10, №1. – P. 82–95.

73. Magnetism from conductors and enhanced nonlinear phenomena / J. B. Pendry [et al.] // IEEE Transactions on Microwave Theory and Techniques. – 1999. – Vol.47, № 11. – P. 2075–2084.

74. Schelkunoff, S.A. Antennas: Theory and Practice / S.A. Schelkunoff, H.T.Friis – N.Y.: John Willey & Sons, 1952. – 584 p.

75. Composite medium with simultaneously negative permeability and permittivity / D.R. Smith [et al.] // Physical Review Letters. – 2000. – Vol.84, № 18. – P. 4184–4187.

76. Shelby, R. A. Experimental verification of a negative index of refraction / R. A. Shelby, D. R. Smith, S. Schultz // Science. – 2001. – 292. – P. 77–79.

77. Simultaneous negative phase and group velocity of light in a metamaterial / G. Dolling [et al.] // Science. – 2006. – 312. – P. 892–894.

78. Negative-index metamaterial at 780 nm wavelength / G. Dolling [et al.]. // Optics Letters. – 2007. – 32. – P. 53–55.

79. Veselago, V. G. Electrodynamics of substances with simultaneously negative values epsilon and nu / V. G. Veselago // UFN. – 1967. – V. 7. – P. 517–526.

80. Agranovich, V. M. Spatial dispersion and negative refraction of light / V. M. Agranovich, Yu. N. Gartstein // UFN. – 2006. – V. 176, № 10. – P. 1051-1068.

81. Sivukhin, D.V. On the energy of the electromagnetic field in dispersing media / D. V. Sivukhin // Optics and Spectroscopy. –1957.– V. 3. - P. 308-312.Pafomov, V.E. On the issue of transition radiation and Vavilov-Cherenkov radiation / V.E. Pafomov // Journal of Experimental and Theoretical Physics. - 1959. - V. 36. - P. 1853-1858.

82. Pafomov, V.E. On the issue of transition radiation and Vavilov-Cherenkov radiation / V.E. Pafomov // Journal of Experimental and Theoretical Physics. - 1959. - V. 36. - p. 1853-1858.

83. Pafomov, V.E. Cherenkov radiation in anisotropic ferrites / V. E. Pafomov // Journal of Experimental and Theoretical Physics. - 1956. - V. 30. – P. 761-765.

84. Pafomov, V.E. Radiation from an electron crossing a plate V. E. Pafomov // Journal of Experimental and Theoretical Physics. - 1957. - V. 33. - P. 1074-1075.

85. Mandelstam, L.I. Group velocity in the crystal lattice / L. . Mandelstam // Journal of Experimental and Theoretical Physics. - 1945. - V. 15. - P. 475-478.

86. Mandelstam, L.I. A complete collection of works / L.I. Mandelstam. - M.: Publishing House of the Academy of Sciences of the USSR, 1950. - V. 5. - 470 p.

87. Mandelstam, L.I. Lectures on optics, the theory of relativity and quantum mechanics / L. I. Mandelstam. - M .: Science, 1972. - 440 p.

88. Agranovich, V. M. Crystal optics with allowance for spatial dispersion and the theory of excitons / V. M. Agranovich, V. L. Ginzburg. - M .: Science, 1965. - 376 p.

89. McDonald, K. T. Negative Group Velocity / K. T. McDonald // American Journal of Physics. – 2001. – Vol. 69. – P. 607-614.

90. Lamb, H. On group-velocity / H. Lamb // Proceedings of the London Mathematical Society. – 1904. – P. 473-479.

91. Laue, M. Die Fortpflanzung der Strahlung in dispergierenden Medien. (The Propagation of Radiation in Dispersive and Absorbing Media.) / M. Laue // Annals of Physics. – 1905. – Vol. 18. – P. 523-566.

92. Veselago, V. G. Waves in metamaterials: their role in modern physics / V. G. Veselago // Uspekhi Fiz. - 2011. - V. 181. - P. 1201-1205.

93. Schuster, A. An Introduction to the Theory of Optics. / A. Schuster. – London: Edward Arnold and Co. – 1928. – 397 p.

94. Pocklington, H. C. Growth of a wave-group when the group velocity is negative / H. C. Pocklington // Nature. – 1905. – Vol. 71. – P. 607-608.

95. Pendry, J. B. Negative refraction makes a perfect lens, / J. B. Pendry // Physical Review Letters. – 2000. – 85. – P. 3966–3969.

96. Metamaterials 2007: proceedings of 1st International Congress on Advanced Electromagnetic Materials in Microwaves and Optics, Rome, Italy, 22-26 October, 2007 / University "Roma Tre" ; ed.: F. Bilotti, L. Vegni. – Rome, Italy, 2007. – 961 p.

97. Grbic, A. Overcoming the Diffraction Limit with a Planar Left-Handed Transmission-Line Lens / A. Grbic, G. Eleftheriades // Physical Review Letters. – 2004. – 92. – P. 1–4.

98. Saturation of the Magnetic Response of Split-Ring Resonators at Optical Frequencies / J. Zhou [et al.] // Physical Review Letters. – 2005. – P. 95.

99. Negative index of refraction in optical metamaterials / V. M. Shalaev [et al.] // Optics Letters. – 2005. – 30. – P. 3356–3358.

100. Near-infrared double negative metamaterials / S. Zhang [et al.] // Optics Express. – 2005. – 13. – P. 4922–4930.

101. Experimental Demonstration of Near-Infrared Negative-Index Metamaterials / S. Zhang [et al.] // Physical Review Letters. – 2005. – 95. – P. 1–4.

102. Shatrov, A.D. Artificial two-dimensional isotropic medium with negative refraction / A.D. Shatrov // Abstracts of Reports and Communications of

the II scientific and technical conference "Physics and technical applications of wave processes." - Samara, 2003. - P. 4–6.

103. Lagarkov, A.N. Electrodynamic properties of simple bodies from materials with negative magnetic and dielectric permeabilities / A.N. Lagarkov, V.N. Kisel // Reports of the Academy of Sciences. - 2001. - V. 377, No. l. - p. 40–43.

104. Kraftmakher, G. A. Compositional environment with simultaneously negative dielectric and magnetic permeabilities / G. A. Kraftmakher, V. S. Butylkin // ZhTPF Letters. - 2003. - V. 29, no. 6. - p. 26−32.

105. Kuzmiak, V. Scattering properties of cylinder fabricated from a left-handed material / V. Kuzmiak, A. A. Maradudin // Physical Review B. − 2002. − Vol. 66, 045116. − P. 1−7.

106. Negative permittivity and permeability of gold square nanosphirals / R. Abdeddaim [et al.] // Applied Physics Letters. − 2009. − T. 94, № 8. − P. 081907/1−081907/3.

107. Shevchenko, V. V. Geometro-optical theory of a flat lens from a chiral metamaterial / V. V. Shevchenko // Radio engineering and electronics. - 2009. - V. 54, № 6. - P. 696−700.

108. Nemec, H. Tunable terahertz metamaterials with negative permeability / H. Nemec // Physical Review B. − 2009. − T. 79, № 24. − P. 241108/1−241108/4.

109. Fabrication of a novel micron scale Y-structure-based chiral metamaterial: simulation and experimental analysis of its chiral land negative index properties in the terahertz and microwave regimes / N. Wongkasem [et al.] // Microsc. Res. and Techn. − 2007. − T. 70, № 6. − P. 497−505.

110. Zhang, C. Negative reflections of electromagnetic waves in a strong chiral medium / C. Zhang, T. J. Cui // Applied Physics Letters. − 2007. − T. 91, № 19. − P. 194101/1−194101/3.

111. Tretyakov, S. Backward-wave regime and negative refraction in chiral composites / S. Tretyakov, A. Sihvola, L. Jylha // Photonics and Nanostructures – Fundamentals and Applications. − 2005. − № 3. − P. 107–115.

112. Maltsev, V.P. Metamaterial on the basis of a two-dimensional two-element lattice of cylinders with surface conductivity along right- and left-handed lines / V.P. Maltsev, A.D. Shatrov // Radioengineering and Electronics. − 2009. − T. 54, No. 7. − p. 832−837.

113. Cai, Optical cloaking with metamaterials / W. Cai [et al.] // Nature Photonics. – 2007. – 1. – P. 224–227.

114. Experimental implementations of masking coatings / A. V. Shelokova [et al.] // Phys. - 2015. - V. 185, №2. - p. 181–206.

115. Metamaterial electromagnetic cloak at microwave frequencies / D. Schurig [et al.] // Science. – 2006. – Vol. 314. – P. 977–980.

116. Pendry, J. B. Controlling electromagnetic fields / J. B. Pendry, D. Schurig, D. R. Smith // Science. – 2006. – Vol. 312. − P. 1780−1782.

117. Sivukhin, D.V. General course of physics. Optics / D.V. Sivukhin. - M.: Science, 1980. - 752 p.

118. Born, M. Optics / M. Born, E. Wolf. - Kharkov-Kiev: ONTI, 1937. - 495 p.

119. Volkenshtein, M. V. Molecular Optics / M. V. Volkenshteyn. - M .: Gostekhteorizdat, 1951. - 744 p.

120. Semchenko, I. V. Transformation of polarization of electromagnetic waves by spiral radiators / I. V. Semchenko, S. A. Khakhomov, A. L. Samofalov // Radioengineering and Electronics. - 2007. - V. 52, № 8. - p. 917–922.

121. Semchenko, I. V. Radiation of Circularly Polarized Electromagnetic Waves by the Artificial Flat Lattice with Two-Turns Helical Elements / I. V. Semchenko, S. A. Khakhomov, A. L. Samofalov // Bianasotropics' 2004: proceedings of the 10th International Conference on Complex Media and Metamaterials, Het Pand, Chent, Belgium, 22-24 September 2004. – Het Pand, Chent, Belgium, 2004. – P. 236-239.

122. Electromagnetics of bianisotropic materials: Theory and Applications / A. N. Serdyukov [et al.].– Gordon and Breach Publishing Group [etc.]: London, 2001. – 337 p.

123. Semchenko, I. V. Optimal helix shape: equality of dielectric, magnetic and chiral susceptibilities / I.V. Semchenko, S.A. Khakhomov, A.L. Samofalov // Proceedings of higher educational institutions. Physics. - 2009. - V. 52, № 5. - P. 30–36.

124. Fedorov, F. I. Theory of gyrotropy / F. I. Fedorov. - Minsk: Science and technology, 1976. - 452 p.

125. Semchenko, I. V. Chiral metamaterial with unit negative refraction index/ I. V. Semchenko, S. A. Khakhomov, S. A. Tretyakov // 1st International Congress on Advanced Electromagnetic Materials in Microwaves and Optics, Metamaterials 2007: proceedings, Rome , Italy, 22-24 October 2007. – Rome, Italy, 2007. – P. 218-221.

126. Semchenko, I. V. Chiral metamaterial with unit negative refraction index/ I. V. Semchenko, S. A. Khakhomov, S. A. Tretyakov // The European

Physical Journal. Applied Physics. – 2009. – Vol. 46, № 3. – P. 32607-1-32607-4.

127. Semchenko, I. V. Polarization plane rotation of electromagnetic waves by the artificial periodic structure with one-turn helical elements / I. V. Semchenko, S. A. Khakhomov, A. L. Samofalov // Electromagnetics. – 2006. – Vol. 26, №. 3–4. – P. 219–233.

128. Semchenko, I. V. Polarization Plane Rotation of Electromagnetic Waves by the Artificial Periodic Structure with One-Turn Helical Elements / I. V. Semchenko, S. A. Khakhomov, A. L. Samofalov // Bianasotropics' 2004: proceedings of the 10th International Conference on Complex Media and Metamaterials, Het Pand, Chent, Belgium, 22-24 September 2004. – Het Pand, Chent, Belgium, 2004. – P. 74-77.

129. Semchenko, I. V. Realistic Spirals of Optimal Shape for Electromagnetic Cloaking / I. V. Semchenko, S. A. Khakhomov, A. L. Samofalov // 2nd International Congress on Advanced Electromagnetic Materials in Microwaves and Optics, Metamaterials 2008: proceedings, Pamplona, Spain, 21-26 September 2008. – Pamplona, Spain, 2008. – P. 1-3.

130. Semchenko, I. V. Helices of optimal shape for nonreflecting covering / I. V. Semchenko, S. A. Khakhomov, A. L. Samofalov // The European Physical Journal. Applied Physics. – 2010. – Vol. 49, № 3. – P. 33002-p1-33002-p5.

131. Broadband Reflectionless Metasheets: Frequency-Selective Transmission and Perfect Absorption / V. S. Asadchy, I. A. Faniayeu, Y. Ra'di, S. A. Khakhomov, I. V. Semchenko, S. A. Tretyakov // Phys. Rev. X. – 2015. – Vol. 5, iss. 3. – P. 031005-1-031005-10.

132. Chiral metamaterials of the terahertz range based on spirals from metal-semiconductor nanofilms / E.V. Naumova [et al.] // Avtometriya. - 2009. - Vol. 45, No. 4. - P.12-22.

133. Free-standing and overgrown InGaAs//GaAs nanotubes, nanohelices and their arrays / Prinz V.Ya. [et al.] // Physica E. – 2000. – Vol. 6, № 1. – P. 828-831.

134. Structure with chiral electromagnetic properties and method of its manufacture (versions): Pat. 2317942 of the Russian Federation: IPC B82B 3/00 (2006) / E. V. Naumova, V. Ya. Prints; date of publication: 27.02.2008.

135. Study of the properties of artificial anisotropic structures with high chirality / I. V. Semchenko, S. A. Khakhomov, E. V. Naumova, V. Ya. Prints, S. V. Golod, V. V. Kubarev // Crystallography. - 2011. - Vol. 56, No. 3. - P. 404–411.

136. Investigation of the properties of low-reflecting metamaterials with compensated chirality / I. V. Semchenko, S. A. Khakhomov, V. S. Asadchiy, E. V. Naumova, V. Ya. Prints, S. V. Golod, A. G. Milekhin, A. M. Goncharenko, G. V. Sinitsin // Crystallography. - 2014, V. 59, № 4. - P. 544–550.

137. Ground-plane-less bidirectional terahertz absorber based on omega resonators / A. Balmakou, M. Podalov, S. Khakhomov, D. Stavenga, I. Semchenko // Optics Letters. –2015, May 1. – Vol. 40, № 9. – P. 2084–2087.

138. Sihvola, A. H. View on the history of electromagnetics of metamaterials: Evolution of the congress series of complex media / A. H. Sihvola, I. V. Semchenko, S. A. Khakhomov // Photonics and Nanostructures - Fundamentals and Applications. – 2014. – 12. – P. 279–283.

139. The potential energy of non-resonant optimal bianisotropic particles in an electromagnetic field does not depend on time / I. Semchenko, S. Khakhomov, A. Balmakou, S. Tretyakov // The European Physical Journal, EPJ Applied Metamaterials. – 2014. – 1. – P. 1–4.

140. Polarization selectivity of artificial anisotropic structures based on DNA of such helices / I. V. Semchenko, A. P. Balmakov, S. A. Khakhomov // Crystallography. - 2010. - V. 55, No. 6. - P. 992–998.

141. Modeling of Spirals with Equal Dielectric, Magnetic, and Chiral Susceptibilities / E. Saenz, I. V. Semchenko, S. A. Khakhomov, K. Guven, R. Gonzalo, E. Ozbay, S. Tretyakov // Electromagnetics. – 2008. – Vol. 28, № 7. – P. 476–493.

142. Semchenko, I .V. Effective electron model of the wire helix excitation at microwaves: first step to optimization of pitch angle of helix / I. V. Semchenko, S. A. Khakhomov, E. A. Fedosenko // Advances in Electromagnetics of Complex Media and Metamaterials, ed. by Said Zouhdi, Ari Sihvola and Mohamed Arsalane, Kluwer Academic Publishers. – 2002. – P. 245–258.

143. Bianisotropics'93: proceedings of the Seminar on Electrodynamics of Chiral and Bianisotropic Media, Gomel, Belarus 12-14 October 1993 / ed.: A. Sihvola, S. Tretyakov, I. Semchenko. – 12-14 October 1993, Gomel, Belarus. – Gomel, 1993. – 120 p.

144. Semchenko, I. V. Optimal Shape of Spiral: Equality of Dielectric, Magnetic and Chiral Properties / I. V. Semchenko, S. A. Khakhomov, A. L. Samofalov // META'08, Metamaterials for Secure Information and Communication Technologies: proceedings, Marrakesh, Morocco, 7−10 May 2008.− Paris, France, 2008. – P. 71–80.

145. Bokut, B. V. Special waves in natural gyrotropic media / B. V. Bokut, V. V. Gvozdev, A. N. Serdyukov // Journal of Applied Spectroscopy. - 1981. - 34. - p. 701–706.

146. Bokut, B. V. To the phenomenological theory of natural optical activity / B. V. Bokut, A. N. Serdyukov // Journal of Experimental and Theoretical Physics. - 1971. - V. 61, № 5. - P. 1808 - 1813.

147. Sihvola, A. H, Bi-isotropic constitutive relations / A. H. Sihvola, I. V. Lindell // Microwave and Optical Technology Letters. – 1991. – Vol. 4, № 8. – P. 195-297.