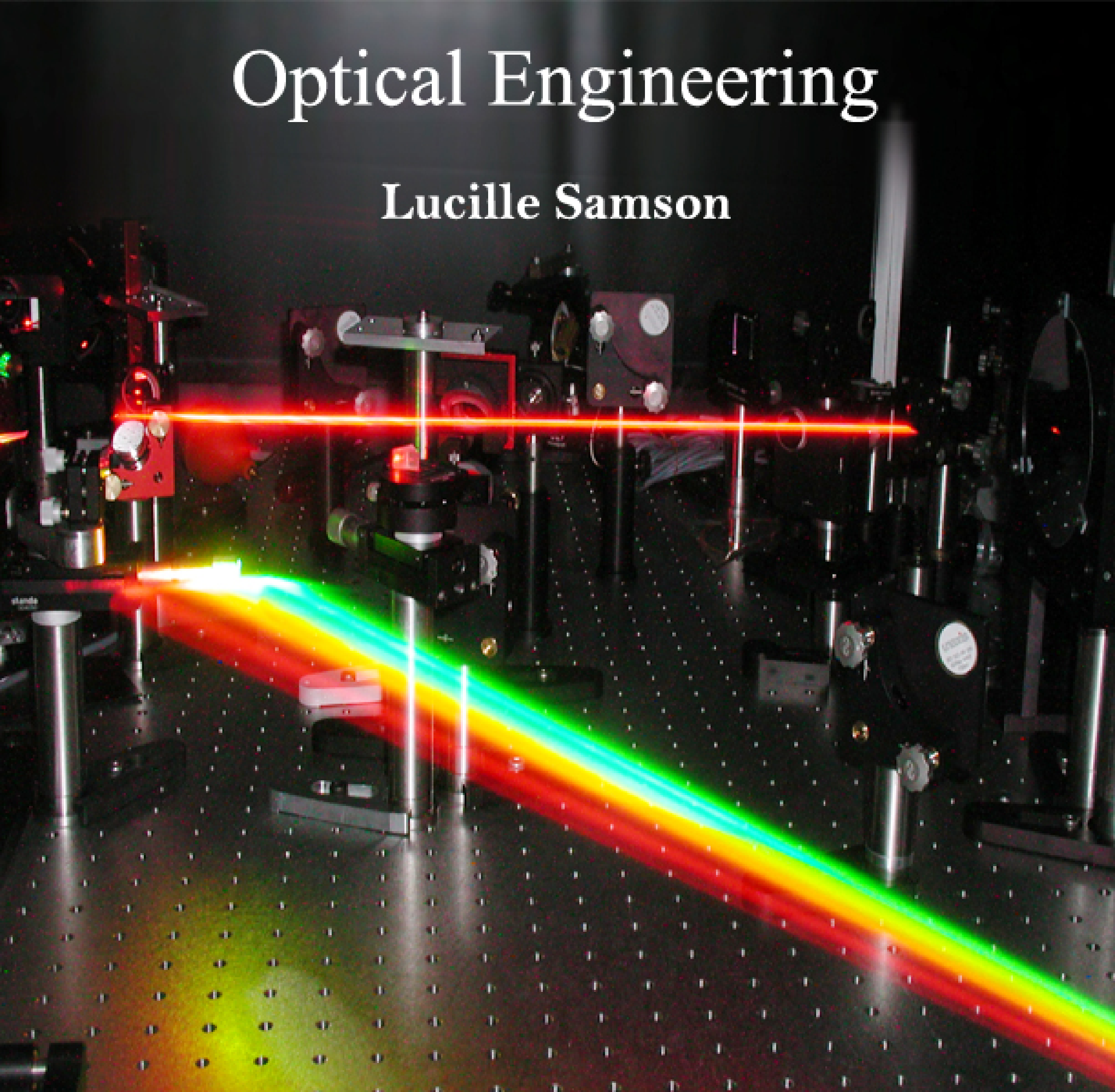
Optical Engineering

Lucille Samson



First Edition, 2012

ISBN 978-81-323-3092-9

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Introduction

Optical engineering is the field of study that focuses on applications of optics.

Optical engineers design components of optical instruments such as lenses, microscopes, telescopes, and other equipment that utilize the properties of light. Other devices include optical sensors and measurement systems, lasers, fiber optic communication systems, optical disc systems (e.g. CD, DVD), etc.

Because optical engineers want to design and build devices that make light do something useful, they must understand and apply the science of optics in substantial detail, in order to know what is physically possible to achieve (physics and chemistry). However, they also must know what is practical in terms of available technology, materials, costs, design methods, etc. As with other fields of engineering, computers are important to many (perhaps most) optical engineers. They are used with instruments, for simulation, in design, and for many other applications. Engineers often use general computer tools such as spreadsheets and programming languages, and they make frequent use of specialized optical software designed specifically for their field.

Optical engineering metrology uses optical methods to measure micro-vibrations with instruments like the laser speckle interferometer or to measure the properties of the various masses with instruments measuring refraction.

Ancient History of Optical Engineering

4,000 years ago there were some signs and indications that early optical engineers used optical applications. People who designed and built the Stonehenge and Pyramid of Cheops used basic optical engineering principles. These structures had a connection with the earth and sun. These early engineers knew light travels in straight lines and understood the cycle of the seasons, which made these structures relative to the calendar and the compass. In 350 BC, Plato and Aristotle argued about the accurate nature of light. Plato thought vision was achieved by the discharge of optical beams from the eyes.

Aristotle believed vision is accomplished when particles from the object releases into the pupil of the eye. In 300 BC, Euclid, who wrote and studied optics and geometry, wrote the book *Optics*, which heavily contributed to the study of the science of optics.

Chapter 1

Optical Lens Design

Optical lens design refers to the calculation of lens construction parameters (variables) that will meet a set of performance requirements and constraints, including cost and schedule limitations.

Construction parameters include surface profile types (spherical, aspheric, holographic, diffractive, etc.), and the parameters for each surface type such as radius of curvature, distance to the next surface, glass type and optionally tilt and decenter.

Design requirements

Performance requirements can include:

1. Optical performance, i.e., image quality: quantified by encircled energy, modulation transfer function, Strehl ratio, ghost reflection control, and pupil performance (size, location and aberration control); the choice of the image quality metric is application specific.
2. Physical requirements such as weight, static volume, dynamic volume, center of gravity and overall configuration requirements.
3. Environmental requirements: ranges for temperature, pressure, vibration and electromagnetic shielding.

Design constraints can include realistic lens element center and edge thicknesses, minimum and maximum air-spaces between lenses, maximum constraints on entrance and exit angles, physically realizable glass index of refraction and dispersion properties.

Manufacturing costs and delivery schedules are also a major part of optical design. The price of an optical glass blank of given dimensions can vary by a factor of fifty or more, depending on the size, glass type, index homogeneity quality, and availability, with BK7 usually being the cheapest. Costs for larger and/or thicker optical blanks of a given material, above 100mm to 150mm or so, usually increase faster than what would be proportional to just the increase in physical volume. This is primarily due to increased blank annealing time required to achieve acceptable index homogeneity and internal stress birefringence levels throughout the blank volume. Availability of glass blanks is driven by how frequently a particular glass type is mixed and poured by a given manufacturer, and can seriously affect manufacturing cost and schedule.

Process

Lenses can first be designed using paraxial theory to position images and pupils, then real surfaces inserted and optimized. Paraxial theory can be skipped in simpler cases and the lens directly optimized using real surfaces. Lenses are first designed using average index of refraction and dispersion properties published in the glass manufacturer's catalog and though glass model calculations. However, the properties of the real glass blanks will vary from this ideal; index of refraction values can vary by as much as 0.0003 or more from catalog values, and dispersion can either remain about the same or vary slightly. These changes in index and dispersion can sometimes be enough to affect the lens focus location and imaging performance in highly corrected systems.

The lens blank manufacturing process is as follows:

1. The glass batch ingredients for a desired glass type are mixed together in a powder state,
2. the powder mixture is melted together in a furnace,
3. the fluid is further mixed while molten to maximize batch homogeneity,
4. poured into lens blanks and
5. annealed according to empirically determined time-temperature schedules.

The glass blank pedigree, or "melt data", can be determined for a given glass batch by making small precision prisms from various locations in the batch and measuring their index of refraction on a spectrometer, typically at five or more wavelengths. Lens design programs have curve fitting routines that can fit the melt data to a selected dispersion curve, from which the index of refraction at any wavelength within the fitted wavelength range can be calculated. A re-optimization, or "melt re-comp", can then be performed on the lens design using measured index of refraction data where available. When manufactured, the resulting lens performance will more closely match the desired requirements than if average glass catalog values for index of refraction were assumed.

Delivery schedules are impacted by glass and mirror blank availability and lead times to acquire, the amount of tooling a shop must fabricate prior to starting on a project, the manufacturing tolerances on the parts (tighter tolerances mean longer fab times), the complexity of any optical coatings that must be applied to the finished parts, further complexities in mounting or bonding lens elements into cells and in the overall lens system assembly, and any post-assembly alignment and quality control testing and tooling required. Tooling costs and delivery schedules can be reduced by using existing tooling at any given shop wherever possible, and by maximizing manufacturing tolerances to the extent possible.

Lens optimization

Optical design is partly a science because ray paths and wavefront structure can be very accurately calculated anywhere along the propagation path through the lens. Glass and coating optical properties can be measured and modeled with sufficient precision for use in lenses. If tolerances are included during the design, parts can usually be manufactured accurately enough that the resulting lens assembly performs acceptably close to the paper design.

Optical design is also partly an art, though, as the multi-dimensional design space within which a constrained lens design is free to roam is literally beyond human imagination if more than a few construction parameters are free to vary. The number, type and placement of optical elements are partly driven by physical requirements, but are also often based on previous similar designs obtained from published data, patents and textbooks. Skill and intuition in lens design are acquired over years of experience spanning hundreds to thousands of different lens design projects, preferably leading to additional experiences (and headaches) dealing with fabricating and aligning systems.

As an example of the complexity of lens-design space, a simple two-element air-spaced lens has nine variables (four radii of curvature, two thicknesses, one airspace thickness, and two glass types). Even for this simplest case, the design space is thus nine-dimensional, and local or global solutions within this space can at least be imagined as smaller or larger bubbles in a sponge-like 9-D foam-scape. A complex multi-configuration lens corrected over a wide spectral band and field of view, at multiple zoomed focal lengths and over a realistic temperature range, can have an extremely complex design volume, having over a hundred dimensions.

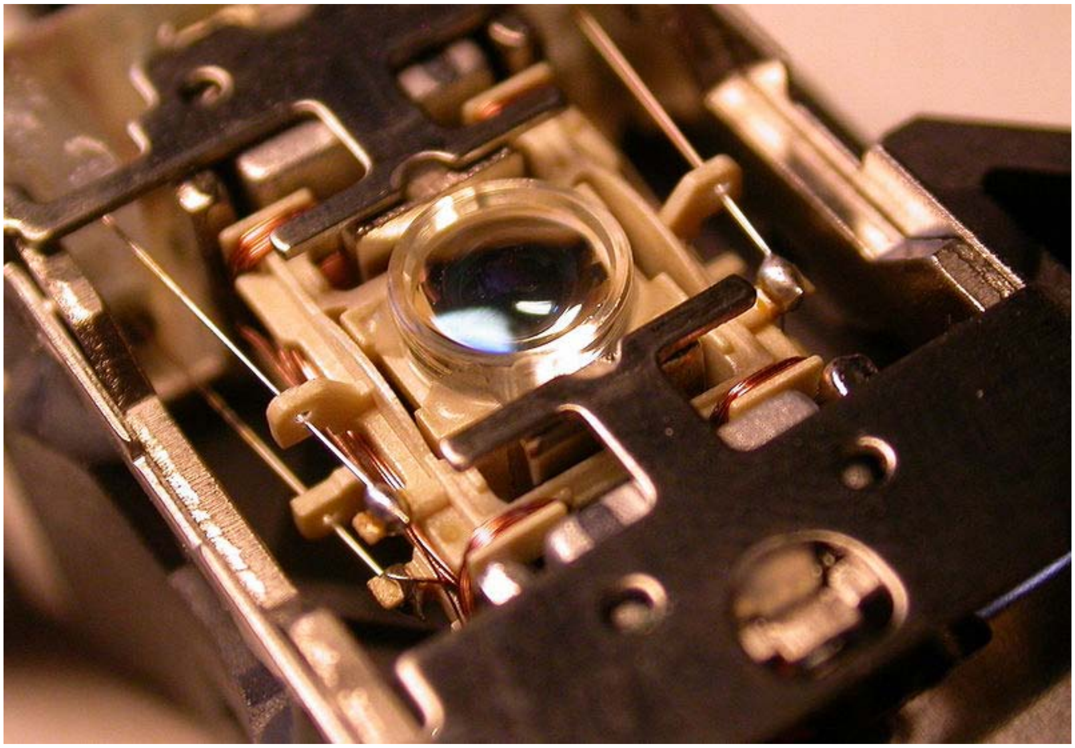
Lens optimization techniques that can navigate this multi-dimensional space and proceed to local minima have been studied since the 1940s, beginning with early work by James G. Baker, and later by D. Feder, Wynne, Glatzel, D. Grey and others. Prior to the advent of digital computers, lens design was an agonizingly slow hand-calculation process requiring high-precision trigonometric and logarithmic tables, reams of paper, plotting 2-D cuts through the multi-dimensional space, and significant patience and understudying from previous masters. Tracing a single ray through a given lens surface could take more than an hour of painstaking calculations and checks, and a lens designer could not design more than a very few complex, high-performance anastigmatic objectives in an entire lifetime.

Modern desktop computers can now raytrace tens to hundreds of millions of rays per second through a lens, and perform hundreds to thousands of optimization cycles per second, rapidly exploring the n-dimensional design volume and even hill-climbing in and out of local minima in the search for the best solution.

However, even with lightning-fast optimizers, seasoned experience is still needed to guide solution trajectories through unacceptably shallow local minima and achieve the desired performance requirements. Experience in the mechanical and physical properties of glass, metals, optical coatings and bonding materials is also needed, especially in systems required to give high sustained performance over wide temperature ranges and harsh environmental conditions.

Chapter 2

Optical Disc



The optical lens of a compact disc drive.

In computing and optical recording, an **optical disc** is a flat, usually circular disc which encodes binary data in the form of pits (binary value of 0 or off, due to lack of reflection when read) and lands (binary value of 1 or on, due to a reflection when read) on a special material (often aluminium) on one of its flat surfaces. The encoding material sits atop a thicker substrate (usually polycarbonate) which makes up the bulk of the disc and forms a dust defocusing layer. The encoding pattern follows a continuous, spiral path covering the entire disc surface and extending from the innermost track to the outermost track. The data is stored on the disc with a laser or stamping machine, and can be accessed when the data path is illuminated with a laser diode in an optical disc drive which spins the disc at speeds of about 200 RPM up to 4000 rpm or more depending on the drive type, disc format, and the distance of the read head from the center of the disc (inner tracks are read at a faster disc speed). The pits or bumps distort the reflected laser light, hence most optical discs (except the black discs of the original PlayStation video game console) characteristically have an iridescent appearance created by the grooves of the reflective layer. The reverse side of an optical disc usually has a printed label, generally made of paper but sometimes printed or stamped onto the disc itself. This side of the disc contains the actual data and is typically coated with a transparent material, usually lacquer. Unlike the 3½-inch floppy disk, most optical discs do not have an integrated protective casing and are therefore susceptible to data transfer problems due to scratches, fingerprints, and other environmental problems.

Optical discs are usually between 7.6 and 30 cm (3 to 12 in) in diameter, with 12 cm (4.75 in) being the most common size. A typical disc is about 1.2 mm (0.05 in) thick, while the track pitch (distance from the center of one track to the center of the next) is typically 1.6 µm.

An optical disc is designed to support one of three recording types: read-only (e.g.: CD and CD-ROM), recordable (write-once, e.g. CD-R), or re-recordable (rewritable, e.g. CD-RW). Write-once optical discs commonly have an organic dye recording layer between the substrate and the reflective layer. Rewritable discs typically contain an alloy recording layer composed of a phase change material, most often AgInSbTe, an alloy of silver, indium, antimony and tellurium.

Optical discs are most commonly used for storing music (e.g. for use in a CD player), video (e.g. for use in a DVD player), or data and programs for personal computers. The Optical Storage Technology Association (OSTA) promotes standardized optical storage formats. Although optical discs are more durable than earlier audio-visual and data storage formats, they are susceptible to environmental and daily-use damage. Libraries and archives enact optical media preservation procedures to ensure continued usability in the computer's optical disc drive or corresponding disc player.

For computer data backup and physical data transfer, optical discs such as CDs and DVDs are gradually being replaced with faster, smaller, and more reliable solid state devices, especially the USB flash drive. This trend is expected to continue as USB flash drives continue to increase in capacity and drop in price. Similarly, personal portable CD players have been supplanted by portable solid state MP3 players, and MP3 music purchased or shared over the internet has significantly reduced the number of audio CDs sold annually.

History

An earlier analog optical disc recorded in 1935 for Licht-Tone Orgel (sampling organ)

The optical disc was invented in 1958. In 1961 and 1969, David Paul Gregg registered a patent for the analog optical disc for video recording. This form of optical disc was a very early form of the DVD U.S. Patent 3,430,966. It is of special interest that U.S. Patent 4,893,297, filed 1989, issued 1990, generated royalty income for Pioneer Corporation’s DVA until 2007 —then compassing the CD, DVD, and Blu-ray Disc systems. In the early 1960s, the Music Corporation of America bought Gregg's patents and his company, Gauss Electrophysics.

Later, in the Netherlands in 1969, Philips Research physicists began their first optical videodisc experiments at Eindhoven. In 1975, Philips and MCA began to work together, and in 1978, commercially much too late, they presented their long-awaited laserdisc in Atlanta. MCA delivered the discs and Philips the players. However, the presentation was a technical and commercial failure and the Philips/MCA cooperation ended.

In Japan and the U.S., Pioneer succeeded with the videodisc until the advent of the DVD. In 1979, Philips and Sony, in consortium, successfully developed the audio compact disc in 1983.

In the mid-1990s, a consortium of manufacturers developed the second generation of the optical disc, the DVD.

The third generation optical disc was developed in 2000-2006, and was introduced as Blu-ray Disk. Developed by the Blu-ray Disc Association (BDA), a group of the world's leading consumer electronics, personal computer and media manufacturers (including Apple, Dell, Hitachi, HP, JVC, LG, Mitsubishi, Panasonic, Pioneer, Philips, Samsung, Sharp, Sony, TDK and Thomson). The format was developed to enable recording, rewriting and playback of high-definition video (HD), as well as storing large amounts of data. The format offers more than five times the storage capacity of traditional DVDs and can hold up to 25 GB on a single-layer disc and 50 GB on a dual-layer disc. This extra capacity combined with the use of advanced video and audio codecs will offer consumers an unprecedented HD experience.

While current optical disc technologies such as DVD, DVD±R, DVD±RW, and DVD-RAM rely on a red laser to read and write data, the new format uses a blue-violet laser instead, hence the name Blu-ray. Despite the different type of lasers used, Blu-ray products can easily be made backwards compatible with CDs and DVDs through the use of a BD/DVD/CD compatible optical pickup unit. The benefit of using a blue-violet laser (405 nm) is that it has a shorter wavelength than a red laser (650 nm), which makes it possible to focus the laser spot with even greater precision. This allows data to be packed more tightly and stored in less space, so it's possible to fit more data on the disc even though it's the same size as a CD/DVD. This together with the change of numerical aperture to 0.85 is what enables Blu-ray Discs to hold 25 GB/50 GB. Recent development by Pioneer has pushed the storage capacity to 500 GB on a single disc by using 20 layers. First movies on Blu-ray discs were released in June 2006. Blu-ray eventually prevailed in a high definition optical disc format war over a competing format, the HD DVD. A standard Blu-ray disc can hold about 25 GB of data, a DVD about 4.7 GB, and a CD about 700 MB.

First-generation

Initially, optical discs were used to store music and computer software. The laser disc format stored analog video signals, but commercially lost to the VHS videotape cassette, due mainly to its high cost and non-re-recordability; other first-generation disc formats were designed only to store digital data and were not initially capable of use as a video medium.

Most first-generation disc devices had an infrared laser reading head. The minimum size of the laser spot is proportional to its wavelength, thus wavelength is a limiting factor against great information density, too little data can be stored so. The infrared range is beyond the long-wavelength end of the visible light spectrum, so, supports less density than any visible light colour. One example of high-density data storage capacity, achieved with an infrared laser, is 700MB of net user data for a 12 cm compact disc.

NOTE: other factors affecting data storage density are, for example, a multi-layered infrared disc would hold more data than an identical single-layer disc; whether CAV, CLV, or zoned-CAV; how the data are encoded; how much clear margin at the center and the edge

- Compact Disc (CD) and derivatives
  - Video CD
  - Super Video CD
- Laserdisc
- GD-ROM
- Phase-change Dual
- Double Density Compact Disc (DDCD)

- Magneto-optical disc
- mini disc

Second-generation

Second-generation optical discs were for storing great amounts of data, including broadcast-quality digital video. Such discs usually are read with a visible-light laser (usually red); the shorter wavelength and greater numerical aperture allow a narrower light beam, permitting smaller pits and lands in the disc. In the DVD format, this allows 4.7 GB storage on a standard 12 cm, single-sided, single-layer disc; alternately, smaller media, such as the MiniDisc and the DataPlay formats, can have capacity comparable to that of the larger, standard compact 12 cm disc.

- Hi-MD
- DVD and derivatives
  - DVD-Audio
  - DualDisc
  - Digital Video Express (DIVX)
- Super Audio CD
- Enhanced Versatile Disc
- DataPlay
- Universal Media Disc
- Ultra Density Optical

Third-generation

Third-generation optical discs are in development, meant for distributing high-definition video and support greater data storage capacities, accomplished with short-wavelength visible-light lasers and greater numerical apertures. The Blu-ray disc uses blue-violet lasers and focusing optics of greater aperture, for use with discs with smaller pits and lands, thereby greater data storage capacity per layer. In practice, the effective multimedia presentation capacity is improved with enhanced video data compression codecs such as H.264, and VC-1.

- Currently shipping:
  - Blu-ray Disc (up to 50 GB)
  - HD VMD Disc
  - CBHD Disc

- In development:
  - Forward Versatile Disc
  - Digital Multilayer Disk or Fluorescent Multilayer Disc

- Abandoned:
  - HD DVD

Fourth-generation

The following formats go beyond the current third-generation discs and have the potential to hold more than one terabyte (1TB) of data:

- Holographic Versatile Disc
- LS-R
- Protein-coated disc

Specifications

Base (1×) and (current) maximum speeds by generation

Generation	Base	Max	
	(Mbit/s)	(Mbit/s)	×
1st (CD)	1.17	65.62	56×
2nd (DVD)	10.55	210.94	20×
3rd (BD)	36	432	12×

Capacity and nomenclature

Designation		Sides	Layers	Diameter	Capacity	
			(total)	(cm)	(GB)	(GiB)
BD	SS SL	1	1	8	7.8	
BD	SS DL	1	2	8	15.6	
BD	SS SL	1	1	12	25	
BD	SS DL	1	2	12	50	
CD–ROM 74 min	SS SL	1	1	12	0.682	0.635
CD–ROM 80 min	SS SL	1	1	12	0.737	0.687
CD–ROM	SS SL	1	1	8	0.194	0.180
DDCD–ROM	SS SL	1	1	12	1.364	1.270
DDCD–ROM	SS SL	1	1	8	0.387	0.360
DVD–1	SS SL	1	1	8	1.46	1.36
DVD–2	SS DL	1	2	8	2.66	2.47
DVD–3	DS SL	2	2	8	2.92	2.72
DVD–4	DS DL	2	4	8	5.32	4.95
DVD–5	SS SL	1	1	12	4.70	4.37
DVD–9	SS DL	1	2	12	8.54	7.95
DVD–10	DS SL	2	2	12	9.40	8.74
DVD–14	DS DL/SL	2	3	12	13.24	12.32
DVD–18	DS DL	2	4	12	17.08	15.90
DVD–R 1.0	SS SL	1	1	12	3.95	3.68

DVD–R (2.0), +R, –RW, +RW	SS SL	1	1	12	4.70	4.37
DVD-R, +R, –RW, +RW	DS SL	2	2	12	9.40	8.75
DVD–RAM	SS SL	1	1	8	1.46	1.36
DVD–RAM	DS SL	2	2	8	2.65	2.47
DVD–RAM 1.0	SS SL	1	1	12	2.58	2.40
DVD–RAM 2.0	SS SL	1	1	12	4.70	4.37
DVD–RAM 1.0	DS SL	2	2	12	5.16	4.80
DVD–RAM 2.0	DS SL	2	2	12	9.40	8.75
HD DVD	SS SL	1	1	8	4.70	
HD DVD	SS DL	1	2	8	9.40	
HD DVD	DS SL	2	2	8	9.40	
HD DVD	DS DL	2	4	8	18.80	
HD DVD	SS SL	1	1	12	15.00	
HD DVD	SS DL	1	2	12	30.00	
HD DVD	DS SL	2	2	12	30.00	
HD DVD	DS DL	2	4	12	60.00	
HD DVD–RAM	SS SL	1	1	12	20.00	

Chapter 3

Optics

Optics includes study of dispersion of light

Optics is the branch of physics which involves the behavior and properties of light, including its interactions with matter and the construction of instruments that use or detect it. Optics usually describes the behavior of visible, ultraviolet, and infrared light. Because light is an electromagnetic wave, other forms of electromagnetic radiation such as X-rays, microwaves, and radio waves exhibit similar properties.

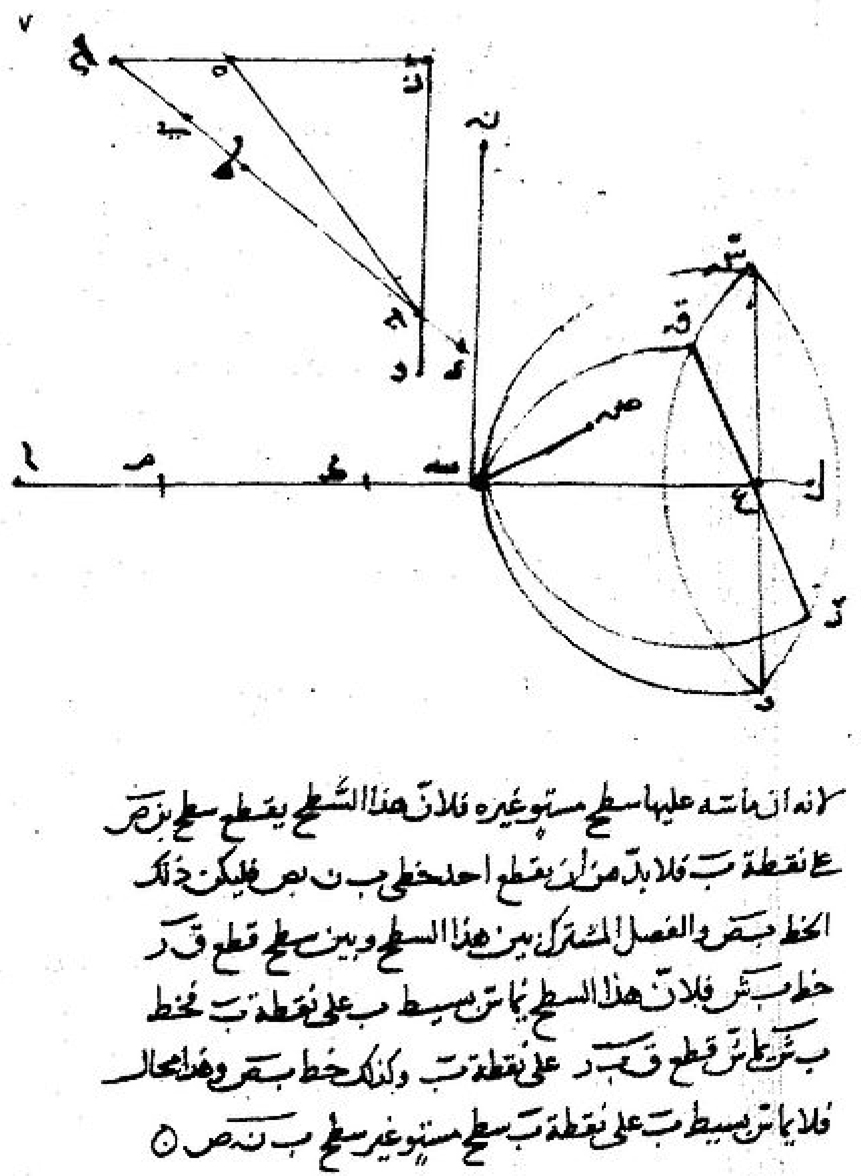
Most optical phenomena can be accounted for using the classical electromagnetic description of light. Complete electromagnetic descriptions of light are, however, often difficult to apply in practice. Practical optics is usually done using simplified models. The most common of these, geometric optics, treats light as a collection of rays that travel in straight lines and bend when they pass through or reflect from surfaces. Physical optics is a more comprehensive model of light, which includes wave effects such as diffraction and interference that cannot be accounted for in geometric optics. Historically, the ray-based model of light was developed first, followed by the wave model of light. Progress in electromagnetic theory in the 19th century led to the discovery that light waves were in fact electromagnetic radiation.

Some phenomena depend on the fact that light has both wave-like and particle-like properties. Explanation of these effects requires quantum mechanics. When considering light's particle-like properties, the light is modeled as a collection of particles called "photons". Quantum optics deals with the application of quantum mechanics to optical systems.

Optical science is relevant to and studied in many related disciplines including astronomy, various engineering fields, photography, and medicine (particularly ophthalmology and optometry). Practical applications of optics are found in a variety of technologies and everyday objects, including mirrors, lenses, telescopes, microscopes, lasers, and fiber optics.

History

Optics began with the development of lenses by the ancient Egyptians and Mesopotamians. The earliest known lenses were made from polished crystal, often quartz, and have been dated as early as 700 BC for Assyrian lenses such as the Layard/Nimrud lens. The ancient Romans and Greeks filled glass spheres with water to make lenses. These practical developments were followed by the development of theories of light and vision by ancient Greek and Indian philosophers, and the development of geometrical optics in the Greco-Roman world. The word *optics* comes from the ancient Greek word *ὀπτική*, meaning *appearance* or *look*. Plato first articulated emission theory, the idea that visual perception is accomplished by rays emitted by the eyes. He also commented on the parity reversal of mirrors in *Timaeus*. Some hundred years later, Euclid wrote a treatise entitled *Optics* wherein he described the mathematical rules of perspective and describes the effects of refraction qualitatively. Ptolemy, in his treatise *Optics*, summarizes much of Euclid and goes on to describe a way to measure the angle of refraction, though he failed to notice the empirical relationship between it and the angle of incidence.



لانه ان ماسته عليها سطح مستو غيره فلان هذا السطح يقطع سطح ب ن ص
على نقطة ب فلا بد من ان يقطع احد خطي ب ن ب ص فليكن ذلك
الخط ب ص والفصل المشترك بين هذا السطح وبين سطح قطع ق د
خط ب ش فلان هذا السطح يماس بسيط ب على نقطة ب فخط
ب ش يماس قطع ق ب د على نقطة ب وكذلك خط ب ص وهذا محال
فلا يماس بسيط ب على نقطة ب سطح مستو غير سطح ب ن ص ○

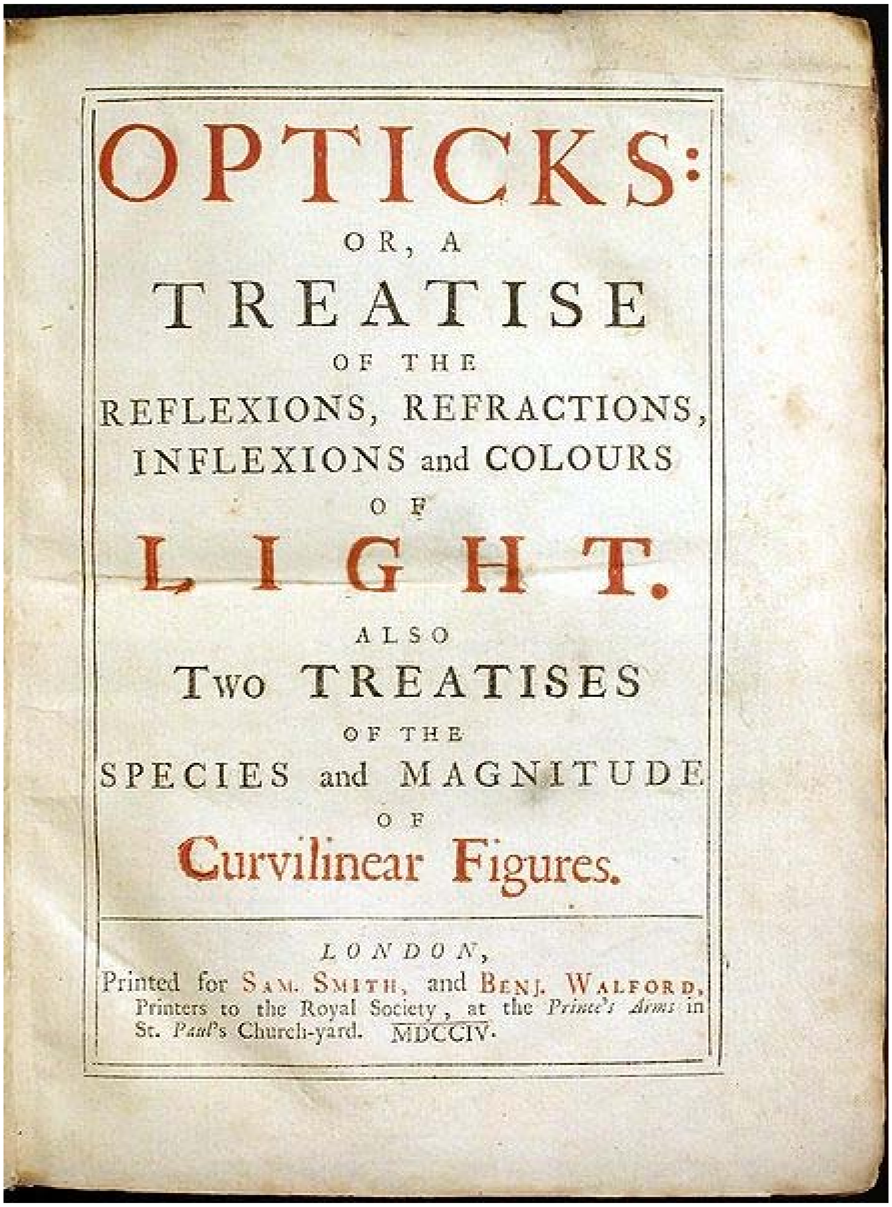
Reproduction of a page of Ibn Sahl's manuscript showing his knowledge of the law of refraction, now known as Snell's law.

During the Middle Ages, Greek ideas about optics were resurrected and extended by writers in the Muslim world. One of the earliest of these was Al-Kindi (c. 801–73). In 984, the Persian mathematician Ibn Sahl wrote the treatise "On burning mirrors and lenses", correctly describing a law of refraction equivalent to Snell's law. He used this law to compute optimum shapes for lenses and curved mirrors. In the early 11th century, Alhazen (Ibn al-Haytham) wrote his *Book of Optics*, which documented the then-current understanding of vision.

In the 13th century, Roger Bacon used parts of glass spheres as magnifying glasses, and discovered that light reflects from objects rather than being released from them. In Italy, around 1284, Salvino D'Armate invented the first wearable eyeglasses.

The earliest known telescopes were refracting telescopes, a type which relies entirely on lenses for magnification. The first rudimentary telescopes were developed independently in the 1570s and 1580s by Leonard Digges, and Giambattista della Porta. Their development in the Netherlands in 1608 was by three individuals: Hans Lippershey and Zacharias Janssen, who were spectacle makers in Middelburg, and Jacob Metius of Alkmaar. In Italy, Galileo greatly improved upon these designs the following year. In 1668, Isaac Newton constructed the first practical reflecting telescope, which bears his name, the Newtonian reflector.

The first microscope was made around 1595, also in Middelburg. Three different eyeglass makers have been given credit for the invention: Lippershey, Janssen, and his father, Hans. The coining of the name "microscope" has been credited to Giovanni Faber, who gave that name to Galileo's compound microscope in 1625.



OPTICKS:
OR, A
TREATISE
OF THE
REFLEXIONS, REFRACTIONS,
INFLEXIONS and COLOURS
OF
LIGHT.
ALSO
Two TREATISES
OF THE
SPECIES and MAGNITUDE
OF
Curvilinear Figures.

LONDON,
Printed for SAM. SMITH, and BENJ. WALFORD,
Printers to the Royal Society, at the *Prince's Arms* in
St. *Paul's* Church-yard. MDCCIV.

Cover of the first edition of Newton's *Opticks*

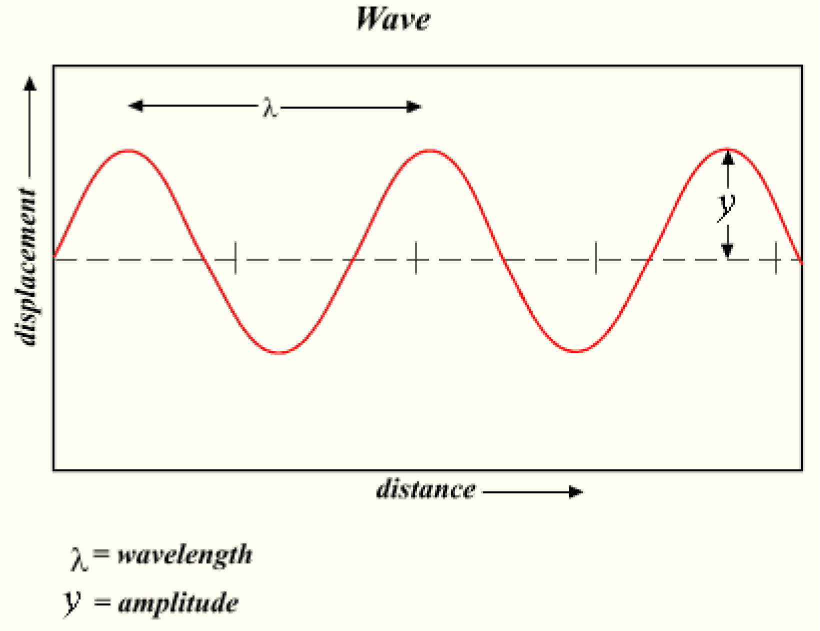
Optical theory progressed in the mid-17th century with treatises written by philosopher René Descartes, which explained a variety of optical phenomena including reflection and refraction by assuming that light was emitted by objects which produced it. This differed substantively from the ancient Greek emission theory. In the late 1660s and early 1670s, Newton expanded Descartes' ideas into a corpuscle theory of light, famously showing that white light, instead of being a unique color, was really a composite of different colors that can be separated into a spectrum with a prism. In 1690, Christian Huygens proposed a wave theory for light based on suggestions that had been made by Robert Hooke in 1664. Hooke himself publicly criticized Newton's theories of light and the feud between the two lasted until Hooke's death. In 1704, Newton published *Opticks* and, at the time, partly because of his success in other areas of physics, he was generally considered to be the victor in the debate over the nature of light.

Newtonian optics was generally accepted until the early 19th century when Thomas Young and Augustin-Jean Fresnel conducted experiments on the interference of light that firmly established light's wave nature. Young's famous double slit experiment showed that light followed the law of superposition, which is a wave-like property not predicted by Newton's corpuscle theory. This work led to a theory of diffraction for light and opened an entire area of study in physical optics. Wave optics was successfully unified with electromagnetic theory by James Clerk Maxwell in the 1860s.

The next development in optical theory came in 1899 when Max Planck correctly modeled blackbody radiation by assuming that the exchange of energy between light and matter only occurred in discrete amounts he called *quanta*. In 1905, Albert Einstein published the theory of the photoelectric effect that firmly established the quantization of light itself. In 1913, Niels Bohr showed that atoms could only emit discrete amounts of energy, thus explaining the discrete lines seen in emission and absorption spectra. The understanding of the interaction between light and matter, which followed from these developments, not only formed the basis of quantum optics but also was crucial for the development of quantum mechanics as a whole. The ultimate culmination was the theory of quantum electrodynamics, which explains all optics and electromagnetic processes in general as being the result of the exchange of real and virtual photons.

Quantum optics gained practical importance with the invention of the maser in 1953 and the laser in 1960. Following the work of Paul Dirac in quantum field theory, George Sudarshan, Roy J. Glauber, and Leonard Mandel applied quantum theory to the electromagnetic field in the 1950s and 1960s to gain a more detailed understanding of photodetection and the statistics of light.

Classical optics



Light propagates through space as a wave with amplitude, wavelength, frequency, and speed that depend on how it was emitted and on the medium through which it travels.

In pre-quantum-mechanical optics, light is an electromagnetic wave composed of oscillating electric and magnetic fields. These fields continually generate each other, as the wave propagates through space and oscillates in time.

The frequency of a light wave is determined by the period of the oscillations. The frequency does not normally change as the wave travels through different materials ("media"), but the speed of the wave depends on the medium. The speed, frequency, and wavelength of a wave are related by the formula

$$v = \lambda f,$$

where v is the speed, λ is the wavelength and f is the frequency. Because the frequency is fixed, a change in the wave's speed produces a change in its wavelength.

The speed of light in a medium is typically characterized by the index of refraction, n, which is the ratio of the speed of light in vacuum, c, to the speed in the medium:

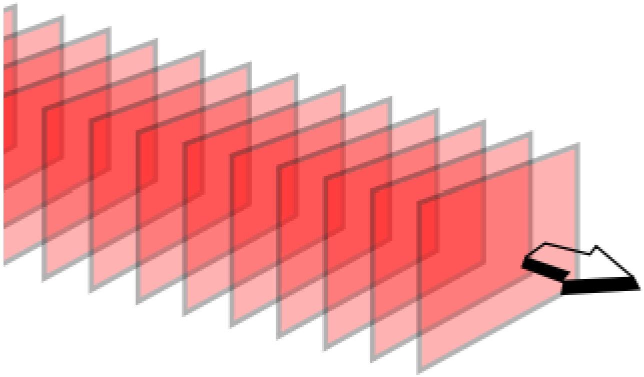
$$n = c / v.$$

The speed of light in vacuum is a constant, which is exactly 299,792,458 metres per second. Thus, a light ray with a wavelength of λ in a vacuum will have a wavelength of λ / n in a material with index of refraction n.

The amplitude of the light wave is related to the intensity of the light, which is related to the energy stored in the wave's electric and magnetic fields.

Traditional optics is divided into two main branches: geometrical optics and physical optics.

Geometrical optics



As a light wave travels through space, it oscillates in amplitude. In this image, each maximum amplitude crest is marked with a plane to illustrate the wavefront. The ray is the arrow perpendicular to these parallel surfaces.

Geometrical optics, or *ray optics*, describes light propagation in terms of "rays". The "ray" in geometric optics is an abstraction, or "instrument", that can be used to predict the path of light. A light ray is a ray that is perpendicular to the light's wavefronts (and therefore collinear with the wave vector). Light rays bend at the interface between two dissimilar media and may be curved in a medium in which the refractive index changes. Geometrical optics provides rules for propagating these rays through an optical system, which indicates how the actual wavefront will propagate. This is a significant simplification of optics that fails to account for optical effects such as diffraction and polarization. It is a good approximation, however, when the wavelength is very small compared with the size of structures with which the light interacts. Geometric optics can be used to describe the geometrical aspects of imaging, including optical aberrations.

A slightly more rigorous definition of a light ray follows from Fermat's principle which states that *the path taken between two points by a ray of light is the path that can be traversed in the least time.*

Approximations

Geometrical optics is often simplified by making the paraxial approximation, or "small angle approximation." The mathematical behavior then becomes linear, allowing optical components and systems to be described by simple matrices. This leads to the techniques of Gaussian optics and *paraxial ray tracing*, which are used to find basic properties of optical systems, such as approximate image and object positions and magnifications.

Reflections

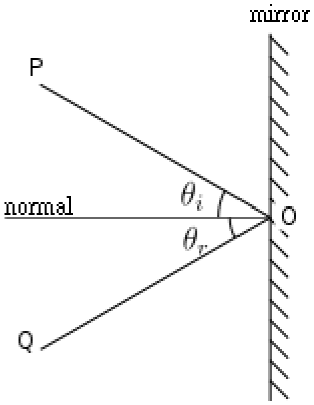


Diagram of specular reflection

Reflections can be divided into two types: specular reflection and diffuse reflection. Specular reflection describes the gloss of surfaces such as mirrors, which reflect light in a simple, predictable way. This allows for production of reflected images that can be associated with an actual (real) or extrapolated (virtual) location in space. Diffuse reflection describes opaque, non limpid materials, such as paper or rock. The reflections from these surfaces can only be described statistically, with the exact distribution of the reflected light depending on the microscopic structure of the material. Many diffuse reflectors are described or can be approximated by Lambert's cosine law, which describes surfaces that have equal luminance when viewed from any angle. Glossy surfaces can give both specular and diffuse reflection.

In specular reflection, the direction of the reflected ray is determined by the angle the incident ray makes with the surface normal, a line perpendicular to the surface at the point where the ray hits. The incident and reflected rays and the normal lie in a single plane, and the angle between the reflected ray and the surface normal is the same as that between the incident ray and the normal. This is known as the Law of Reflection.

For flat mirrors, the law of reflection implies that images of objects are upright and the same distance behind the mirror as the objects are in front of the mirror. The image size is the same as the object size. (The magnification of a flat mirror is unity.) The law also implies that mirror images are parity inverted, which we perceive as a left-right inversion. Images formed from reflection in two (or any even number of) mirrors are not parity inverted. Corner reflectors retroreflect light, producing reflected rays that travel back in the direction from which the incident rays came.

Mirrors with curved surfaces can be modeled by ray-tracing and using the law of reflection at each point on the surface. For mirrors with parabolic surfaces, parallel rays incident on the mirror produce reflected rays that converge at a common focus. Other curved surfaces may also focus light, but with aberrations due to the diverging shape causing the focus to be smeared out in space. In particular, spherical mirrors exhibit spherical aberration. Curved mirrors can form images with magnification greater than or less than one, and the magnification can be negative, indicating that the image is inverted. An upright image formed by reflection in a mirror is always virtual, while an inverted image is real and can be projected onto a screen.

Refractions

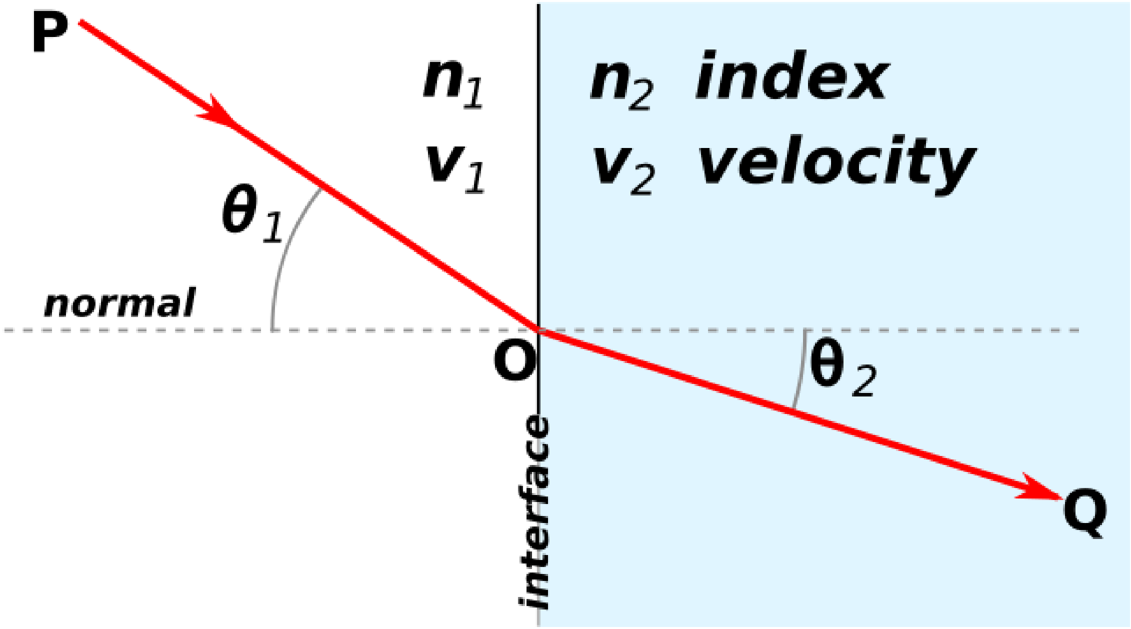


Illustration of Snell's Law for the case $n_1 < n_2$, such as air/water interface

Refraction occurs when light travels through an area of space that has a changing index of refraction; this principle allows for lenses and the focusing of light. The simplest case of refraction occurs when there is an interface between a uniform medium with index of refraction n_1 and another medium with index of refraction n_2. In such situations, Snell's Law describes the resulting deflection of the light ray:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

where θ_1 and θ_2 are the angles between the normal (to the interface) and the incident and refracted waves, respectively. This phenomenon is also associated with a changing speed of light as seen from the definition of index of refraction provided above which implies:

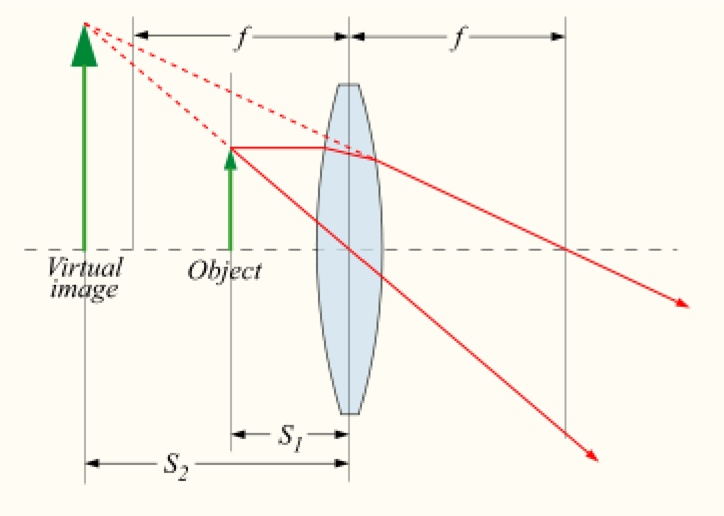
$$v_1 \sin\theta_2 \ = v_2 \sin\theta_1$$

where v_1 and v_2 are the wave velocities through the respective media.

Various consequences of Snell's Law include the fact that for light rays traveling from a material with a high index of refraction to a material with a low index of refraction, it is possible for the interaction with the interface to result in zero transmission. This phenomenon is called total internal reflection and allows for fiber optics technology. As light signals travel down a fiber optic cable, it undergoes total internal reflection allowing for essentially no light lost over the length of the cable. It is also possible to produce polarized light rays using a combination of reflection and refraction: When a refracted ray and the reflected ray form a right angle, the reflected ray has the property of "plane polarization". The angle of incidence required for such a scenario is known as Brewster's angle.

Snell's Law can be used to predict the deflection of light rays as they pass through "linear media" as long as the indexes of refraction and the geometry of the media are known. For example, the propagation of light through a prism results in the light ray being deflected depending on the shape and orientation of the prism. Additionally, since different frequencies of light have slightly different indexes of refraction in most materials, refraction can be used to produce dispersion spectra that appear as rainbows. The discovery of this phenomenon when passing light through a prism is famously attributed to Isaac Newton.

Some media have an index of refraction which varies gradually with position and, thus, light rays curve through the medium rather than travel in straight lines. This effect is what is responsible for mirages seen on hot days where the changing index of refraction of the air causes the light rays to bend creating the appearance of specular reflections in the distance (as if on the surface of a pool of water). Material that has a varying index of refraction is called a gradient-index (GRIN) material and has many useful properties used in modern optical scanning technologies including photocopiers and scanners. The phenomenon is studied in the field of gradient-index optics.

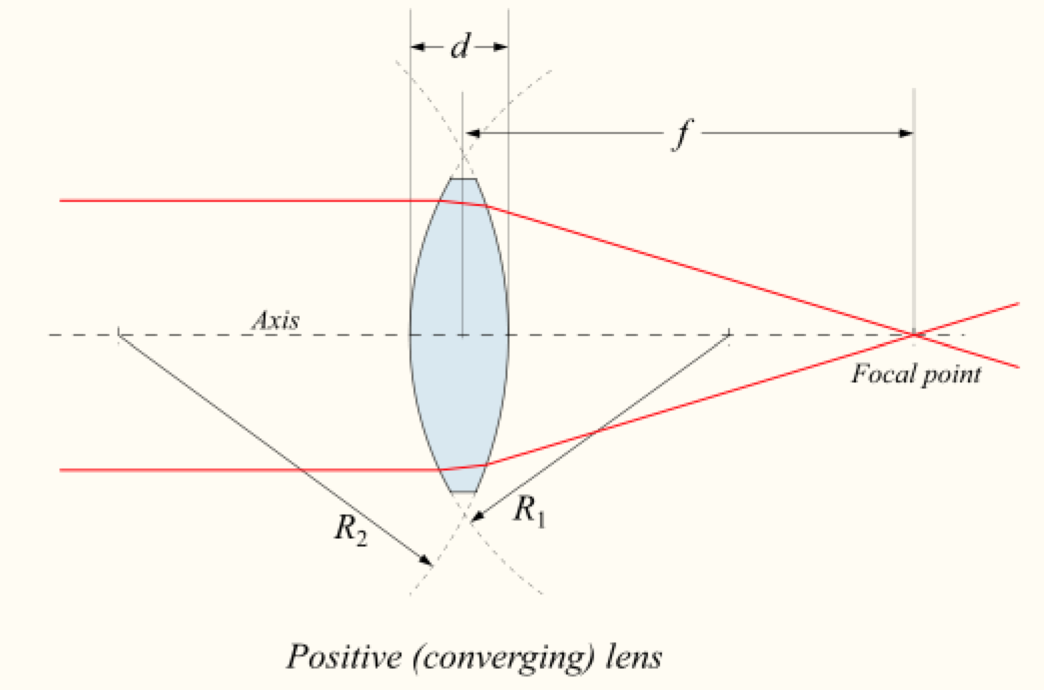


A ray tracing diagram for a converging lens.

A device which produces converging or diverging light rays due to refraction is known as a lens. Thin lenses produce focal points on either side that can be modeled using the lensmaker's equation. In general, two types of lenses exist: convex lenses, which cause parallel light rays to converge, and concave lenses, which cause parallel light rays to diverge. The detailed prediction of how images are produced by these lenses can be made using ray-tracing similar to curved mirrors. Similarly to curved mirrors, thin lenses follow a simple equation that determines the location of the images given a particular focal length (*f*) and object distance (S_1):

$$\frac{1}{S_1}+\frac{1}{S_2}=\frac{1}{f}$$

where S_2 is the distance associated with the image and is considered by convention to be negative if on the same side of the lens as the object and positive if on the opposite side of the lens. The focal length f is considered negative for concave lenses.



Positive (converging) lens

Incoming parallel rays are focused by a convex lens into an inverted real image one focal length from the lens, on the far side of the lens. Rays from an object at finite distance are focused further from the lens than the focal distance; the closer the object is to the lens, the further the image is from the lens. With concave lenses, incoming parallel rays diverge after going through the lens, in such a way that they seem to have originated at an upright virtual image one focal length from the lens, on the same side of the lens that the parallel rays are approaching on. Rays from an object at finite distance are associated with a virtual image that is closer to the lens than the focal length, and on the same side of the lens as the object. The closer the object is to the lens, the closer the virtual image is to the lens.

Likewise, the magnification of a lens is given by

$$M = -\frac{S_2}{S_1} = \frac{f}{f - S_1}$$

where the negative sign is given, by convention, to indicate an upright object for positive values and an inverted object for negative values. Similar to mirrors, upright images produced by single lenses are virtual while inverted images are real.

Lenses suffer from aberrations that distort images and focal points. These are due to both to geometrical imperfections and due to the changing index of refraction for different wavelengths of light (chromatic aberration).

Physical optics

Physical optics or wave optics builds on Huygens's principle, which states that every point on an advancing wavefront is the center of a new disturbance. When combined with the superposition principle, this explains how optical phenomena are manifested when there are multiple sources or obstructions that are spaced at distances similar to the wavelength of the light.

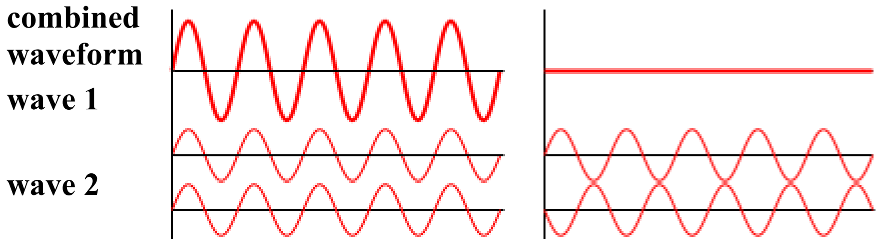
Complex models based on physical optics can account for the propagation of any wavefront through an optical system, including predicting the wavelength, amplitude, and phase of the wave. Additionally, all of the results from geometrical optics can be recovered using the techniques of Fourier optics which apply many of the same mathematical and analytical techniques used in acoustic engineering and signal processing.

Using numerical modeling on a computer, optical scientists can simulate the propagation of light and account for most diffraction, interference, and polarization effects. Such simulations typically still rely on approximations, however, so this is not a full electromagnetic wave theory model of the propagation of light. Such a full model is computationally demanding and is normally only used to solve small-scale problems that require extraordinary accuracy.

Gaussian beam propagation is a simple paraxial physical optics model for the propagation of coherent radiation such as laser beams. This technique partially accounts for diffraction, allowing accurate calculations of the rate at which a laser beam expands with distance, and the minimum size to which the beam can be focused. Gaussian beam propagation thus bridges the gap between geometric and physical optics.

Superposition and interference

In the absence of nonlinear effects, the superposition principle can be used to predict the shape of interacting waveforms through the simple addition of the disturbances. This interaction of waves to produce a resulting pattern is generally termed "interference" and can result in a variety of outcomes. If two waves of the same wavelength and frequency are *in phase*, both the wave crests and wave troughs align. This results in constructive interference and an increase in the amplitude of the wave, which for light is associated with a brightening of the waveform in that location. Alternatively, if the two waves of the same wavelength and frequency are out of phase, then the wave crests will align with wave troughs and vice-versa. This results in destructive interference and a decrease in the amplitude of the wave, which for light is associated with a dimming of the waveform at that location.



Two waves in phase

Two waves 180° out of phase

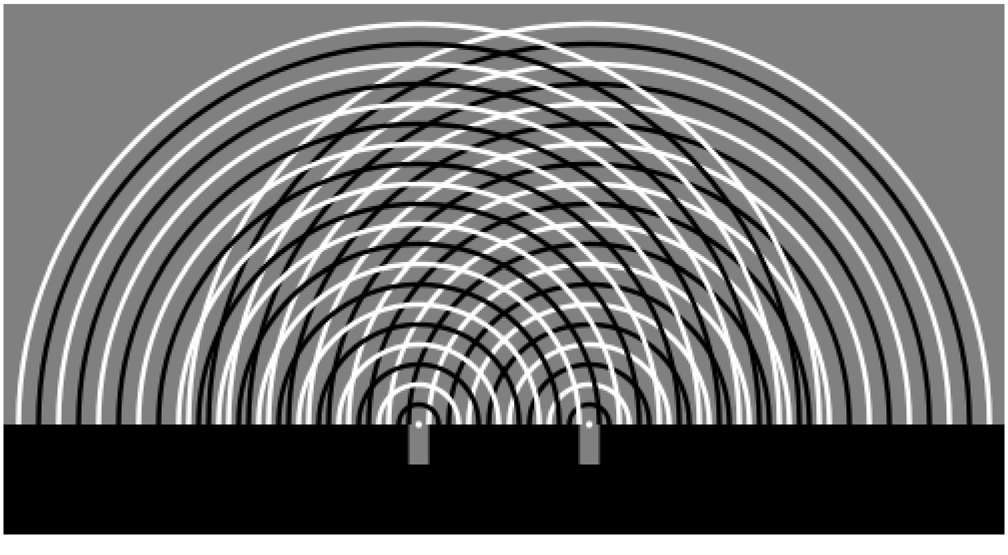
When oil or fuel is spilled, colorful patterns are formed by thin-film interference.

Since Huygens's principle states that every point of a wavefront is associated with the production of a new disturbance, it is possible for a wavefront to interfere with itself constructively or destructively at different locations producing bright and dark fringes in regular and predictable patterns. Interferometry is the science of measuring these patterns, usually as a means of making precise determinations of distances or angular resolutions. The Michelson interferometer was a famous instrument which used interference effects to accurately measure the speed of light.

The appearance of thin films and coatings is directly affected by interference effects. Antireflective coatings use destructive interference to reduce the reflectivity of the surfaces they coat, and can be used to minimize glare and unwanted reflections. The simplest case is a single layer with thickness one-fourth the wavelength of incident light. The reflected wave from the top of the film and the reflected wave from the film/material interface are then exactly 180° out of phase, causing destructive interference. The waves are only exactly out of phase for one wavelength, which would typically be chosen to be near the center of the visible spectrum, around 550 nm. More complex designs using multiple layers can achieve low reflectivity over a broad band, or extremely low reflectivity at a single wavelength.

Constructive interference in thin films can create strong reflection of light in a range of wavelengths, which can be narrow or broad depending on the design of the coating. These films are used to make dielectric mirrors, interference filters, heat reflectors, and filters for color separation in color television cameras. This interference effect is also what causes the colorful rainbow patterns seen in oil slicks.

Diffraction and optical resolution



Diffraction on two slits separated by distance *d*. The bright fringes occur along lines where black lines intersect with black lines and white lines intersect with white lines. These fringes are separated by angle θ and are numbered as order *n*.

Diffraction is the process by which light interference is most commonly observed. The effect was first described in 1665 by Francesco Maria Grimaldi, who also coined the term from the Latin *diffringere*, 'to break into pieces'. Later that century, Robert Hooke and Isaac Newton also described phenomena now known to be diffraction in Newton's rings while James Gregory recorded his observations of diffraction patterns from bird feathers.

The first physical optics model of diffraction that relied on Huygens' Principle was developed in 1803 by Thomas Young in his accounts of the interference patterns of two closely spaced slits. Young showed that his results could only be explained if the two slits acted as two unique sources of waves rather than corpuscles. In 1815 and 1818, Augustin-Jean Fresnel firmly established the mathematics of how wave interference can account for diffraction.

The simplest physical models of diffraction use equations that describe the angular separation of light and dark fringes due to light of a particular wavelength (λ). In general, the equation takes the form

$$m\lambda = d\sin\theta$$

where d is the separation between two wavefront sources (in the case of Young's experiments, it was two slits), θ is the angular separation between the central fringe and the mth order fringe, where the central maximum is $m = 0$.

This equation is modified slightly to take into account a variety of situations such as diffraction through a single gap, diffraction through multiple slits, or diffraction through a diffraction grating that contains a large number of slits at equal spacing. More complicated models of diffraction require working with the mathematics of Fresnel or Fraunhofer diffraction.

X-ray diffraction makes use of the fact that atoms in a crystal have regular spacing at distances that are on the order of one angstrom. To see diffraction patterns, x-rays with similar wavelengths to that spacing are passed through the crystal. Since crystals are three-dimensional objects rather than two-dimensional gratings, the associated diffraction pattern varies in two directions according to Bragg reflection, with the associated bright spots occurring in unique patterns and d being twice the spacing between atoms.

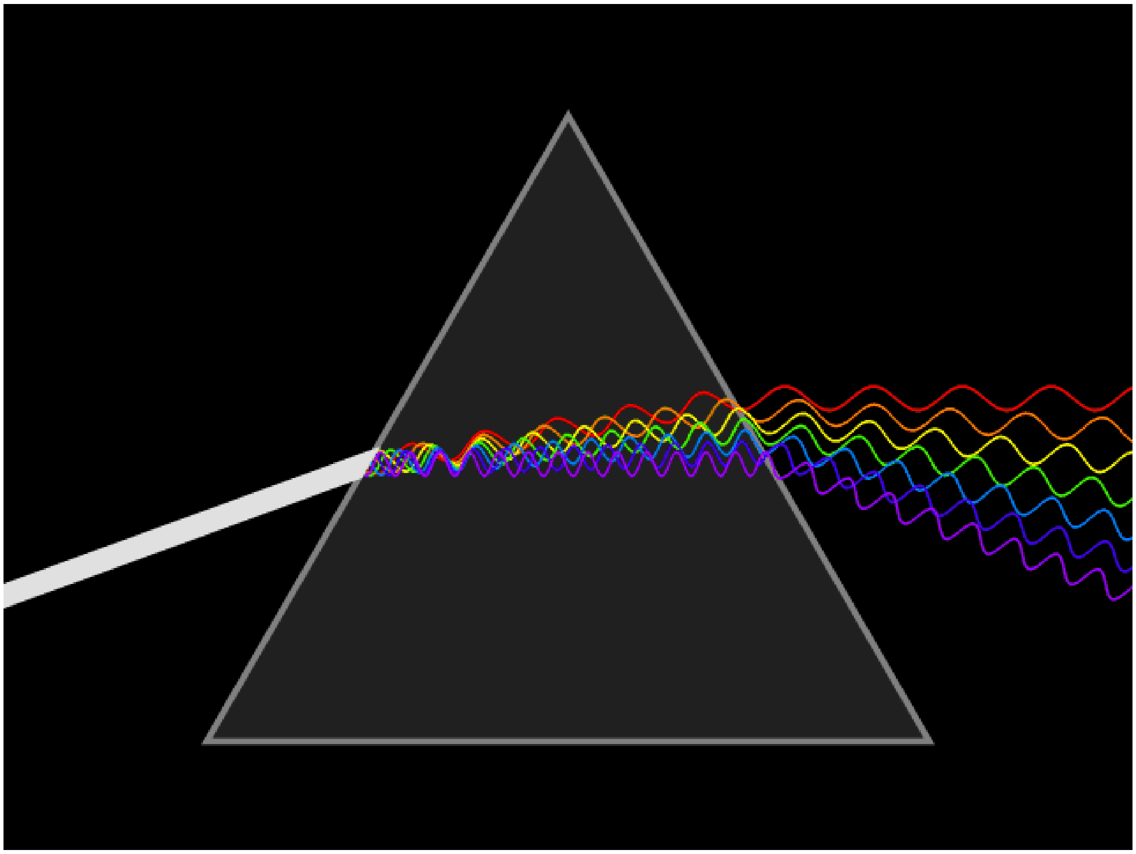
Diffraction effects limit the ability for an optical detector to optically resolve separate light sources. In general, light that is passing through an aperture will experience diffraction and the best images that can be created (as described in diffraction-limited optics) appear as a central spot with surrounding bright rings, separated by dark nulls; this pattern is known as an Airy pattern, and the central bright lobe as an Airy disk. The size of such a disk is given by

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

where θ is the angular resolution, λ is the wavelength of the light, and D is the diameter of the lens aperture. If the angular separation of the two points is significantly less than the Airy disk angular radius, then the two points cannot be resolved in the image, but if their angular separation is much greater than this, distinct images of the two points are formed and they can therefore be resolved. Rayleigh defined the somewhat arbitrary "Rayleigh criterion" that two points whose angular separation is equal to the Airy disk radius (measured to first null, that is, to the first place where no light is seen) can be considered to be resolved. It can be seen that the greater the diameter of the lens or its aperture, the finer the resolution. Interferometry, with its ability to mimic extremely large baseline apertures, allows for the greatest angular resolution possible.

For astronomical imaging, the atmosphere prevents optimal resolution from being achieved in the visible spectrum due to the atmospheric scattering and dispersion which cause stars to twinkle. Astronomers refer to this effect as the quality of astronomical seeing. Techniques known as adaptive optics have been utilized to eliminate the atmospheric disruption of images and achieve results that approach the diffraction limit.

Dispersion and scattering

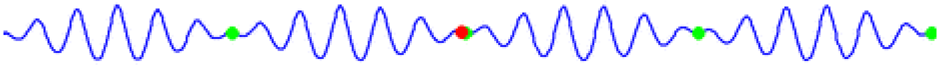


Conceptual animation of light dispersion through a prism. High frequency (blue) light is deflected the most, and low frequency (red) the least.

Refractive processes take place in the physical optics limit, where the wavelength of light is similar to other distances, as a kind of scattering. The simplest type of scattering is Thomson scattering which occurs when electromagnetic waves are deflected by single particles. In the limit of Thompson scattering, in which the wavelike nature of light is evident, light is dispersed independent of the frequency, in contrast to Compton scattering which is frequency-dependent and strictly a quantum mechanical process, involving the nature of light as particles. In a statistical sense, elastic scattering of light by numerous particles much smaller than the wavelength of the light is a process known as Rayleigh scattering while the similar process for scattering by particles that are similar or larger in wavelength is known as Mie scattering with the Tyndall effect being a commonly observed result. A small proportion of light scattering from atoms or molecules may undergo Raman scattering, wherein the frequency changes due to excitation of the atoms and molecules. Brillouin scattering occurs when the frequency of light changes due to local changes with time and movements of a dense material.

Dispersion occurs when different frequencies of light have different phase velocities, due either to material properties (*material dispersion*) or to the geometry of an optical waveguide (*waveguide dispersion*). The most familiar form of dispersion is a decrease in index of refraction with increasing wavelength, which is seen in most transparent materials. This is called "normal dispersion". It occurs in all dielectric materials, in wavelength ranges where the material does not absorb light. In wavelength ranges where a medium has significant absorption, the index of refraction can increase with wavelength. This is called "anomalous dispersion".

The separation of colors by a prism is an example of normal dispersion. At the surfaces of the prism, Snell's law predicts that light incident at an angle θ to the normal will be refracted at an angle arcsin(sin (θ) / *n*) . Thus, blue light, with its higher refractive index, is bent more strongly than red light, resulting in the well-known rainbow pattern.



Dispersion: two sinusoids propagating at different speeds make a moving interference pattern. The red dot moves with the phase velocity, and the green dots propagate with the group velocity. In this case, the phase velocity is twice the group velocity. The red dot overtakes two green dots, when moving from the left to the right of the figure. In effect, the individual waves (which travel with the phase velocity) escape from the wave packet (which travels with the group velocity).

Material dispersion is often characterized by the Abbe number, which gives a simple measure of dispersion based on the index of refraction at three specific wavelengths. Waveguide dispersion is dependent on the propagation constant. Both kinds of dispersion cause changes in the group characteristics of the wave, the features of the wave packet that change with the same frequency as the amplitude of the electromagnetic wave. "Group velocity dispersion" manifests as a spreading-out of the signal "envelope" of the radiation and can be quantified with a group dispersion delay parameter:

$$D = \frac{1}{v_g^2}\frac{dv_g}{d\lambda}$$

where v_g is the group velocity. For a uniform medium, the group velocity is

$$v_g = c\left(n - \lambda\frac{dn}{d\lambda}\right)^{-1}$$

where n is the index of refraction and c is the speed of light in a vacuum. This gives a simpler form for the dispersion delay parameter:

$$D = -\frac{\lambda}{c}\,\frac{d^2n}{d\lambda^2}.$$

If D is less than zero, the medium is said to have *positive dispersion* or normal dispersion. If D is greater than zero, the medium has *negative dispersion*. If a light pulse is propagated through a normally dispersive medium, the result is the higher frequency components slow down more than the lower frequency components. The pulse therefore becomes *positively chirped*, or *up-chirped*, increasing in frequency with time. This causes the spectrum coming out of a prism to appear with red light the least refracted and blue/violet light the most refracted. Conversely, if a pulse travels through an anomalously (negatively) dispersive medium, high frequency components travel faster than the lower ones, and the pulse becomes *negatively chirped*, or *down-chirped*, decreasing in frequency with time.

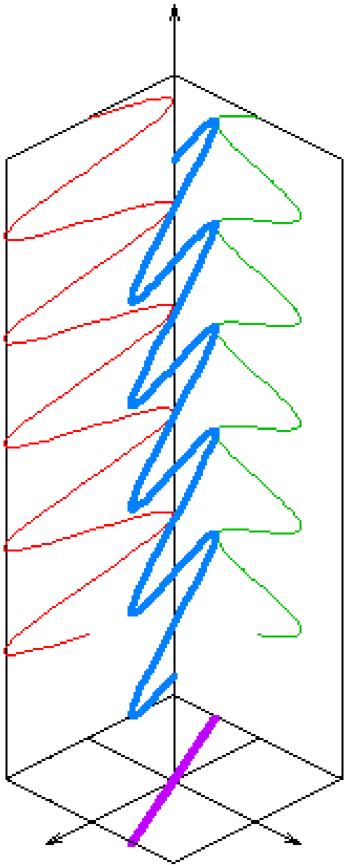
The result of group velocity dispersion, whether negative or positive, is ultimately temporal spreading of the pulse. This makes dispersion management extremely important in optical communications systems based on optical fibers, since if dispersion is too high, a group of pulses representing information will each spread in time and merge together, making it impossible to extract the signal.

Polarization

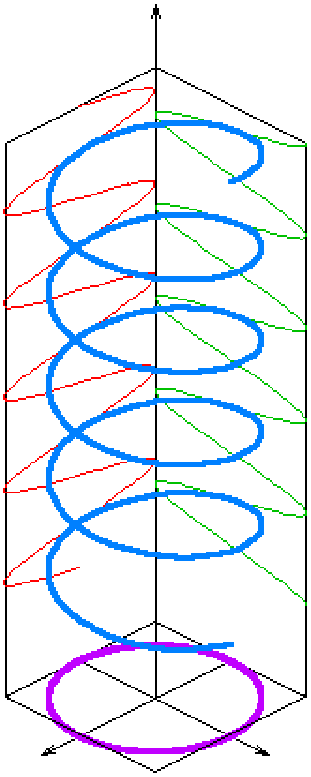
Polarization is a general property of waves that describes the orientation of their oscillations. For transverse waves such as many electromagnetic waves, it describes the orientation of the oscillations in the plane perpendicular to the wave's direction of travel. The oscillations may be oriented in a single direction (linear polarization), or the oscillation direction may rotate as the wave travels (circular or elliptical polarization). Circularly polarized waves can rotate rightward or leftward in the direction of travel, and which of those two rotations is present in a wave is called the wave's chirality.

The typical way to consider polarization is to keep track of the orientation of the electric field vector as the electromagnetic wave propagates. The electric field vector of a plane wave may be arbitrarily divided into two perpendicular components labeled x and y (with

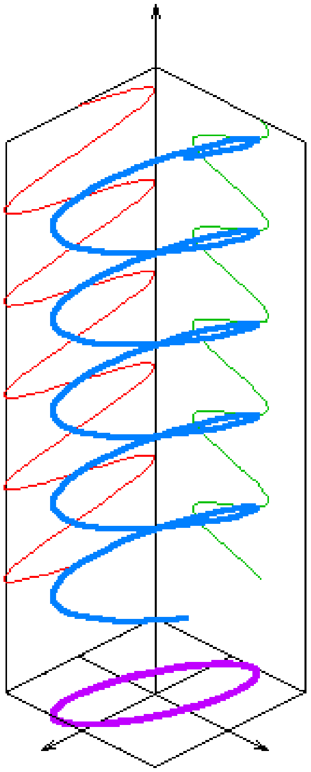
z indicating the direction of travel). The shape traced out in the x-y plane by the electric field vector is a Lissajous figure that describes the *polarization state*. The following figures show some examples of the evolution of the electric field vector (blue), with time (the vertical axes), at a particular point in space, along with its *x* and *y* components (red/left and green/right), and the path traced by the vector in the plane (purple): The same evolution would occur when looking at the electric field at a particular time while evolving the point in space, along the direction opposite to propagation.



Linear



Circular



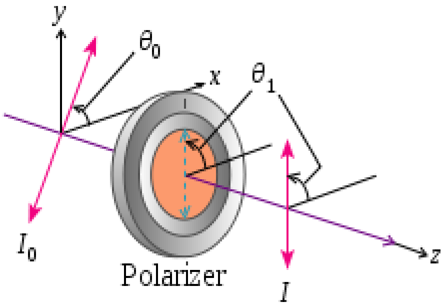
Elliptical polarization

In the leftmost figure above, the x and y components of the light wave are in phase. In this case, the ratio of their strengths is constant, so the direction of the electric vector (the vector sum of these two components) is constant. Since the tip of the vector traces out a single line in the plane, this special case is called linear polarization. The direction of this line depends on the relative amplitudes of the two components.

In the middle figure, the two orthogonal components have the same amplitudes and are 90° out of phase. In this case, one component is zero when the other component is at maximum or minimum amplitude. There are two possible phase relationships that satisfy this requirement: the *x* component can be 90° ahead of the *y* component or it can be 90° behind the *y* component. In this special case, the electric vector traces out a circle in the plane, so this polarization is called circular polarization. The rotation direction in the circle depends on which of the two phase relationships exists and corresponds to *right-hand circular polarization* and *left-hand circular polarization*.

In all other cases, where the two components either do not have the same amplitudes and/or their phase difference is neither zero nor a multiple of 90°, the polarization is called elliptical polarization because the electric vector traces out an ellipse in the plane (the *polarization ellipse*). This is shown in the above figure on the right. Detailed mathematics of polarization is done using Jones calculus and is characterized by the Stokes parameters.

Media that have different indexes of refraction for different polarization modes are called *birefringent*. Well known manifestations of this effect appear in optical wave plates/retarders (linear modes) and in Faraday rotation/optical rotation (circular modes). If the path length in the birefringent medium is sufficient, plane waves will exit the material with a significantly different propagation direction, due to refraction. For example, this is the case with macroscopic crystals of calcite, which present the viewer with two offset, orthogonally polarized images of whatever is viewed through them. It was this effect that provided the first discovery of polarization, by Erasmus Bartholinus in 1669. In addition, the phase shift, and thus the change in polarization state, is usually frequency dependent, which, in combination with dichroism, often gives rise to bright colors and rainbow-like effects. In mineralogy, such properties, known as pleochroism, are frequently exploited for the purpose of identifying minerals using polarization microscopes. Additionally, many plastics that are not normally birefringent will become so when subject to mechanical stress, a phenomenon which is the basis of photoelasticity. Non-birefringent methods, to rotate the linear polarization of light beams, include the use of prismatic polarization rotators which utilize total internal reflection in a prism set designed for efficient colinear transmission.



A polarizer changing the orientation of linearly polarized light.
In this picture, $\theta_1 - \theta_0 = \theta_i$.

Media that reduce the amplitude of certain polarization modes are called *dichroic*. with devices that block nearly all of the radiation in one mode known as *polarizing filters* or simply "polarizers". Malus' law, which is named after Etienne-Louis Malus, says that when a perfect polarizer is placed in a linear polarized beam of light, the intensity, *I*, of the light that passes through is given by

$$I = I_0 \cos^2 \theta_i \quad ,$$

where

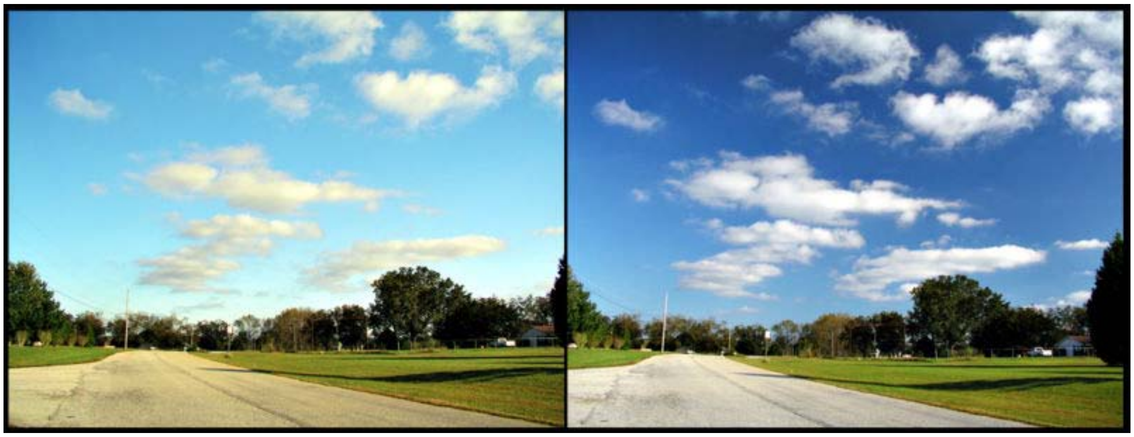
> I_0 is the initial intensity,
> and θ_i is the angle between the light's initial polarization direction and the axis of the polarizer.

A beam of unpolarized light can be thought of as containing a uniform mixture of linear polarizations at all possible angles. Since the average value of $\cos^2\theta$ is 1/2, the transmission coefficient becomes

$$\frac{I}{I_0} = \frac{1}{2}$$

In practice, some light is lost in the polarizer and the actual transmission of unpolarized light will be somewhat lower than this, around 38% for Polaroid-type polarizers but considerably higher (>49.9%) for some birefringent prism types.

In addition to birefringence and dichroism in extended media, polarization effects can also occur at the (reflective) interface between two materials of different refractive index. These effects are treated by the Fresnel equations. Part of the wave is transmitted and part is reflected, with the ratio depending on angle of incidence and the angle of refraction. In this way, physical optics recovers Brewster's angle.



The effects of a polarizing filter on the sky in a photograph. Left picture is taken without polarizer. For the right picture, filter was adjusted to eliminate certain polarizations of the scattered blue light from the sky.

Most sources of electromagnetic radiation contain a large number of atoms or molecules that emit light. The orientation of the electric fields produced by these emitters may not be correlated, in which case the light is said to be *unpolarized*. If there is partial correlation between the emitters, the light is *partially polarized*. If the polarization is consistent across the spectrum of the source, partially polarized light can be described as a superposition of a completely unpolarized component, and a completely polarized one. One may then describe the light in terms of the degree of polarization, and the parameters of the polarization ellipse.

Light reflected by shiny transparent materials is partly or fully polarized, except when the light is normal (perpendicular) to the surface. It was this effect that allowed the mathematician Etienne Louis Malus to make the measurements that allowed for his development of the first mathematical models for polarized light. Polarization occurs when light is scattered in the atmosphere. The scattered light produces the brightness and color in clear skies. This partial polarization of scattered light can be taken advantage of using polarizing filters to darken the sky in photographs. Optical polarization is principally of importance in chemistry due to circular dichroism and optical rotation ("*circular birefringence*") exhibited by optically active (chiral) molecules.

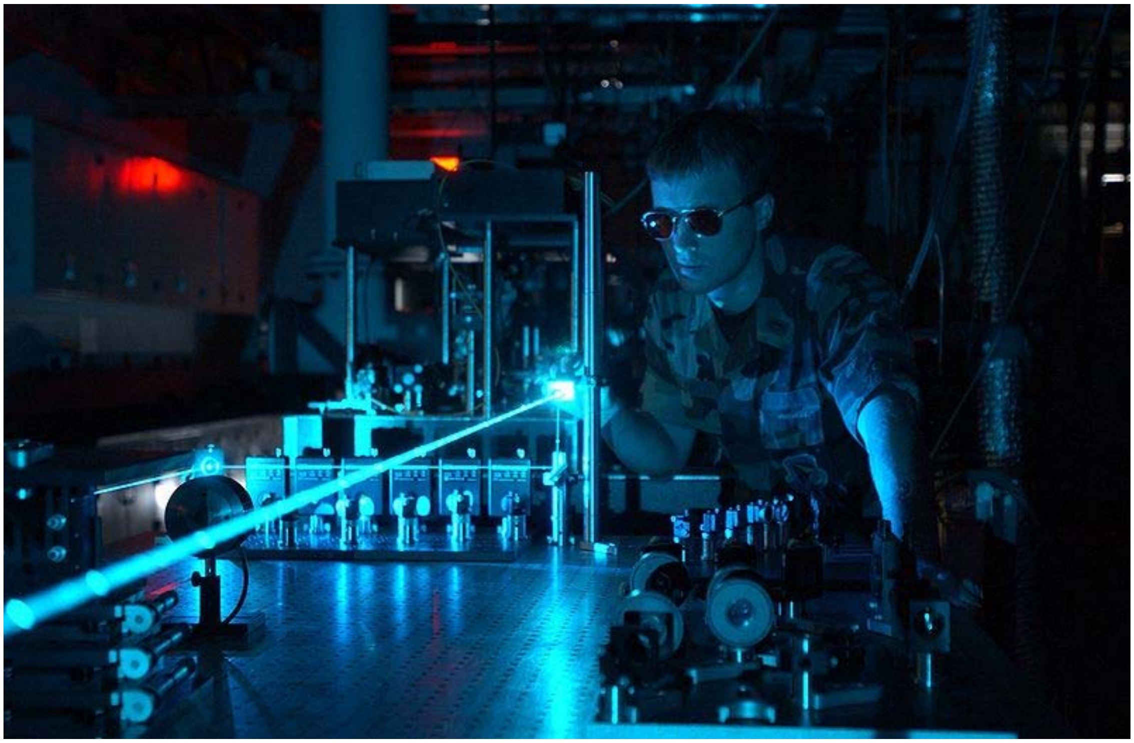
Modern optics

Modern optics encompasses the areas of optical science and engineering that became popular in the 20th century. These areas of optical science typically relate to the electromagnetic or quantum properties of light but do include other topics. A major subfield of modern optics, quantum optics, deals with specifically quantum mechanical properties of light. Quantum optics is not just theoretical; some modern devices, such as lasers, have principles of operation that depend on quantum mechanics. Light detectors, such as photomultipliers and channeltrons, respond to individual photons. Electronic image sensors, such as CCDs, exhibit shot noise corresponding to the statistics of individual photon events. Light-emitting diodes and photovoltaic cells, too, cannot be understood without quantum mechanics. In the study of these devices, quantum optics often overlaps with quantum electronics.

Specialty areas of optics research include the study of how light interacts with specific materials as in crystal optics and metamaterials. Other research focuses on the phenomenology of electromagnetic waves as in singular optics, non-imaging optics, non-linear optics, statistical optics, and radiometry. Additionally, computer engineers have taken an interest in integrated optics, machine vision, and photonic computing as possible components of the "next generation" of computers.

Today, the pure science of optics is called optical science or optical physics to distinguish it from applied optical sciences, which are referred to as optical engineering. Prominent subfields of optical engineering include illumination engineering, photonics, and optoelectronics with practical applications like lens design, fabrication and testing of optical components, and image processing. Some of these fields overlap, with nebulous boundaries between the subjects terms that mean slightly different things in different parts of the world and in different areas of industry. A professional community of researchers in nonlinear optics has developed in the last several decades due to advances in laser technology.

Lasers



Experiments such as this one with high-power lasers are part of the modern optics research.

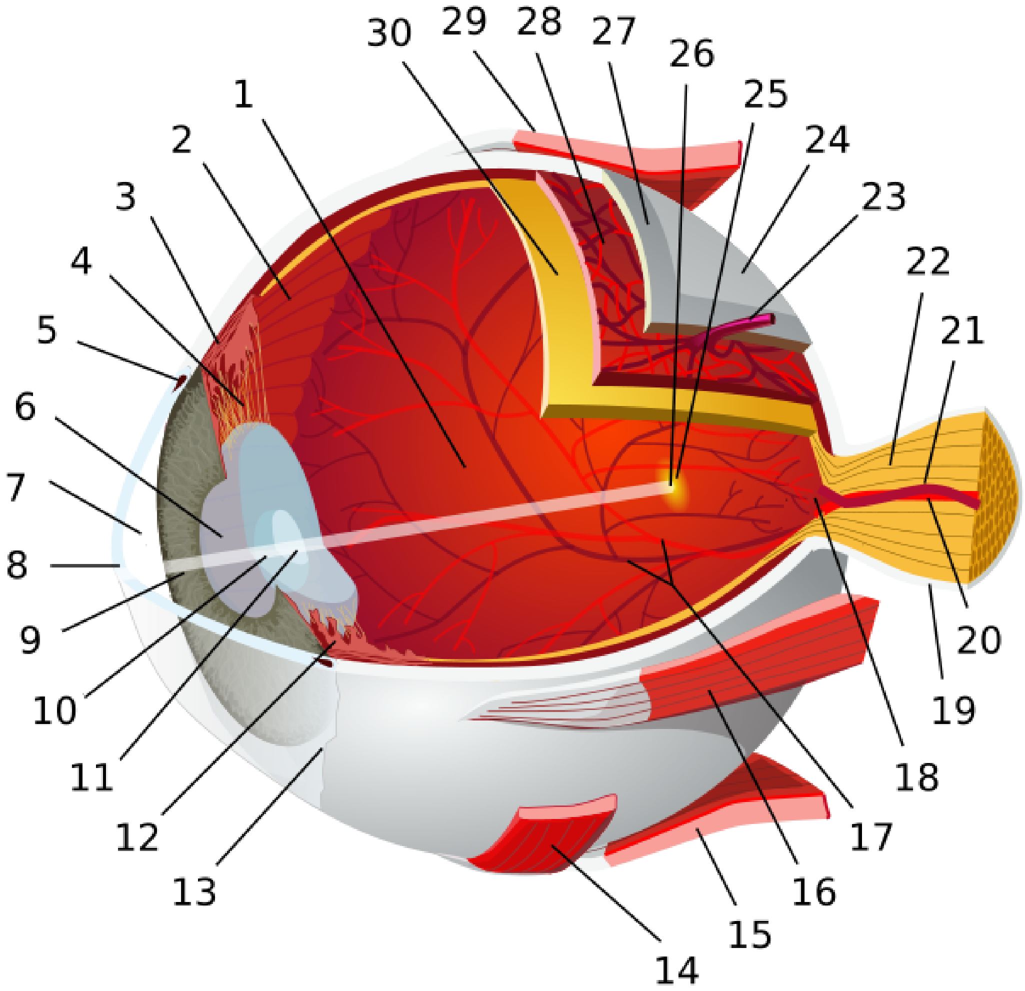
A laser is a device that emits light (electromagnetic radiation) through a process called *stimulated emission*. The term *laser* is an acronym for *Light Amplification by Stimulated Emission of Radiation*. Laser light is usually spatially coherent, which means that the light either is emitted in a narrow, low-divergence beam, or can be converted into one with the help of optical components such as lenses. Because the microwave equivalent of the laser, the *maser*, was developed first, devices that emit microwave and radio frequencies are usually called *masers*.

The first working laser was demonstrated on 16 May 1960 by Theodore Maiman at Hughes Research Laboratories. When first invented, they were called "a solution looking for a problem". Since then, lasers have become a multi-billion dollar industry, finding utility in thousands of highly varied applications. The first application of lasers visible in the daily lives of the general population was the supermarket barcode scanner, introduced in 1974. The laserdisc player, introduced in 1978, was the first successful consumer product to include a laser, but the compact disc player was the first laser-equipped device to become truly common in consumers' homes, beginning in 1982. These optical storage devices use a semiconductor laser less than a millimeter wide to scan the surface of the disc for data retrieval. Fiber-optic communication relies on lasers to transmit large amounts of information at the speed of light. Other common applications of lasers include laser printers and laser pointers. Lasers are used in medicine in areas such as bloodless surgery, laser eye surgery, and laser capture microdissection and in military applications such as missile defense systems, electro-optical countermeasures (EOCM), and LIDAR. Lasers are also used in holograms, bubblegrams, laser light shows, and laser hair removal.

Applications

Optics is part of everyday life. The ubiquity of visual systems in biology indicate the central role optics plays as the science of one of the five senses. Many people benefit from eyeglasses or contact lenses, and optics are integral to the functioning of many consumer goods including cameras. Rainbows and mirages are examples of optical phenomena. Optical communication provides the backbone for both the Internet and modern telephony.

Human eye



Model of a human eye. Features mentioned here are 3. ciliary muscle, 6. pupil, 8. cornea, 10. lens cortex, 22. optic nerve, 26. fovea, 30. retina

The human eye functions by focusing light onto an array of photoreceptor cells called the retina, which covers the back of the eye. The focusing is accomplished by a series of transparent media. Light entering the eye passes first through the cornea, which provides much of the eye's optical power. The light then continues through the fluid just behind the cornea—the anterior chamber, then passes through the pupil. The light then passes through the lens, which focuses the light further and allows adjustment of focus. The light then passes through the main body of fluid in the eye—the vitreous humor, and reaches the retina. The cells in the retina cover the back of the eye, except for where the optic nerve exits; this results in a blind spot.

There are two types of photoreceptor cells, rods and cones, which are sensitive to different aspects of light. Rod cells are sensitive to the intensity of light over a wide frequency range, thus are responsible for black-and-white vision. Rod cells are not present on the fovea, the area of the retina responsible for central vision, and are not as responsive as cone cells to spatial and temporal changes in light. There are, however, twenty times more rod cells than cone cells in the retina because the rod cells are present across a wider area. Because of their wider distribution, rods are responsible for peripheral vision.

In contrast, cone cells are less sensitive to the overall intensity of light, but come in three varieties that are sensitive to different frequency-ranges and thus are used in the perception of color and photopic vision. Cone cells are highly concentrated in the fovea and have a high visual acuity meaning that they are better at spatial resolution than rod cells. Since cone cells are not as sensitive to dim light as rod cells, most night vision is limited to rod cells. Likewise, since cone cells are in the fovea, central vision (including the vision needed to do most reading, fine detail work such as sewing, or careful examination of objects) is done by cone cells.

Ciliary muscles around the lens allow the eye's focus to be adjusted. This process is known as accommodation. The near point and far point define the nearest and farthest distances from the eye at which an object can be brought into sharp focus. For a person with normal vision, the far point is located at infinity. The near point's location depends on how much the muscles can increase the curvature of the lens, and how inflexible the lens has become with age. Optometrists, ophthalmologists, and opticians usually consider an appropriate near point to be closer than normal reading distance—approximately 25 cm.

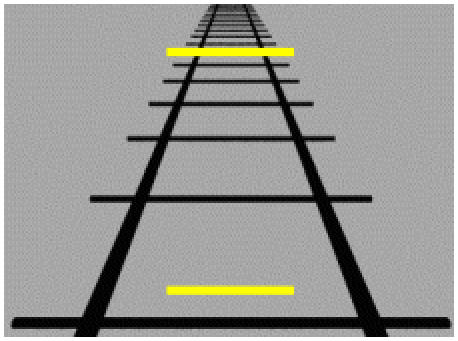
Defects in vision can be explained using optical principles. As people age, the lens becomes less flexible and the near point recedes from the eye, a condition known as presbyopia. Similarly, people suffering from hyperopia cannot decrease the focal length of their lens enough to allow for nearby objects to be imaged on their retina. Conversely, people who cannot increase the focal length of their lens enough to allow for distant objects to be imaged on the retina suffer from myopia and have a far point that is considerably closer than infinity. A condition known as astigmatism results when the cornea is not spherical but instead is more curved in one direction. This causes horizontally extended objects to be focused on different parts of the retina than vertically extended objects, and results in distorted images.

All of these conditions can be corrected using corrective lenses. For presbyopia and hyperopia, a converging lens provides the extra curvature necessary to bring the near point closer to the eye while for myopia a diverging lens provides the curvature necessary to send the far point to infinity. Astigmatism is corrected with a cylindrical surface lens that curves more strongly in one direction than in another, compensating for the non-uniformity of the cornea.

The optical power of corrective lenses is measured in diopters, a value equal to the reciprocal of the focal length measured in meters; with a positive focal length corresponding to a converging lens and a negative focal length corresponding to a diverging lens. For lenses that correct for astigmatism as well, three numbers are given:

one for the spherical power, one for the cylindrical power, and one for the angle of orientation of the astigmatism.

Visual effects



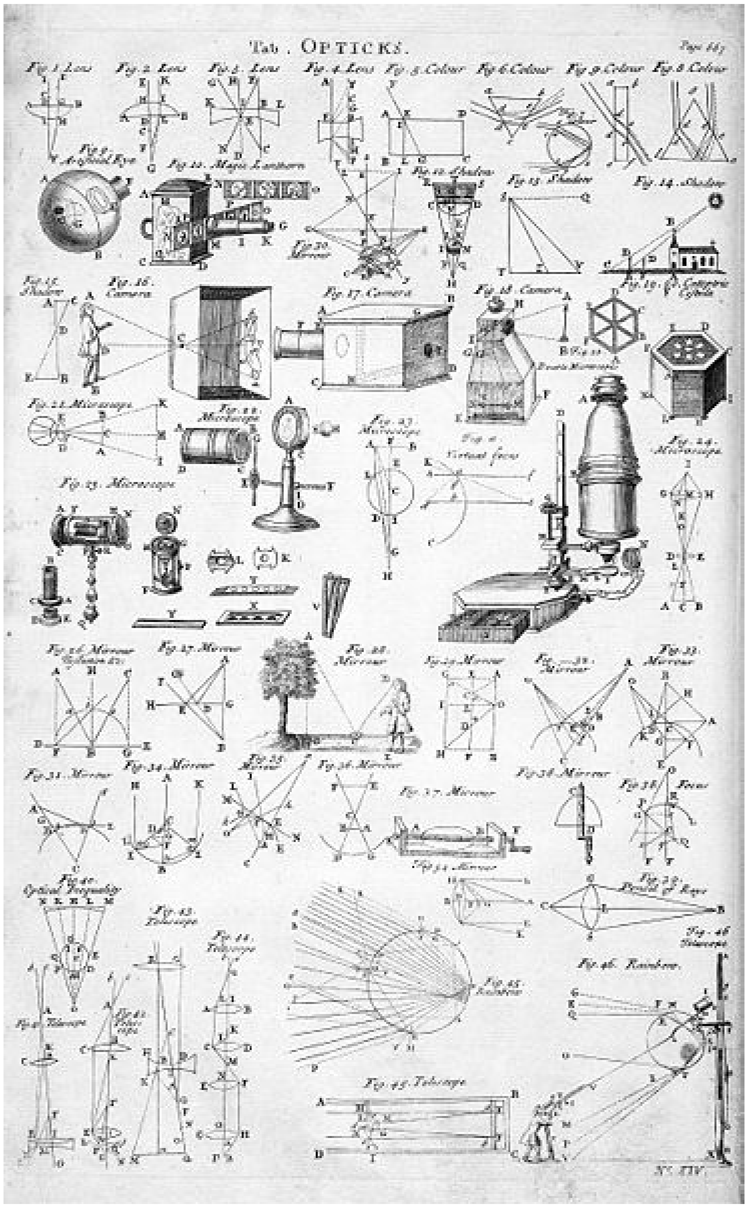
The Ponzo Illusion relies on the fact that parallel lines appear to converge as they approach infinity.

Optical illusions (also called visual illusions) are characterized by visually perceived images that differ from objective reality. The information gathered by the eye is processed in the brain to give a percept that differs from the object being imaged. Optical illusions can be the result of a variety of phenomena including physical effects that create images that are different from the objects that make them, the physiological effects on the eyes and brain of excessive stimulation (e.g. brightness, tilt, color, movement), and cognitive illusions where the eye and brain make unconscious inferences.

Cognitive illusions include some which result from the unconscious misapplication of certain optical principles. For example, the Ames room, Hering, Müller-Lyer, Orbison, Ponzo, Sander, and Wundt illusions all rely on the suggestion of the appearance of distance by using converging and diverging lines, in the same way that parallel light rays (or indeed any set of parallel lines) appear to converge at a vanishing point at infinity in two-dimensionally rendered images with artistic perspective. This suggestion is also responsible for the famous moon illusion where the moon, despite having essentially the same angular size, appears much larger near the horizon than it does at zenith. This illusion so confounded Ptolemy that he incorrectly attributed it to atmospheric refraction when he described it in his treatise, *Optics*.

Another type of optical illusion exploits broken patterns to trick the mind into perceiving symmetries or asymmetries that are not present. Examples include the café wall, Ehrenstein, Fraser spiral, Poggendorff, and Zöllner illusions. Related, but not strictly illusions, are patterns that occur due to the superimposition of periodic structures. For example transparent tissues with a grid structure produce shapes known as moiré patterns, while the superimposition of periodic transparent patterns comprising parallel opaque lines or curves produces line moiré patterns.

Optical instruments



Illustrations of various optical instruments from the 1728 *Cyclopaedia*

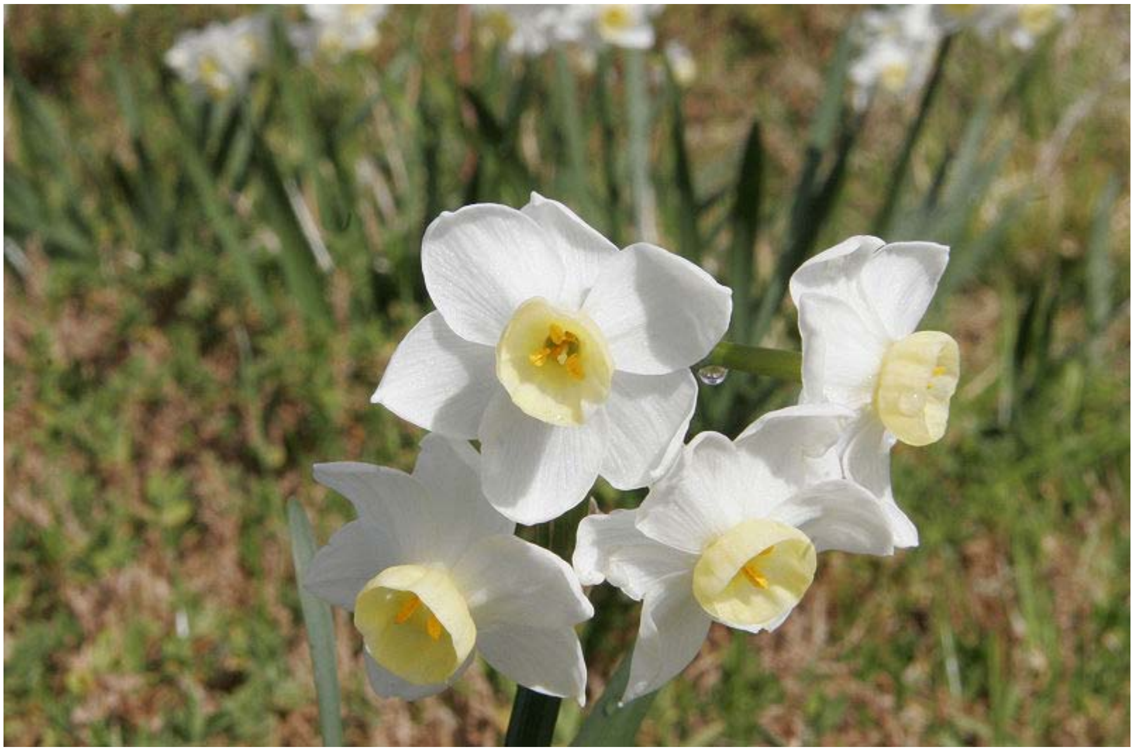
Single lenses have a variety of applications including photographic lenses, corrective lenses, and magnifying glasses while single mirrors are used in parabolic reflectors and rear-view mirrors. Combining a number of mirrors, prisms, and lenses produces compound optical instruments which have practical uses. For example, a periscope is simply two plane mirrors aligned to allow for viewing around obstructions. The most famous compound optical instruments in science are the microscope and the telescope which were both invented by the Dutch in the late 16th century.

Microscopes were first developed with just two lenses: an objective lens and an eyepiece. The objective lens is essentially a magnifying glass and was designed with a very small focal length while the eyepiece generally has a longer focal length. This has the effect of producing magnified images of close objects. Generally, an additional source of illumination is used since magnified images are dimmer due to the conservation of energy and the spreading of light rays over a larger surface area. Modern microscopes, known as *compound microscopes* have many lenses in them (typically four) to optimize the functionality and enhance image stability. A slightly different variety of microscope, the comparison microscope, looks at side-by-side images to produce a stereoscopic binocular view that appears three dimensional when used by humans.

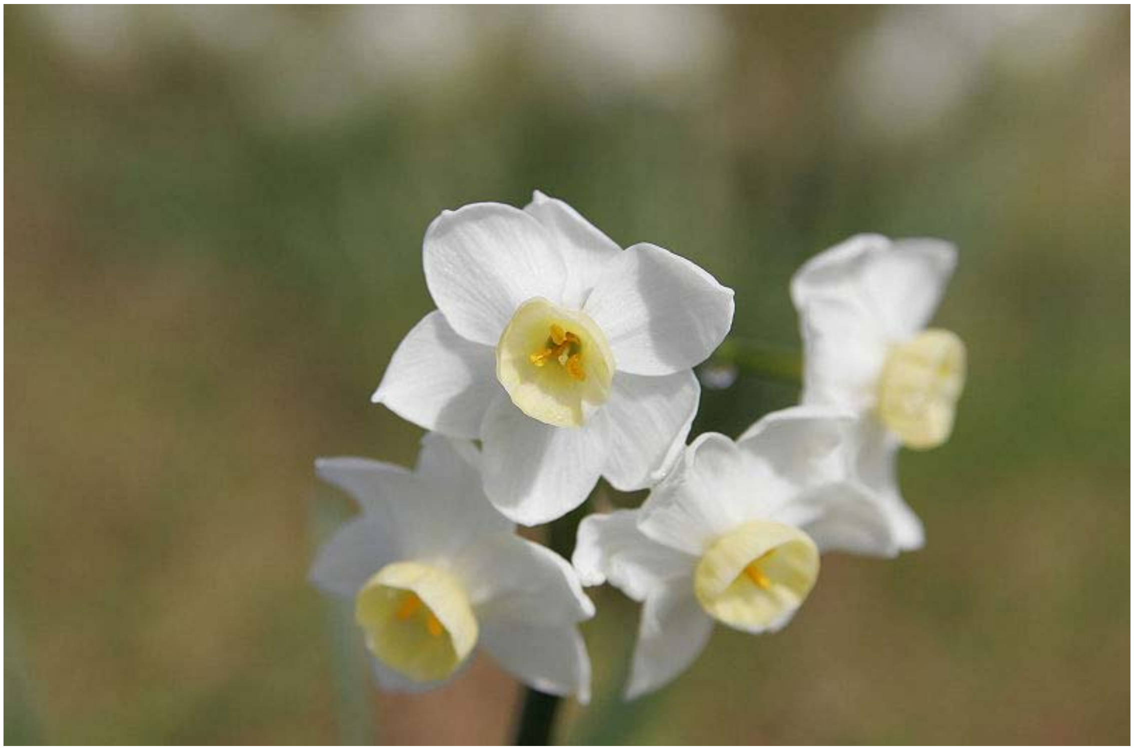
The first telescopes, called *refracting telescopes* were also developed with a single objective and eyepiece lens. In contrast to the microscope, the objective lens of the telescope was designed with a large focal length to avoid optical aberrations. The objective focuses an image of a distant object at its focal point which is adjusted to be at the focal point of an eyepiece of a much smaller focal length. The main goal of a telescope is not necessarily magnification, but rather collection of light which is determined by the physical size of the objective lens. Thus, telescopes are normally indicated by the diameters of their objectives rather than by the magnification which can be changed by switching eyepieces. Because the magnification of a telescope is equal to the focal length of the objective divided by the focal length of the eyepiece, smaller focal-length eyepieces cause greater magnification.

Since crafting large lenses is much more difficult than crafting large mirrors, most modern telescopes are *reflecting telescopes*, that is, telescopes that use a primary mirror rather than an objective lens. The same general optical considerations apply to reflecting telescopes that applied to refracting telescopes, namely, the larger the primary mirror, the more light collected, and the magnification is still equal to the focal length of the primary mirror divided by the focal length of the eyepiece. Professional telescopes generally do not have eyepieces and instead place an instrument (often a charge-coupled device) at the focal point instead.

Photography



Photograph taken with aperture *f*/32



Photograph taken with aperture *f*/5

The optics of photography involves both lenses and the medium in which the electromagnetic radiation is recorded, whether it be a plate, film, or charge-coupled device. Photographers must consider the reciprocity of the camera and the shot which is summarized by the relation

$$\text{Exposure} \propto \text{ApertureArea} \times \text{ExposureTime} \times \text{SceneLuminance}$$

In other words, the smaller the aperture (giving greater depth of focus), the less light coming in, so the length of time has to be increased (leading to possible blurriness if motion occurs). An example of the use of the law of reciprocity is the Sunny 16 rule which gives a rough estimate for the settings needed to estimate the proper exposure in daylight.

A camera's aperture is measured by a unitless number called the f-number or f-stop, *f*/#, often notated as *N*, and given by

$$f/\# = N = \frac{f}{D}$$

where *f* is the focal length, and *D* is the diameter of the entrance pupil. By convention, "*f*/#" is treated as a single symbol, and specific values of *f*/# are written by replacing the number sign with the value. The two ways to increase the f-stop are to either decrease the diameter of the entrance pupil or change to a longer focal length (in the case of a zoom lens, this can be done by simply adjusting the lens). Higher f-numbers also have a larger depth of field due to the lens approaching the limit of a pinhole camera which is able to focus all images perfectly, regardless of distance, but requires very long exposure times.

The field of view that the lens will provide changes with the focal length of the lens. There are three basic classifications based on the relationship to the diagonal size of the film or sensor size of the camera to the focal length of the lens:

- Normal lens: angle of view of about 50° (called *normal* because this angle considered roughly equivalent to human vision) and a focal length approximately equal to the diagonal of the film or sensor.
- Wide-angle lens: angle of view wider than 60° and focal length shorter than a normal lens.
- Long focus lens: angle of view narrow than a normal lens. This is any lens with a focal length longer than the diagonal measure of the film or sensor. The most common type of long focus lens is the telephoto lens, a design that uses a special *telephoto group* to be physically shorter than its focal length.

Modern zoom lenses may have some or all of these attributes.

The absolute value for the exposure time required depends on how sensitive to light the medium being used is (measured by the film speed, or, for digital media, by the quantum efficiency). Early photography used media that had very low light sensitivity, and so exposure times had to be long even for very bright shots. As technology has improved, so has the sensitivity through film cameras and digital cameras.

Other results from physical and geometrical optics apply to camera optics. For example, the maximum resolution capability of a particular camera set-up is determined by the diffraction limit associated with the pupil size and given, roughly, by the Rayleigh criterion.

Atmospheric optics

A colorful sky is often due to scattering of light off particulates and pollution, as in this photograph of a sunset during the October 2007 California wildfires.

The unique optical properties of the atmosphere cause a wide range of spectacular optical phenomena. The blue color of the sky is a direct result of Rayleigh scattering which redirects higher frequency (blue) sunlight back into the field of view of the observer. Because blue light is scattered more easily than red light, the sun takes on a reddish hue when it is observed through a thick atmosphere, as during a sunrise or sunset. Additional particulate matter in the sky can scatter different colors at different angles creating colorful glowing skies at dusk and dawn. Scattering off of ice crystals and other particles in the atmosphere are responsible for halos, afterglows, coronas, rays of sunlight, and sun dogs. The variation in these kinds of phenomena is due to different particle sizes and geometries.

Mirages are optical phenomena in which light rays are bent due to thermal variations in the refraction index of air, producing displaced or heavily distorted images of distant objects. Other dramatic optical phenomena associated with this include the Novaya Zemlya effect where the sun appears to rise earlier than predicted with a distorted shape. A spectacular form of refraction occurs with a temperature inversion called the Fata Morgana where objects on the horizon or even beyond the horizon, such as islands, cliffs, ships or icebergs, appear elongated and elevated, like "fairy tale castles".

Rainbows are the result of a combination of internal reflection and dispersive refraction of light in raindrops. A single reflection off the backs of an array of raindrops produces a rainbow with an angular size on the sky that ranges from 40° to 42° with red on the outside. Double rainbows are produced by two internal reflections with angular size of 50.5° to 54° with violet on the outside. Because rainbows are seen with the sun 180° away from the center of the rainbow, rainbows are more prominent the closer the sun is to the horizon.

Chapter 4

Optical Aberration

Aberrations are departures of the performance of an optical system from the predictions of paraxial optics. Aberration leads to blurring of the image produced by an image-forming optical system. It occurs when light from one point of an object after transmission through the system does not converge into (or does not diverge from) a single point. Instrument-makers need to correct optical systems to compensate for aberration. The articles on reflection, refraction and caustics discuss the general features of reflected and refracted rays.

Overview

Aberrations fall into two classes: *monochromatic* and *chromatic*. Monochromatic aberrations are caused by the geometry of the lens and occur both when light is reflected and when it is refracted. They appear even when using monochromatic light, hence the name.

Chromatic aberrations are caused by dispersion, the variation of a lens's refractive index with wavelength. They do not appear when monochromatic light is used.

Monochromatic aberrations

- Piston
- Tilt
- Defocus
- Spherical aberration
- Coma
- Astigmatism
- Field curvature
- Image distortion

Piston and tilt are not actually true optical aberrations, as they do not represent or model curvature in the wavefront. If an otherwise perfect wavefront is "aberrated" by piston and tilt, it will still form a perfect, aberration-free image, only shifted to a different position. Defocus is the lowest-order true optical aberration.

Chromatic aberrations

- Axial, or longitudinal, chromatic aberration
- Lateral, or transverse, chromatic aberration

Monochromatic aberration

The elementary theory of optical systems leads to the theorem: Rays of light proceeding from any *object point* unite in an *image point*; and therefore an *object space* is reproduced in an *image space.* The introduction of simple auxiliary terms, due to C. F. Gauss (*Dioptrische Untersuchungen*, Göttingen, 1841), named the focal lengths and focal planes, permits the determination of the image of any object for any system. The Gaussian theory, however, is only true so long as the angles made by all rays with the optical axis (the symmetrical axis of the system) are infinitely small, i.e. with infinitesimal objects, images and lenses; in practice these conditions are not realized, and the images projected by uncorrected systems are, in general, ill defined and often completely blurred, if the aperture or field of view exceeds certain limits.

The investigations of James Clerk Maxwell (*Phil.Mag.,* 1856; *Quart. Journ. Math.,* 1858) and Ernst Abbe showed that the properties of these reproductions, i.e. the relative position and magnitude of the images, are not special properties of optical systems, but necessary consequences of the supposition (in Abbe) of the reproduction of all points of a space in image points (Maxwell assumes a less general hypothesis), and are independent of the manner in which the reproduction is effected. These authors proved, however, that no optical system can justify these suppositions, since they are contradictory to the fundamental laws of reflexion and refraction. Consequently the Gaussian theory only supplies a convenient method of approximating to reality; and no constructor would attempt to realize this unattainable ideal. All that at present can be attempted is, to reproduce a single plane in another plane; but even this has not been altogether satisfactorily accomplished, aberrations always occur, and it is improbable that these will ever be entirely corrected.

This, and related general questions, have been treated — besides the above-mentioned authors — by M. Thiesen (*Berlin. Akad. Sitzber.,* 1890, xxxv. 799; *Berlin. Phys. Ges. Verh.,* 1892) and H. Bruns (*Leipzig. Math. Phys. Ber.,* 1895, xxi. 325) by means of Sir W. R. Hamilton's *characteristic function* (Irish Acad. Trans., *Theory of Systems of Rays*, 1828, et seq.). Reference may also be made to the treatise of Czapski-Eppenstein, pp. 155–161.

A review of the simplest cases of aberration will now be given.

Aberration of axial points (spherical aberration in the restricted sense)

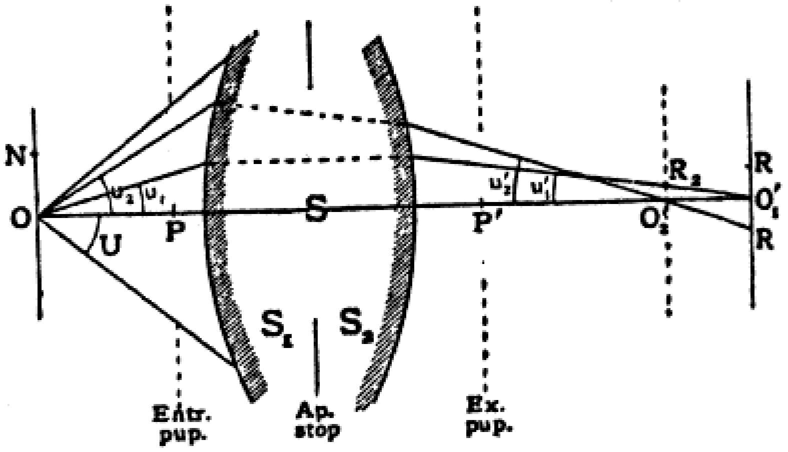


FIG. 5.

Figure 5

Let S (fig.5) be any optical system, rays proceeding from an axis point O under an angle u1 will unite in the axis point O'1; and those under an angle u2 in the axis point O'2. If there is refraction at a collective spherical surface, or through a thin positive lens, O'2 will lie in front of O'1 so long as the angle u2 is greater than u1 (*under correction*); and conversely with a dispersive surface or lenses (*over correction*). The caustic, in the first case, resembles the sign > (greater than); in the second < (less than). If the angle u1 is very small, O'1 is the Gaussian image; and O'1 O'2 is termed the *longitudinal aberration,* and O'1R the *lateral aberration* of the pencils with aperture u2. If the pencil with the angle u2 is that of the maximum aberration of all the pencils transmitted, then in a plane perpendicular to the axis at O'1 there is a circular *disk of confusion* of radius O'1R, and in a parallel plane at O'2 another one of radius O'2R2; between these two is situated the *disk of least confusion.*

The largest opening of the pencils, which take part in the reproduction of O, i.e. the angle u, is generally determined by the margin of one of the lenses or by a hole in a thin plate placed between, before, or behind the lenses of the system. This hole is termed the *stop* or *diaphragm*; Abbe used the term *aperture stop* for both the hole and the limiting margin of the lens. The component S1 of the system, situated between the aperture stop and the object O, projects an image of the diaphragm, termed by Abbe the *entrance pupil*; the *exit pupil* is the image formed by the component S2, which is placed behind the aperture stop. All rays which issue from O and pass through the aperture stop also pass through the entrance and exit pupils, since these are images of the aperture stop. Since the maximum aperture of the pencils issuing from O is the angle u subtended by the entrance pupil at this point, the magnitude of the aberration will be determined by the position and diameter of the entrance pupil. If the system be entirely behind the aperture stop, then this is itself the entrance pupil (*front stop*); if entirely in front, it is the exit pupil (*back stop*).

If the object point be infinitely distant, all rays received by the first member of the system are parallel, and their intersections, after traversing the system, vary according to their *perpendicular height of incidence,* i.e. their distance from the axis. This distance replaces the angle u in the preceding considerations; and the aperture, i.e. the radius of the entrance pupil, is its maximum value.

Aberration of elements, i.e. smallest objects at right angles to the axis

If rays issuing from O (fig. 5) be concurrent, it does not follow that points in a portion of a plane perpendicular at O to the axis will be also concurrent, even if the part of the plane be very small. With a considerable aperture, the neighboring point N will be reproduced, but attended by aberrations comparable in magnitude to ON. These aberrations are avoided if, according to Abbe, the *sine condition,* sin u'1/sin u1=sin u'2/sin u2, holds for all rays reproducing the point O. If the object point O is infinitely distant, u1 and u2 are to be replaced by h1 and h2, the perpendicular heights of incidence; the *sine condition* then becomes sin u'1/h1 sin u'2/h2. A system fulfilling this condition and free from spherical aberration is called *aplanatic* (Greek a-, privative, plann, a wandering). This word was first used by Robert Blair (d. 1828), professor of practical astronomy at Edinburgh University, to characterize a superior achromatism, and, subsequently, by many writers to denote freedom from spherical aberration. Both the aberration of axis points, and the deviation from the sine condition, rapidly increase in most (uncorrected) systems with the aperture.

Aberration of lateral object points (points beyond the axis) with narrow pencils. Astigmatism.

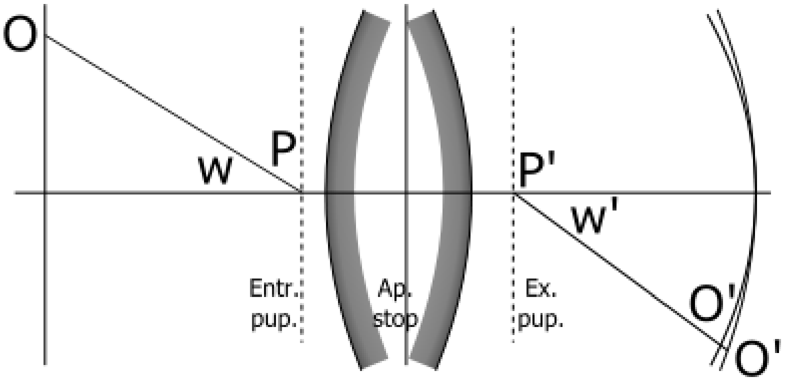


Figure 6

A point O (fig. 6) at a finite distance from the axis (or with an infinitely distant object, a point which subtends a finite angle at the system) is, in general, even then not sharply reproduced, if the pencil of rays issuing from it and traversing the system is made infinitely narrow by reducing the aperture stop; such a pencil consists of the rays which can pass from the object point through the now infinitely small entrance pupil. It is seen (ignoring exceptional cases) that the pencil does not meet the refracting or reflecting surface at right angles; therefore it is astigmatic (Gr. a-, privative, stigmia, a point). Naming the central ray passing through the entrance pupil the *axis of the pencil* or *principal ray,* it can be said: the rays of the pencil intersect, not in one point, but in two focal lines, which can be assumed to be at right angles to the principal ray; of these, one lies in the plane containing the principal ray and the axis of the system, i.e. in the *first principal section* or *meridional section*, and the other at right angles to it, i.e. in the second principal section or sagittal section. We receive, therefore, in no single intercepting plane behind the system, as, for example, a focusing screen, an image of the object point; on the other hand, in each of two planes lines O' and O" are separately formed (in neighboring planes ellipses are formed), and in a plane between O' and O" a circle of least confusion. The interval O'O", termed the astigmatic difference, increases, in general, with the angle W made by the principal ray OP with the axis of the system, i.e. with the field of view. Two *astigmatic image surfaces* correspond to one object plane; and these are in contact at the axis point; on the one lie the focal lines of the first kind, on the other those of the second. Systems in which the two astigmatic surfaces coincide are termed anastigmatic or stigmatic.

Sir Isaac Newton was probably the discoverer of astigmation; the position of the astigmatic image lines was determined by Thomas Young (*A Course of Lectures on Natural Philosophy,* 1807); and the theory was developed by Allvar Gullstrand. A bibliography by P. Culmann is given in Moritz von Rohr's *Die Bilderzeugung in optischen Instrumenten*.

Aberration of lateral object points with broad pencils. Coma.

By opening the stop wider, similar deviations arise for lateral points as have been already discussed for axial points; but in this case they are much more complicated. The course of the rays in the meridional section is no longer symmetrical to the principal ray of the pencil; and on an intercepting plane there appears, instead of a luminous point, a patch of light, not symmetrical about a point, and often exhibiting a resemblance to a comet having its tail directed towards or away from the axis. From this appearance it takes its name. The unsymmetrical form of the meridional pencil—formerly the only one considered—is coma in the narrower sense only; other errors of coma have been treated by Arthur König and Moritz von Rohr, and later by Allvar Gullstrand.

Curvature of the field of the image

If the above errors be eliminated, the two astigmatic surfaces united, and a sharp image obtained with a wide aperture—there remains the necessity to correct the curvature of the image surface, especially when the image is to be received upon a plane surface, e.g. in photography. In most cases the surface is concave towards the system.

Distortion of the image

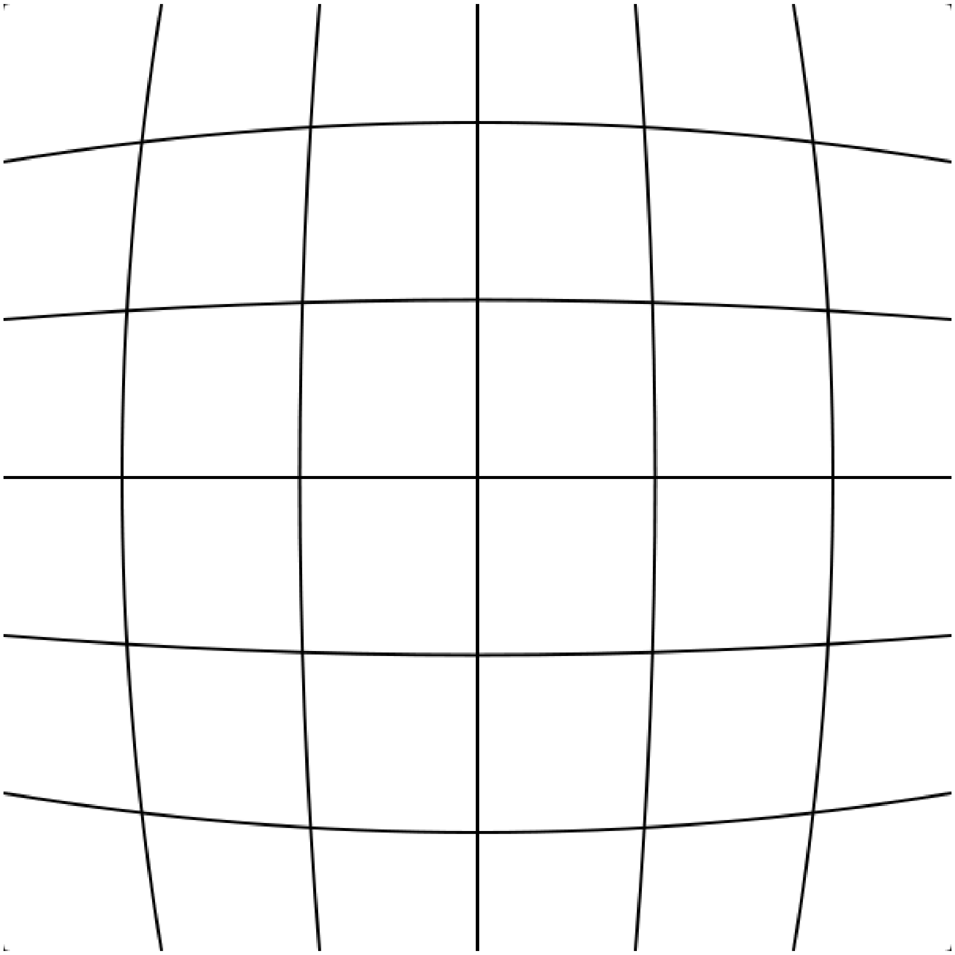


Fig. 7a: Barrel distortion

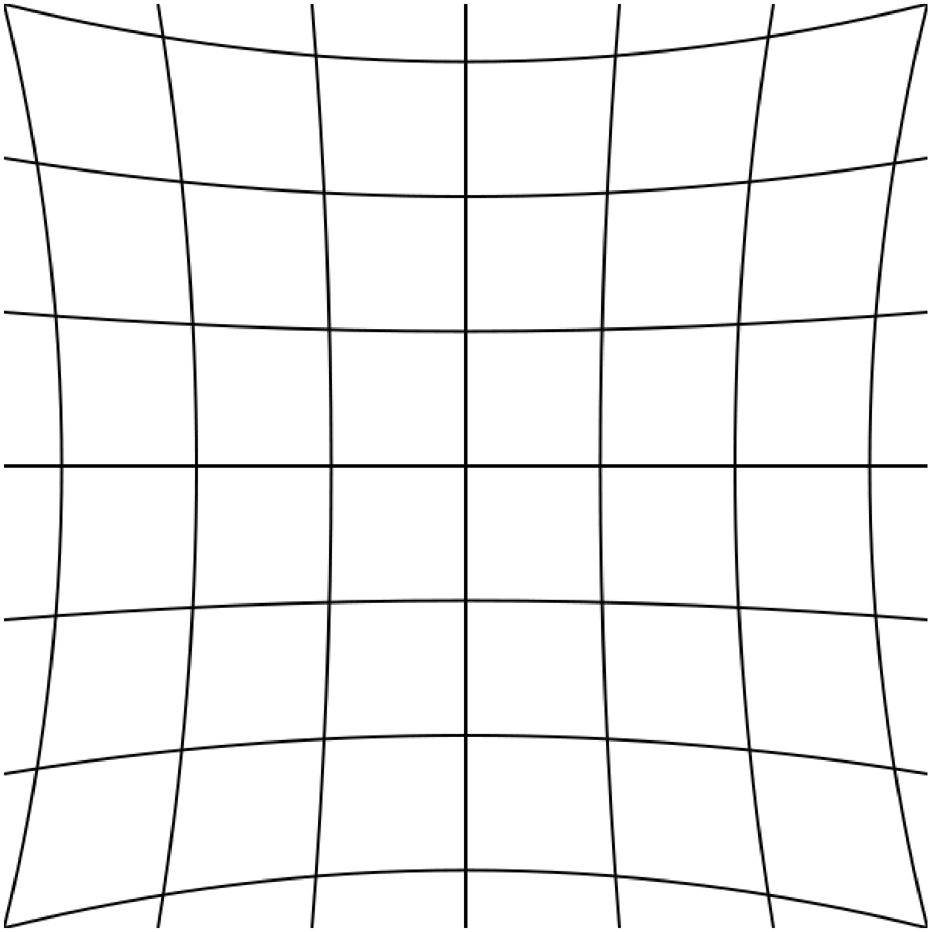


Fig. 7b: Pincushion distortion

Even if the image is sharp, it may be distorted compared to ideal pinhole projection. In pinhole projection, the magnification of an object is inversely proportional to its distance to the camera along the optical axis so that a camera pointing directly at a flat surface reproduces that flat surface. Distortion can be thought of as stretching the image non-uniformly, or, equivalently, as a variation in magnification across the field. While "distortion" can include arbitrary deformation of an image, the most pronounced modes of distortion produced by conventional imaging optics is "barrel distortion", in which the center of the image is magnified more than the perimeter (figure 7a). The reverse, in which the perimeter is magnified more than the center, is known as "pincushion distortion" (figure 7b). This effect is called lens distortion or image distortion, and there are algorithms to correct it.

Systems free of distortion are called *orthoscopic* (orthos, right, skopein to look) or *rectilinear* (straight lines).

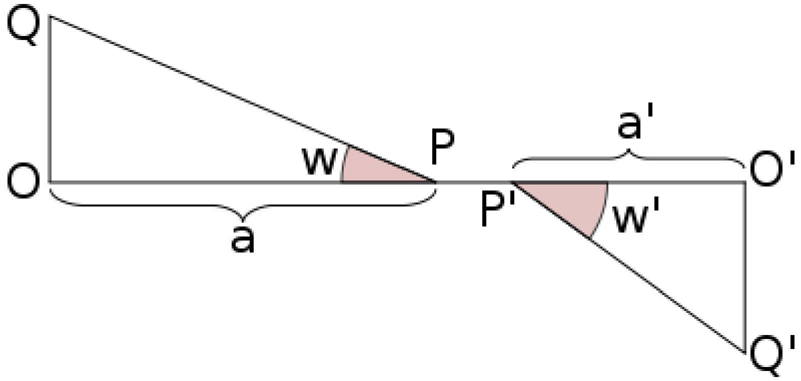


Figure 8

This aberration is quite distinct from that of the sharpness of reproduction; in unsharp, reproduction, the question of distortion arises if only parts of the object can be recognized in the figure. If, in an unsharp image, a patch of light corresponds to an object point, the *center of gravity* of the patch may be regarded as the image point, this being the point where the plane receiving the image, e.g., a focusing screen, intersects the ray passing through the middle of the stop. This assumption is justified if a poor image on the focusing screen remains stationary when the aperture is diminished; in practice, this generally occurs. This ray, named by Abbe a *principal ray* (not to be confused with the *principal rays* of the Gaussian theory), passes through the center of the entrance pupil before the first refraction, and the center of the exit pupil after the last refraction. From this it follows that correctness of drawing depends solely upon the principal rays; and is independent of the sharpness or curvature of the image field. Referring to fig. 8, we have O'Q'/OQ = a' tan w'/a tan w = 1/N, where N is the *scale* or magnification of the image. For N to be constant for all values of w, a' tan w'/a tan w must also be constant. If the ratio a'/a be sufficiently constant, as is often the case, the above relation reduces to the *condition of Airy,* i.e. tan w'/ tan w= a constant. This simple relation is fulfilled in all systems which are symmetrical with respect to their diaphragm (briefly named *symmetrical or holosymmetrical objectives*), or which consist of two like, but different-sized, components, placed from the diaphragm in the ratio of their size, and presenting the same curvature to it (hemisymmetrical objectives); in these systems tan w' / tan w = 1.

The constancy of a'/a necessary for this relation to hold was pointed out by R. H. Bow (Brit. Journ. Photog., 1861), and Thomas Sutton (Photographic Notes, 1862); it has been treated by O. Lummer and by M. von Rohr (Zeit. f. Instrumentenk., 1897, 17, and 1898, 18, p. 4). It requires the middle of the aperture stop to be reproduced in the centers of the entrance and exit pupils without spherical aberration. M. von Rohr showed that for systems fulfilling neither the Airy nor the Bow-Sutton condition, the ratio a' cos w'/a tan w will be constant for one distance of the object. This combined condition is exactly fulfilled by holosymmetrical objectives reproducing with the scale 1, and by hemisymmetrical, if the scale of reproduction be equal to the ratio of the sizes of the two components.

Zernike model of aberrations

Circular wavefront profiles associated with aberrations may be mathematically modeled using Zernike polynomials. Developed by Frits Zernike in the 1930s, Zernike's polynomials are orthogonal over a circle of unit radius. A complex, aberrated wavefront profile may be curve-fitted with Zernike polynomials to yield a set of fitting coefficients that individually represent different types of aberrations. These Zernike coefficients are linearly independent, thus individual aberration contributions to an overall wavefront may be isolated and quantified separately.

There are even and odd Zernike polynomials. The even Zernike polynomials are defined as

$$Z_n^m(\rho,\phi) = R_n^m(\rho)\,\cos(m\,\phi)$$

and the odd Zernike polynomials as

$$Z_n^{-m}(\rho,\phi) = R_n^m(\rho)\,\sin(m\,\phi),$$

where *m* and *n* are nonnegative integers with $n \geq m$, φ is the azimuthal angle in radians, and ρ is the normalized radial distance. The radial polynomials R_n^m have no azimuthal dependence, and are defined as

$$R_n^m(\rho) = \sum_{k=0}^{(n-m)/2} \frac{(-1)^k\,(n-k)!}{k!\,((n+m)/2-k)!\,((n-m)/2-k)!}\;\rho^{n-2\,k} \quad \text{if } n-m \text{ is even}$$

and $R_n^m(\rho) = 0$if *n* − *m* is odd.

The first few Zernike polynomials are:

a_0	"Piston", equal to the mean value of the wavefront
$a_1 \times \rho\cos(\theta)$	"X-Tilt", the deviation of the overall beam in the sagittal direction
$a_2 \times \rho\sin(\theta)$	"Y-Tilt", the deviation of the overall beam in the tangential direction
$a_3 \times (2\rho^2 - 1)$	"Defocus", a parabolic wavefront resulting from being out of focus
$a_4 \times \rho^2\cos(2\theta)$	"0° Astigmatism", a cylindrical shape along the X or Y axis
$a_5 \times \rho^2\sin(2\theta)$	"45° Astigmatism", a cylindrical shape oriented at ±45° from the X axis
$a_6 \times (3\rho^2 - 2)\rho\cos(\theta)$	"X-Coma", comatic image flaring in the horizontal direction

$a_7 \times (3\rho^2 - 2)\rho\sin(\theta)$ "Y-Coma", comatic image flaring in the vertical direction

$a_8 \times (6\rho^4 - 6\rho^2 + 1)$ "Third order spherical aberration"

where ρ is the normalized pupil radius with $0 \leq \rho \leq 1$, θ is the azimuthal angle around the pupil with $0 \leq \theta \leq 2\pi$, and the fitting coefficients $a_0, \dots, a_8$are the wavefront errors in wavelengths.

As in Fourier synthesis using sines and cosines, a wavefront may be perfectly represented by a sufficiently large number of higher-order Zernike polynomials. However, wavefronts with very steep gradients or very high spatial frequency structure, such as produced by propagation through atmospheric turbulence or aerodynamic flowfields, are not well modeled by Zernike polynomials, which tend to low-pass filter fine spatial definition in the wavefront. In this case, other fitting methods such as fractals or singular value decomposition may yield improved fitting results.

The circle polynomials were introduced by Fritz Zernike to evaluate the point image of an aberrated optical system taking into account the effects of diffraction. The perfect point image in the presence of diffraction had already been described by Airy, as early as 1835. It took almost hundred years to arrive at a comprehensive theory and modeling of the point image of aberrated systems (Zernike and Nijboer). The analysis by Nijboer and Zernike describes the intensity distribution close to the optimum focal plane. An extended theory that allows the calculation of the point image amplitude and intensity over a much larger volume in the focal region was recently developed (Extended Nijboer-Zernike theory). This Extended Nijboer-Zernike theory of point image or ‘point-spread function’ formation has found applications in general research on image formation, especially for systems with a high numerical aperture, and in characterizing optical systems with respect to their aberrations.

Analytic treatment of aberrations

The preceding review of the several errors of reproduction belongs to the *Abbe theory of aberrations,* in which definite aberrations are discussed separately; it is well suited to practical needs, for in the construction of an optical instrument certain errors are sought to be eliminated, the selection of which is justified by experience. In the mathematical sense, however, this selection is arbitrary; the reproduction of a finite object with a finite aperture entails, in all probability, an infinite number of aberrations. This number is only finite if the object and aperture are assumed to be *infinitely small of a certain order*; and with each order of infinite smallness, i.e. with each degree of approximation to reality (to finite objects and apertures), a certain number of aberrations is associated. This connection is only supplied by theories which treat aberrations generally and analytically by means of indefinite series.

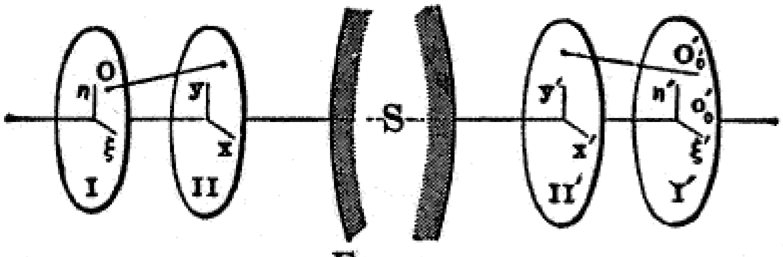


FIG. 9.

Figure 9

A ray proceeding from an object point O (fig. 9) can be defined by the coordinates (ξ, η). Of this point O in an object plane I, at right angles to the axis, and two other coordinates (x, y), the point in which the ray intersects the entrance pupil, i.e. the plane II. Similarly the corresponding image ray may be defined by the points (ξ', η'), and (x', y'), in the planes I' and II'. The origins of these four plane coordinate systems may be collinear with the axis of the optical system; and the corresponding axes may be parallel. Each of the four coordinates ξ', η', x', y' are functions of ξ, η, x, y; and if it be assumed that the field of view and the aperture be infinitely small, then ξ, η, x, y are of the same order of infinitesimals; consequently by expanding ξ', η', x', y' in ascending powers of ξ, η, x, y, series are obtained in which it is only necessary to consider the lowest powers. It is readily seen that if the optical system be symmetrical, the origins of the coordinate systems collinear with the optical axis and the corresponding axes parallel, then by changing the signs of ξ, η, x, y, the values ξ', η', x', y' must likewise change their sign, but retain their arithmetical values; this means that the series are restricted to odd powers of the unmarked variables.

The nature of the reproduction consists in the rays proceeding from a point O being united in another point O'; in general, this will not be the case, for ξ', η' vary if ξ, η be constant, but x, y variable. It may be assumed that the planes I' and II' are drawn where the images of the planes I and II are formed by rays near the axis by the ordinary Gaussian rules; and by an extension of these rules, not, however, corresponding to reality, the Gauss image point O'_0, with coordinates ξ'_0, η'_0, of the point O at some distance from the axis could be constructed. Writing $D\xi'=\xi'-\xi'_0$ and $D\eta'=\eta'-\eta'_0$, then $D\xi'$ and $D\eta'$ are the aberrations belonging to ξ, η and x, y, and are functions of these magnitudes which, when expanded in series, contain only odd powers, for the same reasons as given above. On account of the aberrations of all rays which pass through O, a patch of light, depending in size on the lowest powers of ξ, η, x, y which the aberrations contain, will be formed in the plane I'. These degrees, named by (J. Petzval (*Bericht uber die Ergebnisse einiger dioptrischer Untersuchungen*, Buda Pesth, 1843; *Akad. Sitzber., Wien,* 1857, vols. xxiv. xxvi.) *the numerical orders of the image,* are consequently only odd powers; the condition for the formation of an image of the mth order is that in the series for $D\xi'$ and $D\eta'$ the coefficients of the powers of the 3rd, 5th…(m-2)th degrees must vanish. The images of the Gauss theory being of the third order, the next problem is to obtain an image of 5th order, or to make the coefficients of the powers of 3rd degree zero. This necessitates the satisfying of five equations; in other words, there are five alterations of the 3rd order, the vanishing of which produces an image of the 5th order.

The expression for these coefficients in terms of the constants of the optical system, i.e. the radii, thicknesses, refractive indices and distances between the lenses, was solved by L. Seidel (Astr. Nach., 1856, p. 289); in 1840, J. Petzval constructed his portrait objective, from similar calculations which have never been published. The theory was elaborated by S. Finterswalder (Munchen. Acad. Abhandl., 1891, 17, p. 519), who also published a posthumous paper of Seidel containing a short view of his work (*München. Akad. Sitzber.,* 1898, 28, p. 395); a simpler form was given by A. Kerber (*Beiträge zur Dioptrik*, Leipzig, 1895-6-7-8-9). A. Konig and M. von Rohr have represented Kerber's method, and have deduced the Seidel formulae from geometrical considerations based on the Abbe method, and have interpreted the analytical results geometrically (pp. 212–316).

The aberrations can also be expressed by means of the *characteristic function* of the system and its differential coefficients, instead of by the radii, &c., of the lenses; these formulae are not immediately applicable, but give, however, the relation between the number of aberrations and the order. Sir William Rowan Hamilton (British Assoc. Report, 1833, p. 360) thus derived the aberrations of the third order; and in later times the method was pursued by Clerk Maxwell (*Proc. London Math. Soc.,* 1874–1875; M. Thiesen (*Berlin. Akad. Sitzber.,* 1890, 35, p. 804), H. Bruns (*Leipzig. Math. Phys. Ber.,* 1895, 21, p. 410), and particularly successfully by K. Schwarzschild (*Göttingen. Akad. Abhandl.,* 1905, 4, No. 1), who thus discovered the aberrations of the 5th order (of which there are nine), and possibly the shortest proof of the practical (Seidel) formulae. A. Gullstrand (vide supra, and *Ann. d. Phys.,* 1905, 18, p. 941) founded his theory of aberrations on the differential geometry of surfaces.

The aberrations of the third order are: (1) aberration of the axis point; (2) aberration of points whose distance from the axis is very small, less than of the third order — the deviation from the sine condition and coma here fall together in one class; (3) astigmatism; (4) curvature of the field; (5) distortion.

> **(1)** Aberration of the third order of axis points is dealt with in all text-books on optics. It is very important in telescope design. In telescopes aperture is usually taken as the linear diameter of the objective. It is not the same as microscope aperture which is based on the entrance pupil or field of view as seen from the object and is expressed as an angular measurement. Higher order aberrations in telescope design can be mostly neglected. For microscopes it cannot be neglected. For a single lens of very small thickness and given power, the aberration depends upon the ratio of the radii r:r', and is a minimum (but never zero) for a certain value of this ratio; it varies inversely with the refractive index (the power of the lens remaining constant). The total aberration of two or more very thin lenses in contact, being the sum of the individual aberrations, can be zero. This is also possible if the lenses have the same algebraic sign. Of thin positive lenses with n=1.5, four are necessary to correct spherical aberration of the third order. These systems, however, are not of great practical importance. In most cases, two thin

lenses are combined, one of which has just so strong a positive aberration (*under-correction,* vide supra) as the other a negative; the first must be a positive lens and the second a negative lens; the powers, however: may differ, so that the desired effect of the lens is maintained. It is generally an advantage to secure a great refractive effect by several weaker than by one high-power lens. By one, and likewise by several, and even by an infinite number of thin lenses in contact, no more than two axis points can be reproduced without aberration of the third order. Freedom from aberration for two axis points, one of which is infinitely distant, is known as *Herschel's condition.* All these rules are valid, inasmuch as the thicknesses and distances of the lenses are not to be taken into account.
(2) The condition for freedom from coma in the third order is also of importance for telescope objectives; it is known as *Fraunhofer's condition.* (4) After eliminating the aberration On the axis, coma and astigmatism, the relation for the flatness of the field in the third order is expressed by the *Petzval equation,* S1/r(n'-n) = 0, where r is the radius of a refracting surface, n and n' the refractive indices of the neighboring media, and S the sign of summation for all refracting surfaces.

Practical elimination of aberrations

The classical imaging problem is to reproduce perfectly a finite plane (the object) onto another plane (the image) through a finite aperture. It is impossible to do so perfectly for *more than one* such pairs of planes (this was proven with increasing generality by Maxwell in 1858, by Bruns in 1895, and by Carathéodory in 1926, see summary in Walther, A., J. Opt. Soc. Am. A **6**, 415–422 (1989)). For a single pair of planes (e.g. for a single focus setting of an objective), however, the problem can in principle be solved perfectly. Examples of such a theoretically perfect system include the Luneburg lens and the Maxwell fish-eye.

Practical methods solve this problem with an accuracy which mostly suffices for the special purpose of each species of instrument. The problem of finding a system which reproduces a given object upon a given plane with given magnification (insofar as aberrations must be taken into account) could be dealt with by means of the approximation theory; in most cases, however, the analytical difficulties were too great for older calculation methods but may be ameliorated by application of modern computer systems. Solutions, however, have been obtained in special cases. At the present time constructors almost always employ the inverse method: they compose a system from certain, often quite personal experiences, and test, by the trigonometrical calculation of the paths of several rays, whether the system gives the desired reproduction (examples are given in A. Gleichen, *Lehrbuch der geometrischen Optik*, Leipzig and Berlin, 1902). The radii, thicknesses and distances are continually altered until the errors of the image become sufficiently small. By this method only certain errors of reproduction are investigated, especially individual members, or all, of those named above. The analytical approximation theory is often employed provisionally, since its accuracy does not generally suffice.

In order to render spherical aberration and the deviation from the sine condition small throughout the whole aperture, there is given to a ray with a finite angle of aperture u* (width infinitely distant objects: with a finite height of incidence h*) the same distance of intersection, and the same sine ratio as to one neighboring the axis (u* or h* may not be much smaller than the largest aperture U or H to be used in the system). The rays with an angle of aperture smaller than u* would not have the same distance of intersection and the same sine ratio; these deviations are called *zones,* and the constructor endeavors to reduce these to a minimum. The same holds for the errors depending upon the angle of the field of view, w: astigmatism, curvature of field and distortion are eliminated for a definite value, w*, *zones of astigmatism, curvature of field and distortion,* attend smaller values of w. The practical optician names such systems: *corrected for the angle of aperture u* (the height of incidence h*) or the angle of field of view w*.* Spherical aberration and changes of the sine ratios are often represented graphically as functions of the aperture, in the same way as the deviations of two astigmatic image surfaces of the image plane of the axis point are represented as functions of the angles of the field of view.

The final form of a practical system consequently rests on compromise; enlargement of the aperture results in a diminution of the available field of view, and vice versa. But the larger aperture will give the larger resolution. The following may be regarded as typical:

> (1) Largest aperture; necessary corrections are — for the axis point, and sine condition; errors of the field of view are almost disregarded; example — high-power microscope objectives.
> (2) Wide angle lens; necessary corrections are — for astigmatism, curvature of field and distortion; errors of the aperture only slightly regarded; examples — photographic widest angle objectives and oculars.
> Between these extreme examples stands the normal lens: this is corrected more with regard to aperture; objectives for groups more with regard to the field of view.
> (3) Long focus lenses have small fields of view and aberrations on axis are very important. Therefore zones will be kept as small as possible and design should emphasize simplicity. Because of this these lenses are the best for analytical computation.

Chromatic or color aberration

In optical systems composed of lenses, the position, magnitude and errors of the image depend upon the refractive indices of the glass employed. Since the index of refraction varies with the color or wavelength of the light, it follows that a system of lenses (uncorrected) projects images of different colors in somewhat different places and sizes and with different aberrations; i.e. there are *chromatic differences* of the distances of intersection, of magnifications, and of monochromatic aberrations. If mixed light be employed (e.g. white light) all these images are formed; and since they are all ultimately intercepted by a plane (the retina of the eye, a focusing screen of a camera, etc.), they cause a confusion, named chromatic aberration; for instance, instead of a white margin on a dark background, there is perceived a colored margin, or narrow spectrum. The absence of this error is termed achromatism, and an optical system so corrected is termed achromatic. A system is said to be *chromatically under-corrected* when it shows the same kind of chromatic error as a thin positive lens, otherwise it is said to be *over-corrected.*

If, in the first place, monochromatic aberrations be neglected — in other words, the Gaussian theory be accepted — then every reproduction is determined by the positions of the focal planes, and the magnitude of the focal lengths, or if the focal lengths, as ordinarily happens, be equal, by three constants of reproduction. These constants are determined by the data of the system (radii, thicknesses, distances, indices, etc., of the lenses); therefore their dependence on the refractive index, and consequently on the color, are calculable. The refractive indices for different wavelengths must be known for each kind of glass made use of. In this manner the conditions are maintained that any one constant of reproduction is equal for two different colors, i.e. this constant is achromatized. For example, it is possible, with one thick lens in air, to achromatize the position of a focal plane of the magnitude of the focal length. If all three constants of reproduction be achromatized, then the Gaussian image for all distances of objects is the same for the two colors, and the system is said to be in *stable achromatism.*

In practice it is more advantageous (after Abbe) to determine the chromatic aberration (for instance, that of the distance of intersection) for a fixed position of the object, and express it by a sum in which each component conlins the amount due to each refracting surface. In a plane containing the image point of one color, another colour produces a disk of confusion; this is similar to the confusion caused by two *zones* in spherical aberration. For infinitely distant objects the radius Of the chromatic disk of confusion is proportional to the linear aperture, and independent of the focal length (*vide supra*, *Monochromatic Aberration of the Axis Point*); and since this disk becomes the less harmful with an increasing image of a given object, or with increasing focal length, it follows that the deterioration of the image is proportional to the ratio of the aperture to the focal length, i.e. the *relative aperture.* (This explains the gigantic focal lengths in vogue before the discovery of achromatism.)

Examples:

(a) In a very thin lens, in air, only one constant of reproduction is to be observed, since the focal length and the distance of the focal point are equal. If the refractive index for one color be n, and for another $n + dn$, and the powers, or reciprocals of the focal lengths, be f and $f + df$, then (1) $df / f = dn / (n - 1) = 1 / n$; dn is called the dispersion, and n the dispersive power of the glass.
(b) Two thin lenses in contact: let f_1 and f_2 be the powers corresponding to the lenses of refractive indices n_1 and n_2 and radii r'_1, r''_1, and r'_2, r''_2 respectively; let f denote the total power, and df, dn_1, dn_2 the changes of f, n_1, and n_2 with the color. Then the following relations hold:
(2) $f = f_1 - f_2 = (n_1 - 1)(1 / r'_1 - 1 / r''_1) + (n2 - 1)(1 / r'_2 - 1 / r''_2) = (n_1 - 1)k_1 + (n_2 - 1)k_2$; and
(3) $df = k_1dn_1 + k_2dn_2$. For achromatism $df = 0$, hence, from (3),

(4) $k_1 / k_2 = - dn_2 / dn_1$, or $f_1 / f_2 = - n_1 / n_2$. Therefore f_1 and f_2 must have different algebraic signs, or the system must be composed of a collective and a dispersive lens. Consequently the powers of the two must be different (in order that f be not zero (equation 2)), and the dispersive powers must also be different (according to 4).

Newton failed to perceive the existence of media of different dispersive powers required by achromatism; consequently he constructed large reflectors instead of refractors. James Gregory and Leonhard Euler arrived at the correct view from a false conception of the achromatism of the eye; this was determined by Chester More Hall in 1728, Klingenstierna in 1754 and by Dollond in 1757, who constructed the celebrated achromatic telescopes.

Glass with weaker dispersive power (greater v) is named *crown glass*; that with greater dispersive power, *flint glass*. For the construction of an achromatic collective lens (f positive) it follows, by means of equation (4), that a collective lens I. of crown glass and a dispersive lens II. of flint glass must be chosen; the latter, although the weaker, corrects the other chromatically by its greater dispersive power. For an achromatic dispersive lens the converse must be adopted. This is, at the present day, the ordinary type, e.g., of telescope objective (fig. 10); the values of the four radii must satisfy the equations (2) and (4). Two other conditions may also be postulated: one is always the elimination of the aberration on the axis; the second either the *Herschel* or *Fraunhofer Condition,* the latter being the best vide supra, *Monochromatic Aberration*). In practice, however, it is often more useful to avoid the second condition by making the lenses have contact, i.e. equal radii. According to P. Rudolph (Eder's Jahrb. f. Photog., 1891, 5, p. 225; 1893, 7, p. 221), cemented objectives of thin lenses permit the elimination of spherical aberration on the axis, if, as above, the collective lens has a smaller refractive index; on the other hand, they permit the elimination of astigmatism and curvature of the field, if the collective lens has a greater refractive index. Should the cemented system be positive, then the more powerful lens must be positive; and, according to (4), to the greater power belongs the weaker dispersive power (greater v), that is to say, crown glass; consequently the crown glass must have the greater refractive index for astigmatic and plane images. In all earlier kinds of glass, however, the dispersive power increased with the refractive index; that is, v decreased as n increased; but some of the Jena glasses by E. Abbe and O. Schott were crown glasses of high refractive index, and achromatic systems from such crown glasses, with flint glasses of lower refractive index, are called the *new achromats,* and were employed by P. Rudolph in the first *anastigmats* (photographic objectives).

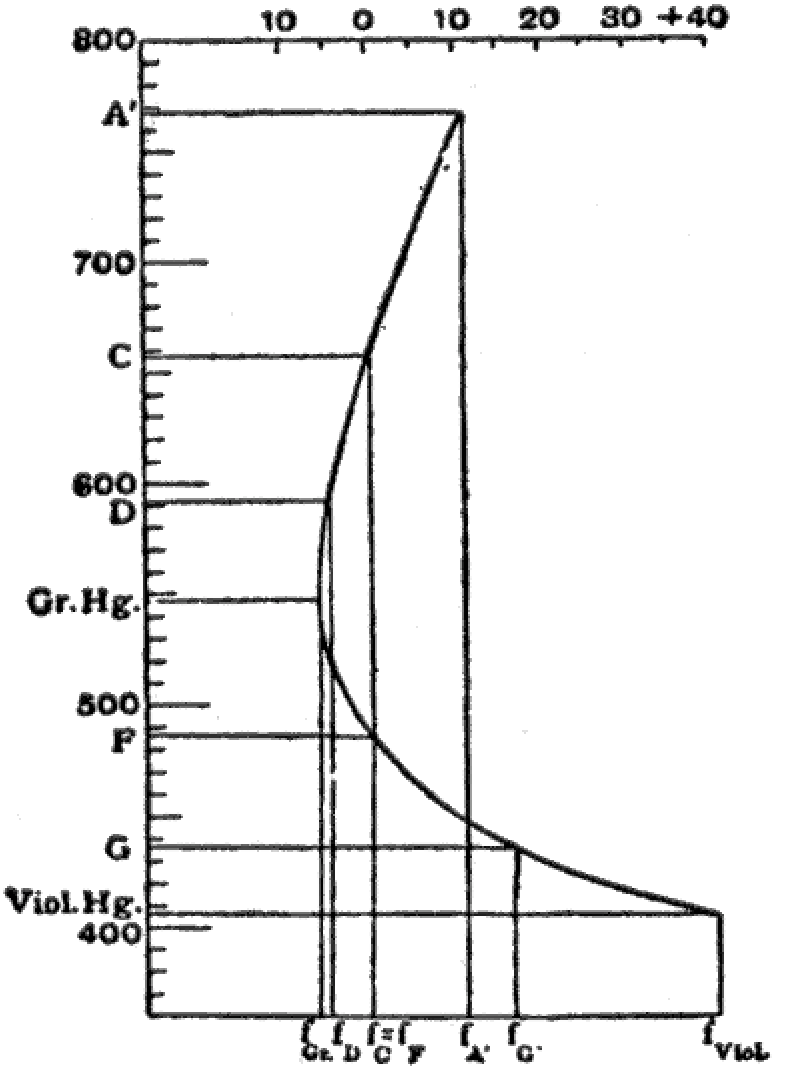
Instead of making df vanish, a certain value can be assigned to it which will produce, by the addition of the two lenses, any desired chromatic deviation, e.g. sufficient to eliminate one present in other parts of the system. If the lenses I. and II. be cemented and have the same refractive index for one color, then its effect for that one color is that of a lens of one piece; by such decomposition of a lens it can be made chromatic or achromatic at will, without altering its spherical effect. If its chromatic effect (df / f) be greater than that of the same lens, this being made of the more dispersive of the two glasses employed, it is termed *hyper-chromatic.*

For two thin lenses separated by a distance D the condition for achromatism is $D = v_1f_1 + v_2f_2$; if $v_1 = v_2$ (e.g. if the lenses be made of the same glass), this reduces to $D = (f_1 + f_2) / 2$, known as the *condition for oculars.*

If a constant of reproduction, for instance the focal length, be made equal for two colors, then it is not the same for other colors, if two different glasses are employed. For example, the condition for achromatism (4) for two thin lenses in contact is fulfilled in only one part of the spectrum, since dn_2 / dn_1 varies within the spectrum. This fact was first ascertained by J. Fraunhofer, who defined the colors by means of the dark lines in the solar spectrum; and showed that the ratio of the dispersion of two glasses varied about 20% from the red to the violet (the variation for glass and water is about 50%). If, therefore, for two colors, a and b, $f_a = f_b = f$, then for a third color, c, the focal length is different; that is, if c lies between a and b, then $f_c < f$, and vice versa; these algebraic results follow from the fact that towards the red the dispersion of the positive crown glass preponderates, towards the violet that of the negative flint. These chromatic errors of systems, which are achromatic for two colors, are called the *secondary spectrum,* and depend upon the aperture and focal length in the same manner as the primary chromatid errors do.

In fig. 11, taken from M. von Rohr's *Theorie und Geschichte des photographischen Objectivs*, the abscissae are focal lengths, and the ordinates wavelengths. The Fraunhofer lines used are shown in the table to the right of the figure.

A'	C	D	Green Hg.	F	G'	Violet Hg.
767.7	656.3	589.3	546.1	486.2	454.1	405.1 nm



and the focal lengths are made equal for the lines C and F. In the neighborhood of 550 nm the tangent to the curve is parallel to the axis of wave-lengths; and the focal length varies least over a fairly large range of color, therefore in this neighborhood the color union is at its best. Moreover, this region of the spectrum is that which appears brightest to the human eye, and consequently this curve of the secondary on spectrum, obtained by making $f_C = f_F$, is, according to the experiments of Sir G. G. Stokes (Proc. Roy. Soc., 1878), the most suitable for visual instruments (*optical achromatism,*). In a similar manner, for systems used in photography, the vertex of the color curve must be placed in the position of the maximum sensibility of the plates; this is generally supposed to be at G'; and to accomplish this the F and violet mercury lines are united. This artifice is specially adopted in objectives for astronomical photography (*pure actinic achromatism*). For ordinary photography, however, there is this disadvantage: the image on the focusing-screen and the correct adjustment of the photographic sensitive plate are not in register; in astronomical photography this difference is constant, but in other kinds it depends on the distance of the objects. On this account the lines D and G' are united for ordinary photographic objectives; the optical as well as the actinic image is chromatically inferior, but both lie in the same place; and consequently the best correction lies in F (this is known as the *actinic correction* or *freedom from chemical focus*).

Should there be in two lenses in contact the same focal lengths for three colours a, b, and c, i.e. $f_a = f_b = f_c = f$, then the relative partial dispersion $(n_c - n_b)(n_a - n_b)$ must be equal for the two kinds of glass employed. This follows by considering equation (4) for the two pairs of colors ac and bc. Until recently no glasses were known with a proportional degree of absorption; but R. Blair (Trans. Edin. Soc., 1791, 3, p. 3), P. Barlow, and F. S. Archer overcame the difficulty by constructing fluid lenses between glass walls. Fraunhofer prepared glasses which reduced the secondary spectrum; but permanent success was only assured on the introduction of the Jena glasses by E. Abbe and O. Schott. In using glasses not having proportional dispersion, the deviation of a third colour can be eliminated by two lenses, if an interval be allowed between them; or by three lenses in contact, which may not all consist of the old glasses. In uniting three colors an *achromatism of a higher order* is derived; there is yet a residual *tertiary spectrum,* but it can always be neglected.

The Gaussian theory is only an approximation; monochromatic or spherical aberrations still occur, which will be different for different colors; and should they be compensated for one color, the image of another color would prove disturbing. The most important is the chromatic difference of aberration of the axis point, which is still present to disturb the image, after par-axial rays of different colors are united by an appropriate combination of glasses. If a collective system be corrected for the axis point for a definite wave-length, then, on account of the greater dispersion in the negative components — the flint glasses, — over-correction will arise for the shorter wavelengths (this being the error of the negative components), and under-correction for the longer wave-lengths (the error of crown glass lenses preponderating in the red). This error was treated by Jean le Rond d'Alembert, and, in special detail, by C. F. Gauss. It increases rapidly with the aperture, and is more important with medium apertures than the secondary spectrum of par-axial rays; consequently, spherical aberration must be eliminated for two colors, and if this be impossible, then it must be eliminated for those particular wave-lengths which are most effectual for the instrument in question (a graphical representation of this error is given in M. von Rohr, *Theorie und Geschichte des photographischen Objectivs*).

The condition for the reproduction of a surface element in the place of a sharply reproduced point — the constant of the sine relationship must also be fulfilled with large apertures for several colors. E. Abbe succeeded in computing microscope objectives free from error of the axis point and satisfying the sine condition for several colors, which therefore, according to his definition, were *aplanatic for several colors*; such systems he termed *apochromatic*. While, however, the magnification of the individual zones is the same, it is not the same for red as for blue; and there is a chromatic difference of magnification. This is produced in the same amount, but in the opposite sense, by the oculars, which Abbe used with these objectives (*compensating oculars*), so that it is eliminated in the image of the whole microscope. The best telescope objectives, and photographic objectives intended for three-color work, are also apochromatic, even if they do not possess quite the same quality of correction as microscope objectives do. The chromatic differences of other errors of reproduction have seldom practical importances.

Chapter 5

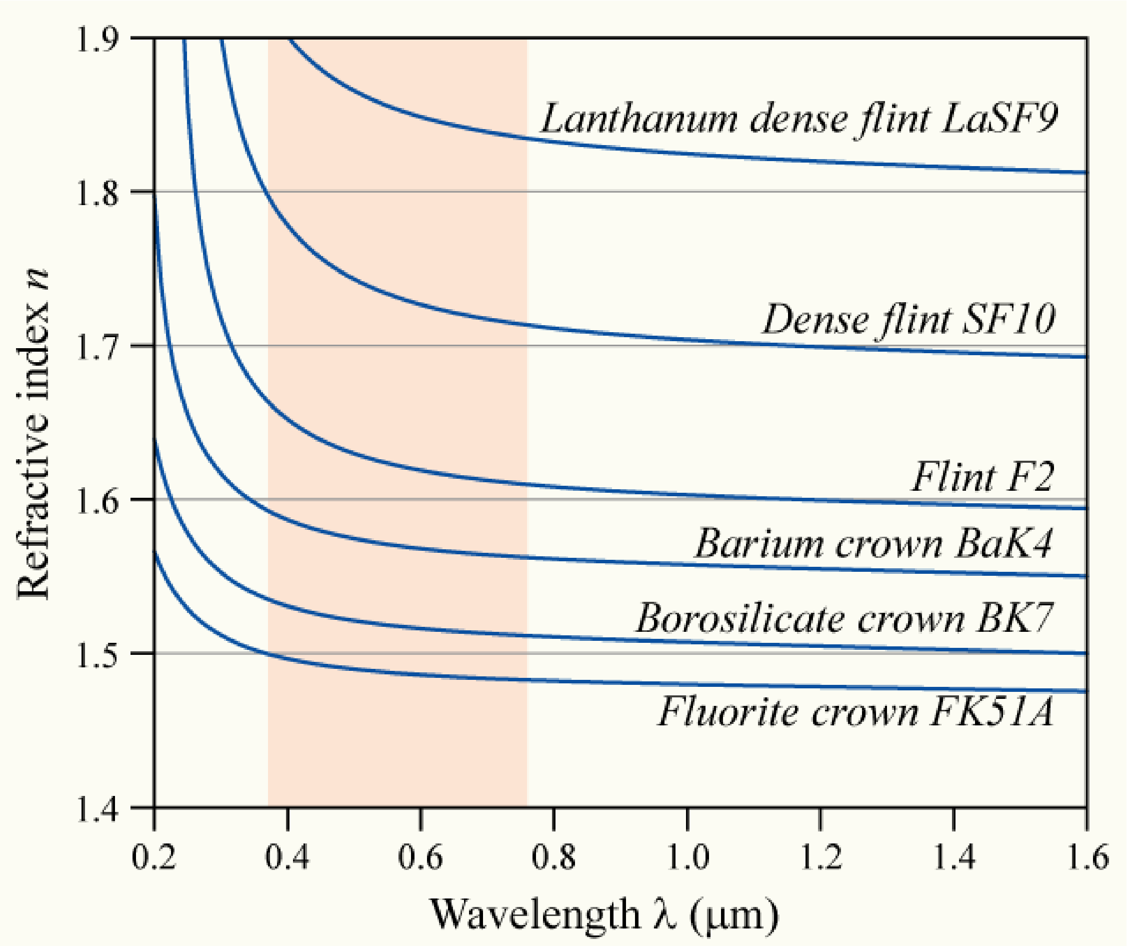
Dispersion (Optics)

In optics, **dispersion** is the phenomenon in which the phase velocity of a wave depends on its frequency, or alternatively when the group velocity depends on the frequency. Media having such a property are termed *dispersive media*. Dispersion is sometimes called ***chromatic*** **dispersion** to emphasize its wavelength-dependent nature, or **group-velocity dispersion** (**GVD**) to emphasize the role of the group velocity.

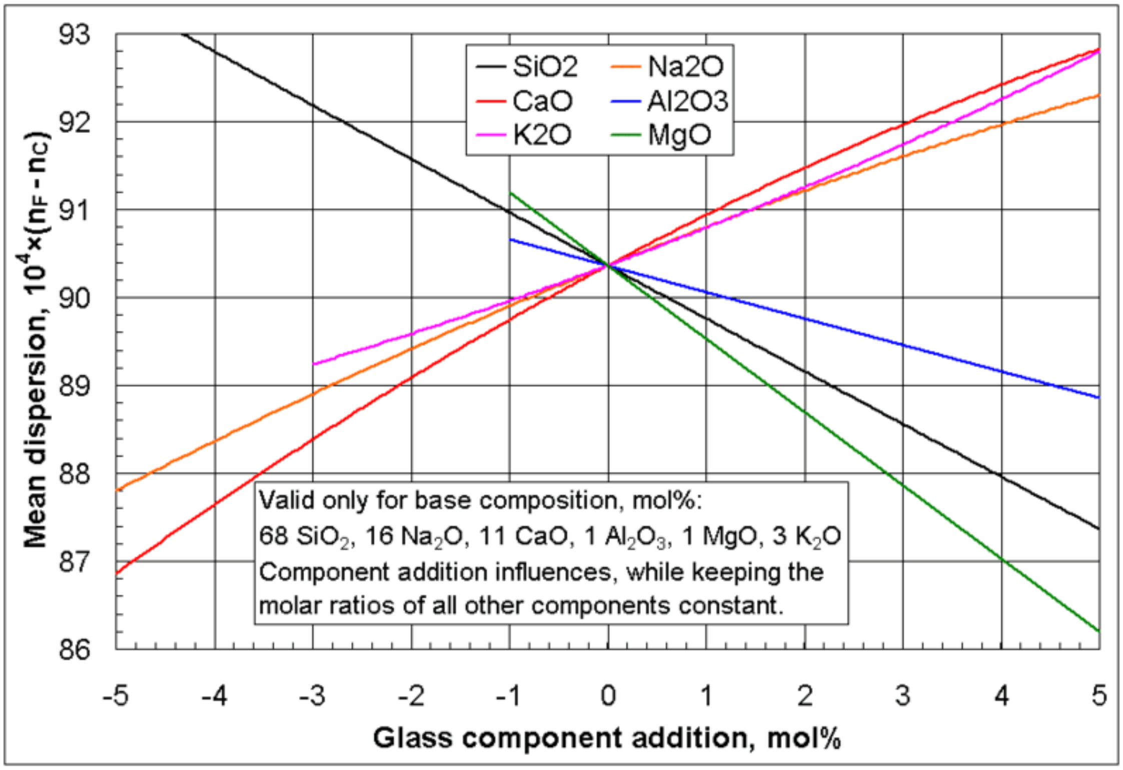
The most familiar example of dispersion is probably a rainbow, in which dispersion causes the spatial separation of a white light into components of different wavelengths (different colors). However, dispersion also has an effect in many other circumstances: for example, GVD causes pulses to spread in optical fibers, degrading signals over long distances; also, a cancellation between group-velocity dispersion and nonlinear effects leads to soliton waves. Dispersion is most often described for light waves, but it may occur for any kind of wave that interacts with a medium or passes through an inhomogeneous geometry (e.g., a waveguide), such as sound waves.

There are generally two sources of dispersion: **material dispersion** and **waveguide dispersion**. Material dispersion comes from a frequency-dependent response of a material to waves. For example, material dispersion leads to undesired chromatic aberration in a lens or the separation of colors in a prism. Waveguide dispersion occurs when the speed of a wave in a waveguide (such as an optical fiber) depends on its frequency for geometric reasons, independent of any frequency dependence of the materials from which it is constructed. More generally, "waveguide" dispersion can occur for waves propagating through any inhomogeneous structure (e.g., a photonic crystal), whether or not the waves are confined to some region. In general, *both* types of dispersion may be present, although they are not strictly additive. Their combination leads to signal degradation in optical fibers for telecommunications, because the varying delay in arrival time between different components of a signal "smears out" the signal in time.

Material dispersion in optics



The variation of refractive index vs. vacuum wavelength for various glasses. The wavelengths of visible light are shaded in red.



Influences of selected glass component additions on the mean dispersion of a specific base glass (n_F valid for λ = 486 nm (blue), n_C valid for λ = 656 nm (red))

Material dispersion can be a desirable or undesirable effect in optical applications. The dispersion of light by glass prisms is used to construct spectrometers and spectroradiometers. Holographic gratings are also used, as they allow more accurate discrimination of wavelengths. However, in lenses, dispersion causes chromatic aberration, an undesired effect that may degrade images in microscopes, telescopes and photographic objectives.

The *phase velocity*, *v*, of a wave in a given uniform medium is given by

$$v = \frac{c}{n}$$

where *c* is the speed of light in a vacuum and *n* is the refractive index of the medium.

In general, the refractive index is some function of the frequency *f* of the light, thus $n = n(f)$, or alternatively, with respect to the wave's wavelength $n = n(\lambda)$. The wavelength dependence of a material's refractive index is usually quantified by an empirical formula, the Cauchy or Sellmeier equations.

Because of the Kramers–Kronig relations, the wavelength dependence of the real part of the refractive index is related to the material absorption, described by the imaginary part of the refractive index (also called the extinction coefficient). In particular, for non-magnetic materials ($\mu = \mu_0$), the susceptibility χ that appears in the Kramers–Kronig relations is the electric susceptibility $\chi_e = n^2 - 1$.

The most commonly seen consequence of dispersion in optics is the separation of white light into a color spectrum by a prism. From Snell's law it can be seen that the angle of refraction of light in a prism depends on the refractive index of the prism material. Since that refractive index varies with wavelength, it follows that the angle that the light is refracted by will also vary with wavelength, causing an angular separation of the colors known as *angular dispersion*.

For visible light, most transparent materials (e.g., glasses) have:

$$1 < n(\lambda_{\text{red}}) < n(\lambda_{\text{yellow}}) < n(\lambda_{\text{blue}}) \ ,$$

or alternatively:

$$\frac{\mathrm{d}n}{\mathrm{d}\lambda} < 0,$$

that is, refractive index *n* decreases with increasing wavelength λ. In this case, the medium is said to have *normal dispersion*. Whereas, if the index increases with increasing wavelength the medium has *anomalous dispersion*.

At the interface of such a material with air or vacuum (index of ~1), Snell's law predicts that light incident at an angle θ to the normal will be refracted at an angle $\arcsin(\sin(\theta)/n)$. Thus, blue light, with a higher refractive index, will be bent more strongly than red light, resulting in the well-known rainbow pattern.

Group and phase velocity

Another consequence of dispersion manifests itself as a temporal effect. The formula $v = c / n$ calculates the *phase velocity* of a wave; this is the velocity at which the *phase* of any one frequency component of the wave will propagate. This is not the same as the *group velocity* of the wave, that is the rate at which changes in amplitude (known as the *envelope* of the wave) will propagate. For a homogeneous medium, the group velocity v_g is related to the phase velocity by (here λ is the wavelength in vacuum, not in the medium):

$$v_g = c\left(n - \lambda\frac{dn}{d\lambda}\right)^{-1} .$$

The group velocity v_g is often thought of as the velocity at which energy or information is conveyed along the wave. In most cases this is true, and the group velocity can be thought of as the *signal velocity* of the waveform. In some unusual circumstances, called cases of anomalous dispersion, the rate of change of the index of refraction with respect to the wavelength changes sign, in which case it is possible for the group velocity to exceed the speed of light ($v_g > c$). Anomalous dispersion occurs, for instance, where the wavelength of the light is close to an absorption resonance of the medium. When the dispersion is anomalous, however, group velocity is no longer an indicator of signal velocity. Instead, a signal travels at the speed of the wavefront, which is c irrespective of the index of refraction. Recently, it has become possible to create gases in which the group velocity is not only larger than the speed of light, but even negative. In these cases, a pulse can appear to exit a medium before it enters. Even in these cases, however, a signal travels at, or less than, the speed of light, as demonstrated by Stenner, et al.

The group velocity itself is usually a function of the wave's frequency. This results in **group velocity dispersion** (GVD), which causes a short pulse of light to spread in time as a result of different frequency components of the pulse travelling at different velocities. GVD is often quantified as the *group delay dispersion parameter* (again, this formula is for a uniform medium only):

$$D = -\frac{\lambda}{c}\,\frac{d^2 n}{d\lambda^2}.$$

If D is less than zero, the medium is said to have *positive dispersion*. If D is greater than zero, the medium has *negative dispersion*. If a light pulse is propagated through a normally dispersive medium, the result is the higher frequency components travel slower than the lower frequency components. The pulse therefore becomes *positively chirped*, or *up-chirped*, increasing in frequency with time. Conversely, if a pulse travels through an anomalously dispersive medium, high frequency components travel faster than the lower ones, and the pulse becomes *negatively chirped*, or *down-chirped*, decreasing in frequency with time.

The result of GVD, whether negative or positive, is ultimately temporal spreading of the pulse. This makes dispersion management extremely important in optical communications systems based on optical fiber, since if dispersion is too high, a group of pulses representing a bit-stream will spread in time and merge together, rendering the bit-stream unintelligible. This limits the length of fiber that a signal can be sent down without regeneration. One possible answer to this problem is to send signals down the optical fibre at a wavelength where the GVD is zero (e.g., around 1.3–1.5 μm in silica fibres), so pulses at this wavelength suffer minimal spreading from dispersion—in practice, however, this approach causes more problems than it solves because zero GVD unacceptably amplifies other nonlinear effects (such as four wave mixing). Another possible option is to use soliton pulses in the regime of anomalous dispersion, a form of optical pulse which uses a nonlinear optical effect to self-maintain its shape—solitons have the practical problem, however, that they require a certain power level to be maintained in the pulse for the nonlinear effect to be of the correct strength. Instead, the solution that is currently used in practice is to perform dispersion compensation, typically by matching the fiber with another fiber of opposite-sign dispersion so that the dispersion effects cancel; such compensation is ultimately limited by nonlinear effects such as self-phase modulation, which interact with dispersion to make it very difficult to undo.

Dispersion control is also important in lasers that produce short pulses. The overall dispersion of the optical resonator is a major factor in determining the duration of the pulses emitted by the laser. A pair of prisms can be arranged to produce net negative dispersion, which can be used to balance the usually positive dispersion of the laser medium. Diffraction gratings can also be used to produce dispersive effects; these are often used in high-power laser amplifier systems. Recently, an alternative to prisms and gratings has been developed: chirped mirrors. These dielectric mirrors are coated so that different wavelengths have different penetration lengths, and therefore different group delays. The coating layers can be tailored to achieve a net negative dispersion.

Dispersion in waveguides

Optical fibers, which are used in telecommunications, are among the most abundant types of waveguides. Dispersion in these fibers is one of the limiting factors that determine how much data can be transported on a single fiber.

The transverse modes for waves confined laterally within a waveguide generally have different speeds (and field patterns) depending upon their frequency (that is, on the relative size of the wave, the wavelength) compared to the size of the waveguide.

In general, for a waveguide mode with an angular frequency ω(β) at a propagation constant β (so that the electromagnetic fields in the propagation direction *(z)* oscillate proportional to $e^{i(\beta z - \omega t)}$), the group-velocity dispersion parameter *D* is defined as:

$$D = -\frac{2\pi c}{\lambda^2}\frac{d^2\beta}{d\omega^2} = \frac{2\pi c}{v_g^2\lambda^2}\frac{dv_g}{d\omega}$$

where $\lambda = 2\pi c / \omega$ is the vacuum wavelength and $v_g = d\omega / d\beta$ is the group velocity. This formula generalizes the one in the previous section for homogeneous media, and includes both waveguide dispersion and material dispersion. The reason for defining the dispersion in this way is that $|D|$ is the (asymptotic) temporal pulse spreading Δt per unit bandwidth $\Delta\lambda$ per unit distance travelled, commonly reported in ps / nm km for optical fibers.

A similar effect due to a somewhat different phenomenon is modal dispersion, caused by a waveguide having multiple modes at a given frequency, each with a different speed. A special case of this is polarization mode dispersion (PMD), which comes from a superposition of two modes that travel at different speeds due to random imperfections that break the symmetry of the waveguide.

Higher-order dispersion over broad bandwidths

When a broad range of frequencies (a broad bandwidth) is present in a single wavepacket, such as in an ultrashort pulse or a chirped pulse or other forms of spread spectrum transmission, it may not be accurate to approximate the dispersion by a constant over the entire bandwidth, and more complex calculations are required to compute effects such as pulse spreading.

In particular, the dispersion parameter *D* defined above is obtained from only one derivative of the group velocity. Higher derivatives are known as *higher-order dispersion*. These terms are simply a Taylor series expansion of the dispersion relation β(ω) of the medium or waveguide around some particular frequency. Their effects can be computed via numerical evaluation of Fourier transforms of the waveform, via integration of higher-order slowly varying envelope approximations, by a split-step method (which can use the exact dispersion relation rather than a Taylor series), or by direct simulation of the full Maxwell's equations rather than an approximate envelope equation.

Dispersion in gemology

In the technical terminology of gemology, *dispersion* is the difference in the refractive index of a material at the B and G Fraunhofer wavelengths of 686.7 nm and 430.8 nm and is meant to express the degree to which a prism cut from the gemstone shows "fire", or color. Dispersion is a material property. Fire depends on the dispersion, the cut angles, the lighting environment, the refractive index, and the viewer.

Dispersion in imaging

In photographic and microscopic lenses, dispersion causes chromatic aberration, distorting the image, and various techniques have been developed to counteract it such as the use of multielement lenses with glasses with different dispersion characteristics: the net effect is to recombine (at least approximately) all colors.

Dispersion in pulsar timing

Pulsars are spinning neutron stars that emit pulses at very regular intervals ranging from milliseconds to seconds. Astronomers believe that the pulses are emitted simultaneously over a wide range of frequencies. However, as observed on Earth, the components of each pulse emitted at higher radio frequencies arrive before those emitted at lower frequencies. This dispersion occurs because of the ionised component of the interstellar medium, which makes the group velocity frequency dependent. The extra delay added at a frequency ν is

$$t = k_{\mathrm{DM}} \times \left(\frac{\mathrm{DM}}{\nu^2}\right)$$

where the dispersion constant k_{DM} is given by

$$k_{\mathrm{DM}} = \frac{e^2}{2\pi m_{\mathrm{e}} c} \simeq 4.149 \mathrm{GHz}^2 \mathrm{pc}^{-1} \mathrm{cm}^3 \mathrm{ms}$$
,

and the dispersion measure *DM* is the free electron column density (total electron content) n_e integrated along the path traveled by the photon from the pulsar to the Earth, and is given by

$$\mathrm{DM} = \int_0^d n_e \, dl$$

with units of parsecs per cubic centimetre ($1 \mathrm{pc/cm}^3 = 30.857 \times 10^{21}\ \mathrm{m}^{-2}$).

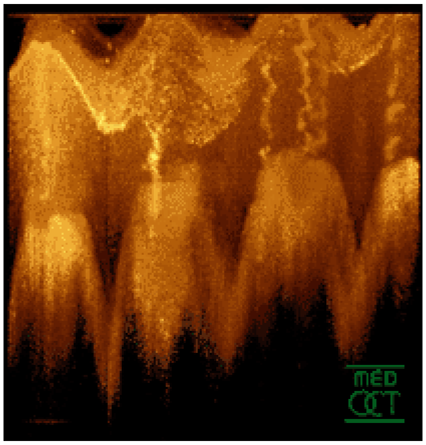
Typically for astronometric observations, this delay cannot be measured directly, since the emission time is unknown. What *can* be measured is the difference in arrival times at two different frequencies. The delay ΔT between a high frequency ν_{hi} and a low frequency ν_{lo} component of a pulse will be

$$\Delta t = k_{\mathrm{DM}} \times \mathrm{DM} \times \left(\frac{1}{\nu_{\mathrm{lo}}^2} - \frac{1}{\nu_{\mathrm{hi}}^2}\right)$$

Re-writing the above equation in terms of *DM* allows one to determine the *DM* by measuring pulse arrival times at multiple frequencies. This in turn can be used to study the interstellar medium, as well as allow for observations of pulsars at different frequencies to be combined.

Chapter 6

Optical Coherence Tomography



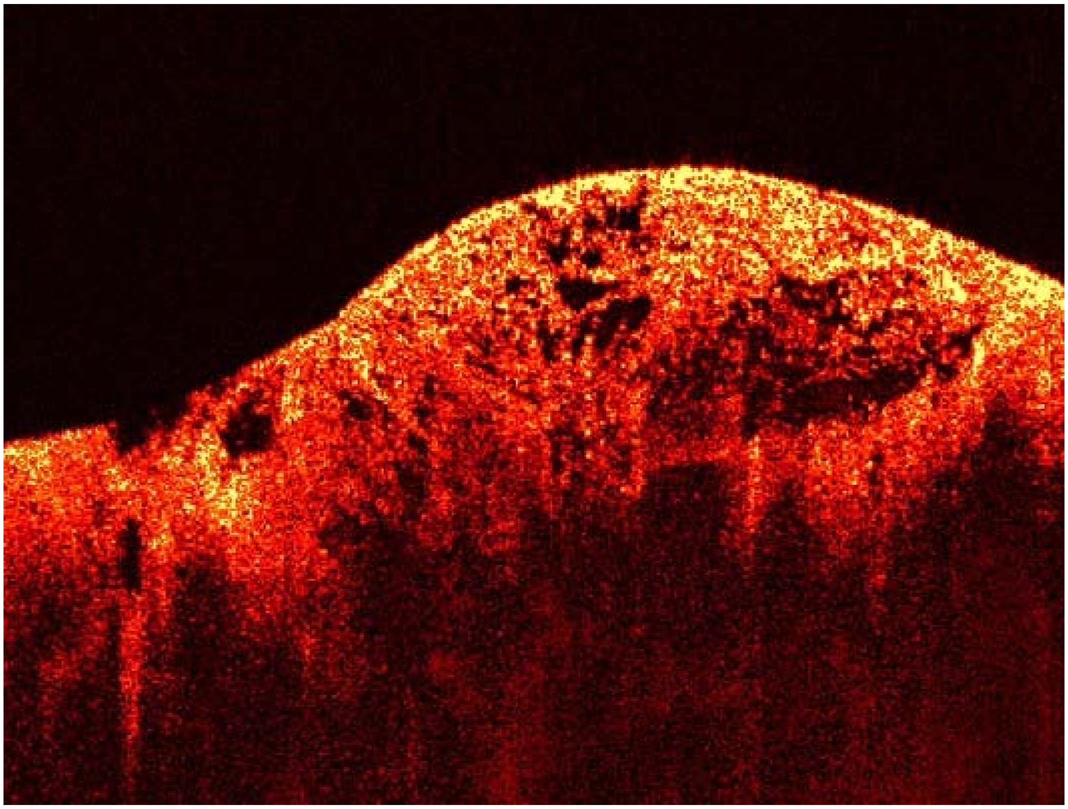
Optical coherence tomography tomogram of a fingertip.

Optical coherence tomography (OCT) is an optical signal acquisition and processing method. It captures micrometer-resolution, three-dimensional images from within optical scattering media (e.g., biological tissue). Optical coherence tomography is an interferometric technique, typically employing near-infrared light. The use of relatively long wavelength light allows it to penetrate into the scattering medium. Confocal microscopy, another similar technique, typically penetrates less deeply into the sample.

Depending on the properties of the light source (superluminescent diodes and ultrashort pulsed lasers have been employed), Optical coherence tomography has achieved sub-micrometer resolution (with very wide-spectrum sources emitting over a ~100 nm wavelength range)

Optical coherence tomography is one of a class of optical tomographic techniques. A relatively recent implementation of optical coherence tomography, frequency-domain optical coherence tomography, provides advantages in signal-to-noise ratio, permitting faster signal acquisition. Commercially available optical coherence tomography systems are employed in diverse applications, including art conservation and diagnostic medicine, notably in ophthalmology where it can be used to obtain detailed images from within the retina. Recently it has also begun to be used in interventional cardiology to help diagnose coronary artery disease

Introduction



Optical Coherence Tomography (OCT) image of a sarcoma

Starting from white-light interferometry for *in vivo* ocular eye measurements imaging of biological tissue, especially of the human eye, was investigated by multiple groups worldwide. A first two-dimensional *in vivo* depiction of a human eye fundus along a horizontal meridian based on white light interferometric depth scans was presented at the ICO-15 SAT conference in 1990. Further developed in 1990 by Naohiro Tanno , then a professor at Yamagata University, and in particular since 1991 by Huang et al., optical coherence tomography (OCT) with micrometer resolution and cross-sectional imaging capabilities has become a prominent biomedical tissue-imaging technique; it is particularly suited to ophthalmic applications and other tissue imaging requiring micrometer resolution and millimeter penetration depth. First *in vivo* OCT images – displaying retinal structures – were published in 1993. OCT has also been used for various art conservation projects, where it is used to analyze different layers in a painting. OCT has critical advantages over other medical imaging systems. Medical ultrasonography, magnetic resonance imaging (MRI) and confocal microscopy are not suited to morphological tissue imaging: the first two have poor resolution; the last lacks millimeter penetration depth.

OCT bases itself upon low coherence interferometry. In conventional interferometry with long coherence length (laser interferometry), interference of light occurs over a distance of meters. In OCT, this interference is shortened to a distance of micrometers, thanks to the use of broadband light sources (sources that can emit light over a broad range of frequencies). Light with broad bandwidths can be generated by using superluminescent diodes (superbright LEDs) or lasers with extremely short pulses (femtosecond lasers). White light is also a broadband source with lower power.

Light in an OCT system is broken into two arms—a sample arm (containing the item of interest) and a reference arm (usually a mirror). The combination of reflected light from the sample arm and reference light from the reference arm gives rise to an interference pattern, but only if light from both arms have travelled the "same" optical distance ("same" meaning a difference of less than a coherence length). By scanning the mirror in the reference arm, a reflectivity profile of the sample can be obtained (this is time domain OCT). Areas of the sample that reflect back a lot of light will create greater interference than areas that don't. Any light that is outside the short coherence length will not interfere. This reflectivity profile, called an A-scan, contains information about the spatial dimensions and location of structures within the item of interest. A cross-sectional tomograph (B-scan) may be achieved by laterally combining a series of these axial depth scans (A-scan). En face imaging (C-scan) at an acquired depth is possible depending on the imaging engine used.

Laypersons explanation

Optical Coherence Tomography, or ‘OCT’, is a technique for obtaining sub-surface images of translucent or opaque materials at a resolution equivalent to a low-power microscope. It is effectively ‘optical ultrasound’, imaging reflections from within tissue to provide cross-sectional images.

OCT is attracting interest among the medical community, because it provides tissue morphology imagery at much higher resolution (better than 10 µm) than other imaging modalities such as MRI or ultrasound.

The key benefits of OCT are:

- Live sub-surface images at near-microscopic resolution
- Instant, direct imaging of tissue morphology
- No preparation of the sample or subject
- No ionizing radiation

OCT delivers high resolution because it is based on light, rather than sound or radio frequency. An optical beam is directed at the tissue, and a small portion of this light that reflects from sub-surface features is collected. Note that most light is not reflected but, rather, scatters. The scattered light has lost its original direction and does not contribute to forming an image but rather contributes to *glare*. The glare of scattered light causes optically scattering materials (e.g., biological tissue, candle wax, or certain plastics) to appear opaque or translucent even while they do not strongly absorb light (as can be ascertained through a simple experiment — e.g., shining a red laser pointer through one's finger). Using the OCT technique, scattered light can be filtered out, completely removing the glare. Even the very tiny proportion of reflected light that is not scattered can then be detected and used to form the image in, e.g., a scanning OCT system employing a microscope.

The physics principle allowing the filtering of scattered light is optical coherence. *Only* the reflected (non-scattered) light is coherent (i.e., retains the optical phase that causes light rays to propagate in one or another direction). In the OCT instrument, an optical interferometer is used in such a manner as to detect *only* coherent light. Essentially, the interferometer strips off scattered light from the reflected light needed to generate an image. In the process depth and intensity of light reflected from a sub-surface feature is obtained. A three-dimensional image can be built up by scanning, as in a sonar or radar system.

Within the range of noninvasive three-dimensional imaging techniques that have been introduced to the medical research community, OCT as an echo technique is similar to ultrasound imaging. Other medical imaging techniques such as computerized axial tomography, magnetic resonance imaging, or positron emission tomography do not utilize the echo-location principle.

The technique is limited to imaging 1 to 2 mm below the surface in biological tissue, because at greater depths the proportion of light that escapes without scattering is too small to be detected. No special preparation of a biological specimen is required, and images can be obtained ‘non-contact’ or through a transparent window or membrane. It is also important to note that the laser output from the instruments is low – eye-safe near-infra-red light is used – and no damage to the sample is therefore likely.

Theory

The principle OCT is white light or low coherence interferometry. The optical setup typically consists of an interferometer (Fig. 1, typically Michelson type) with a low coherence, broad bandwidth light source. Light is split into and recombined from reference and sample arm, respectively.

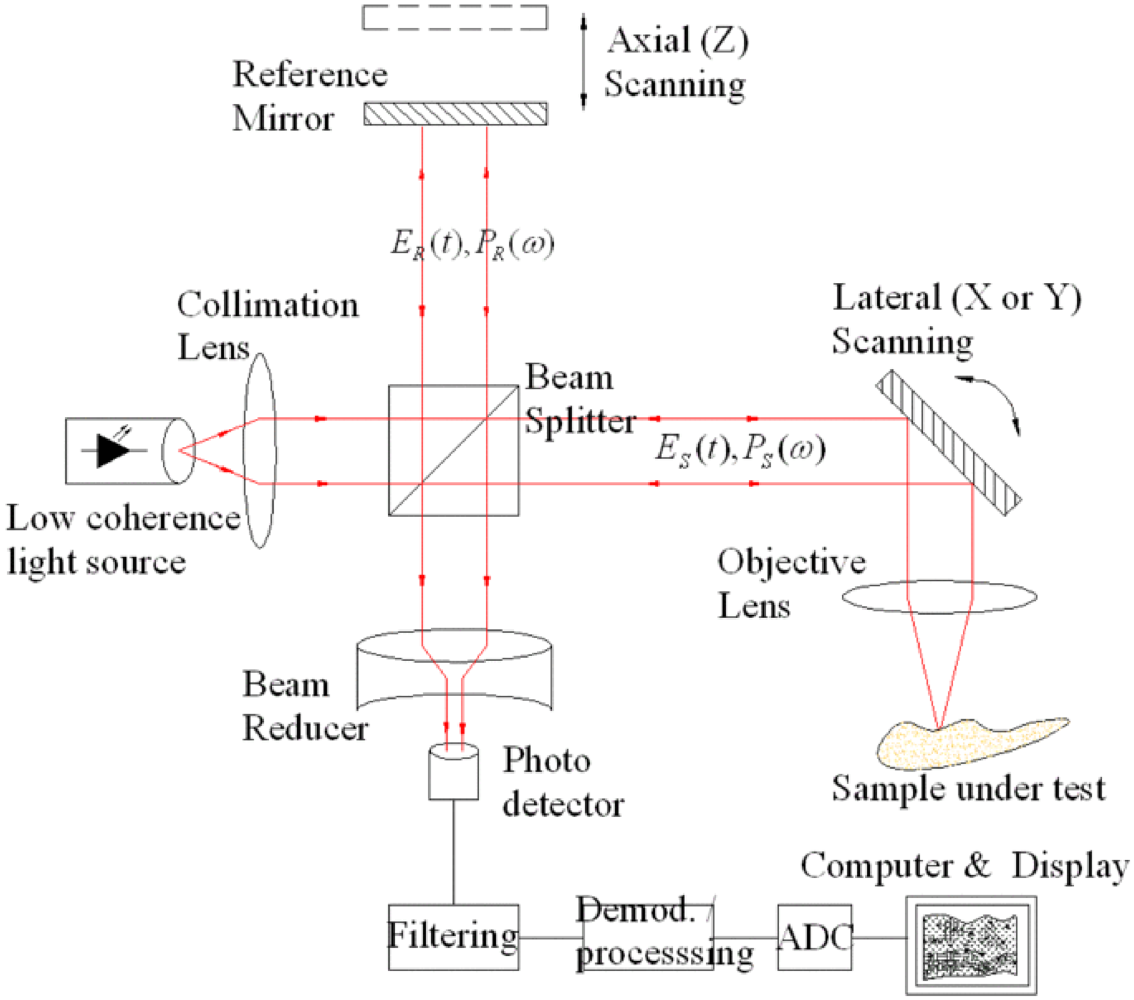


Fig. 2 Typical optical setup of single point OCT. Scanning the light beam on the sample enables non-invasive cross-sectional imaging up to 3 mm in depth with micrometer resolution.

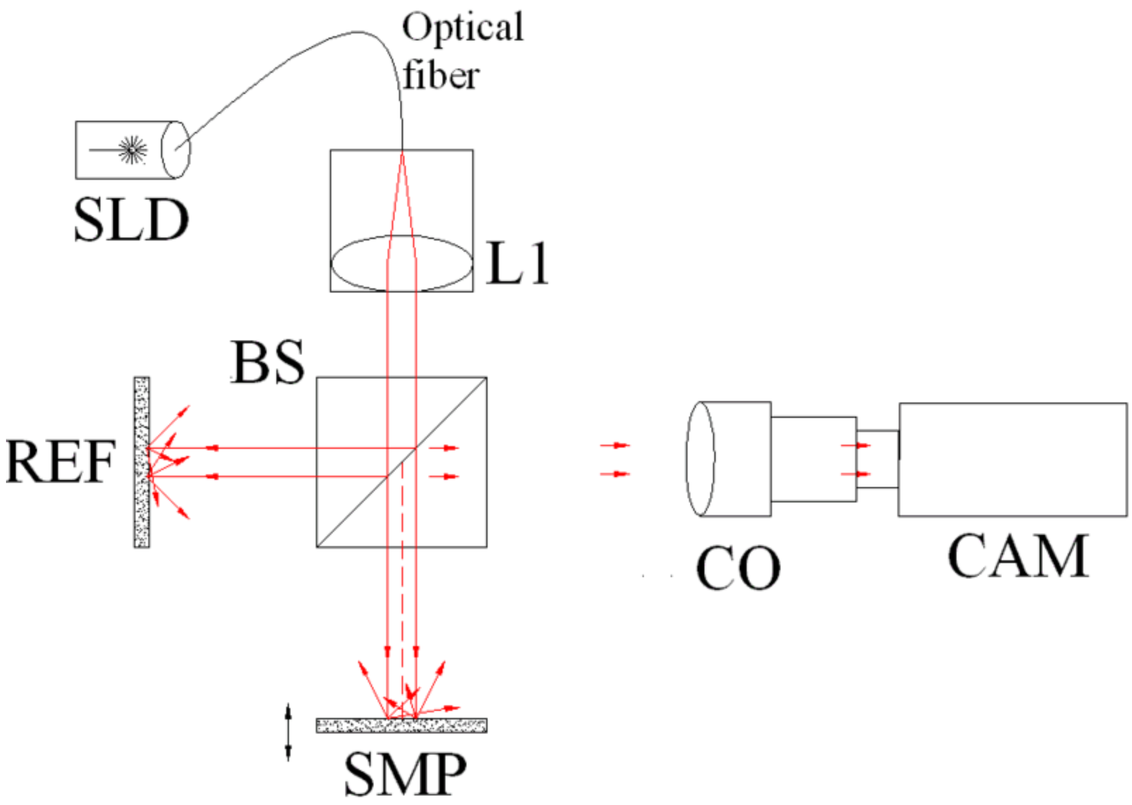


Fig. 1 Full-field OCT optical setup. Components include: super-luminescent diode (SLD), convex lens (L1), 50/50 beamsplitter (BS), camera objective (CO), CMOS-DSP camera (CAM), reference (REF) and sample (SMP). The camera functions as a two-dimensional detector array, and with the OCT technique facilitating scanning in depth, a non-invasive three dimensional imaging device is achieved.

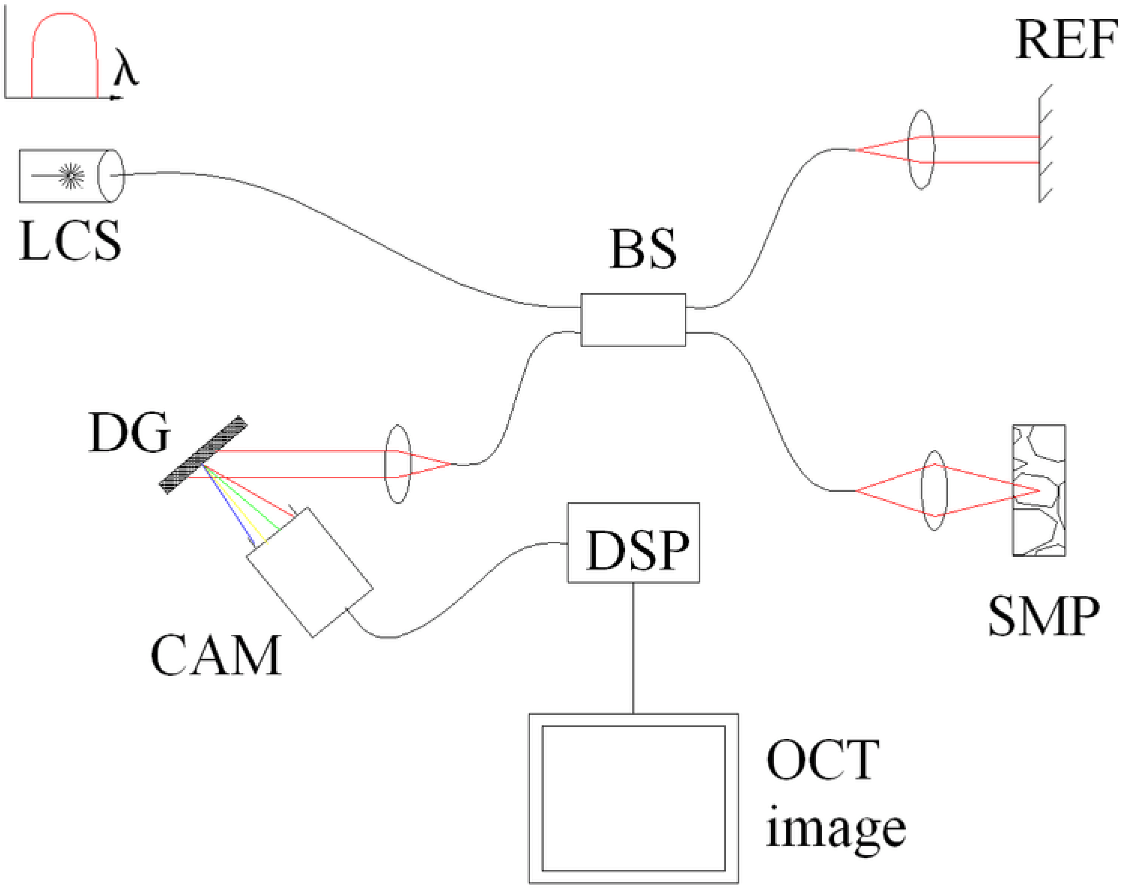


Fig. 4 Spectral discrimination by fourier-domain OCT. Components include: low coherence source (LCS), beamsplitter (BS), reference mirror (REF), sample (SMP), diffraction grating (DG) and full-field detector (CAM) act as a spectrometer, and digital signal processing (DSP)

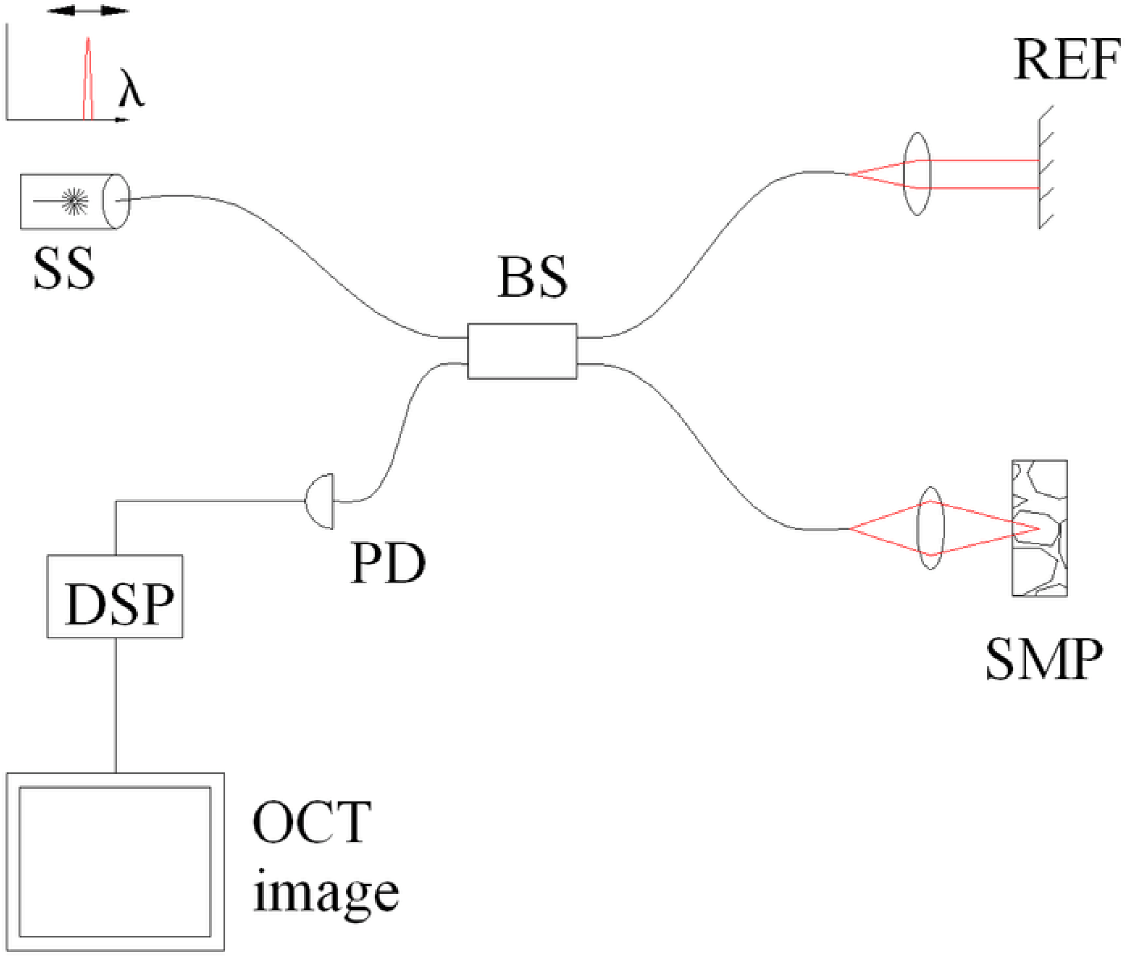


Fig. 3 Spectral discrimination by swept-source OCT. Components include: swept source or tunable laser (SS), beamsplitter (BS), reference mirror (REF), sample (SMP), photodetector (PD), digital signal processing (DSP)

Time domain OCT

In time domain OCT the pathlength of the reference arm is translated longitudinally in time. A property of low coherence interferometry is that interference, i.e. the series of dark and bright fringes, is only achieved when the path difference lies within the coherence length of the light source. This interference is called auto correlation in a symmetric interferometer (both arms have the same reflectivity), or cross-correlation in the common case. The envelope of this modulation changes as pathlength difference is varied, where the peak of the envelope corresponds to pathlength matching.

The interference of two partially coherent light beams can be expressed in terms of the source intensity, I_S, as

$$I = k_1 I_S + k_2 I_S + 2\sqrt{(k_1 I_S) \cdot (k_2 I_S)} \cdot Re\left[\gamma\left(\tau\right)\right] \qquad (1)$$

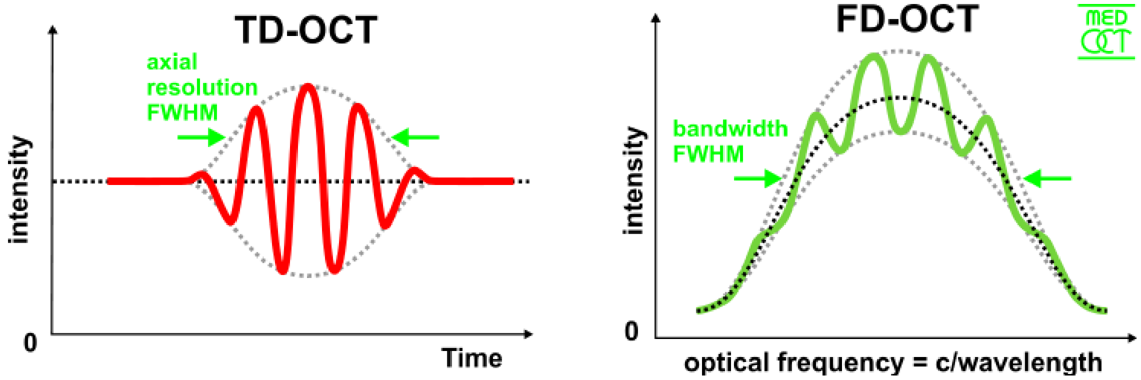
where $k_1 + k_2 < 1$ represents the interferometer beam splitting ratio, and γ(τ) is called the complex degree of coherence, i.e. the interference envelope and carrier dependent on reference arm scan or time delay τ, and whose recovery of interest in OCT. Due to the coherence gating effect of OCT the complex degree of coherence is represented as a Gaussian function expressed as

$$\gamma\left(\tau\right) = \exp\left[-\left(\frac{\pi\Delta\nu\tau}{2\sqrt{\ln 2}}\right)^2\right]\cdot\exp\left(-j2\pi\nu_0\tau\right) \qquad (2)$$

where Δν represents the spectral width of the source in the optical frequency domain, and ν_0 is the centre optical frequency of the source. In equation (2), the Gaussian envelope is amplitude modulated by an optical carrier. The peak of this envelope represents the location of sample under test microstructure, with an amplitude dependent on the reflectivity of the surface. The optical carrier is due to the Doppler effect resulting from scanning one arm of the interferometer, and the frequency of this modulation is controlled by the speed of scanning. Therefore translating one arm of the interferometer has two functions; depth scanning and a Doppler-shifted optical carrier are accomplished by pathlength variation. In OCT, the Doppler-shifted optical carrier has a frequency expressed as

$$f_{Dopp} = \frac{2\cdot\nu_0\cdot v_s}{c} \qquad (3)$$

where ν_0 is the central optical frequency of the source, v_s is the scanning velocity of the pathlength variation, and c is the speed of light.



interference signals in TD vs. FD-OCT

The axial and lateral resolutions of OCT are decoupled from one another; the former being an equivalent to the coherence length of the light source and the latter being a function of the optics. The coherence length of a source and hence the axial resolution of OCT is defined as

$$l_c = \frac{2\ln 2}{\pi} \cdot \frac{\lambda_0^2}{\Delta\lambda}$$
$$\approx 0.44 \cdot \frac{\lambda_0^2}{\Delta\lambda} \qquad (4)$$

Frequency domain OCT (FD-OCT)

In frequency domain OCT the broadband interference is acquired with spectrally separated detectors (either by encoding the optical frequency in time with a spectrally scanning source or with a dispersive detector, like a grating and a linear detector array). Due to the Fourier relation (Wiener-Khintchine theorem between the auto correlation and the spectral power density) the depth scan can be immediately calculated by a Fourier-transform from the acquired spectra, without movement of the reference arm. This feature improves imaging speed dramatically, while the reduced losses during a single scan improve the signal to noise proportional to the number of detection elements. The parallel detection at multiple wavelength ranges limits the scanning range, while the full spectral bandwidth sets the axial resolution.

Spatially encoded frequency domain OCT (spectral domain or Fourier domain OCT)

SEFD-OCT extracts spectral information by distributing different optical frequencies onto a detector stripe (line-array CCD or CMOS) via a dispersive element (see Fig. 4). Thereby the information of the full depth scan can be acquired within a single exposure. However, the large signal to noise advantage of FD-OCT is reduced due the lower dynamic range of stripe detectors in respect to single photosensitive diodes, resulting in an SNR (signal to noise ratio) advantage of ~10 dB at much higher speeds. This is not much of a problem when working at 1300 nm, however, since dynamic range is not a serious problem at this wavelength range.

The drawbacks of this technology are found in a strong fall-off of the SNR, which is proportional to the distance from the zero delay and a sinc-type reduction of the depth dependent sensitivity because of limited detection linewidth. (One pixel detects a quasi-rectangular portion of an optical frequency range instead of a single frequency, the Fourier-transform leads to the sinc(z) behavior). Additionally the dispersive elements in the spectroscopic detector usually do not distribute the light equally spaced in frequency on the detector, but mostly have an inverse dependence. Therefore the signal has to be resampled before processing, which can not take care of the difference in local (pixelwise) bandwidth, which results in further reduction of the signal quality. However, the fall-off is not a serious problem with the development of new generation CCD or photodiode array with a larger number of pixels.

Synthetic array heterodyne detection offers another approach to this problem without the need for high dispersion.

Time encoded frequency domain OCT (also swept source OCT)

TEFD-OCT tries to combine some of the advantages of standard TD and SEFD-OCT. Here the spectral components are not encoded by spatial separation, but they are encoded in time. The spectrum either filtered or generated in single successive frequency steps and reconstructed before Fourier-transformation. By accommodation of a frequency scanning light source (i.e. frequency scanning laser) the optical setup (see Fig. 5) becomes simpler than SEFD, but the problem of scanning is essentially translated from the TD-OCT reference-arm into the TEFD-OCT light source. Here the advantage lies in the proven high SNR detection technology, while swept laser sources achieve very small instantaneous bandwidths (=linewidth) at very high frequencies (20–200 kHz). Drawbacks are the nonlinearities in the wavelength, especially at high scanning frequencies. The broadening of the linewidth at high frequencies and a high sensitivity to movements of the scanning geometry or the sample (below the range of nanometers within successive frequency steps).

Scanning schemes

Focusing the light beam to a point on the surface of the sample under test, and recombining the reflected light with the reference will yield an interferogram with sample information corresponding to a single A-scan (Z axis only). Scanning of the sample can be accomplished by either scanning the light on the sample, or by moving the sample under test. A linear scan will yield a two-dimensional data set corresponding to a cross-sectional image (X-Z axes scan), whereas an area scan achieves a three-dimensional data set corresponding to a volumetric image (X-Y-Z axes scan), also called full-field OCT.

Single point (confocal) OCT

Systems based on single point, or flying-spot time domain OCT, must scan the sample in two lateral dimensions and reconstruct a three-dimensional image using depth information obtained by coherence-gating through an axially scanning reference arm (Fig. 2). Two-dimensional lateral scanning has been electromechanically implemented by moving the sample using a translation stage, and using a novel micro-electro-mechanical system scanner.

Parallel (or full field) OCT

Parallel OCT using a charge-coupled device (CCD) camera has been used in which the sample is full-field illuminated and en face imaged with the CCD, hence eliminating the electromechanical lateral scan. By stepping the reference mirror and recording successive *en face* images a three-dimensional representation can be reconstructed. Three-dimensional OCT using a CCD camera was demonstrated in a phase-stepped technique, using geometric phase-shifting with a Linnik interferometer, utilising a pair of CCDs and heterodyne detection, and in a Linnik interferometer with an oscillating reference mirror and axial translation stage. Central to the CCD approach is the necessity for either very fast CCDs or carrier generation separate to the stepping reference mirror to track the high frequency OCT carrier.

Smart detector array for parallel TD-OCT

A two-dimensional smart detector array, fabricated using a 2 µm complementary metal-oxide-semiconductor (CMOS) process, was used to demonstrate full-field OCT. Featuring an uncomplicated optical setup (Fig. 3), each pixel of the 58x58 pixel smart detector array acted as an individual photodiode and included its own hardware demodulation circuitry.

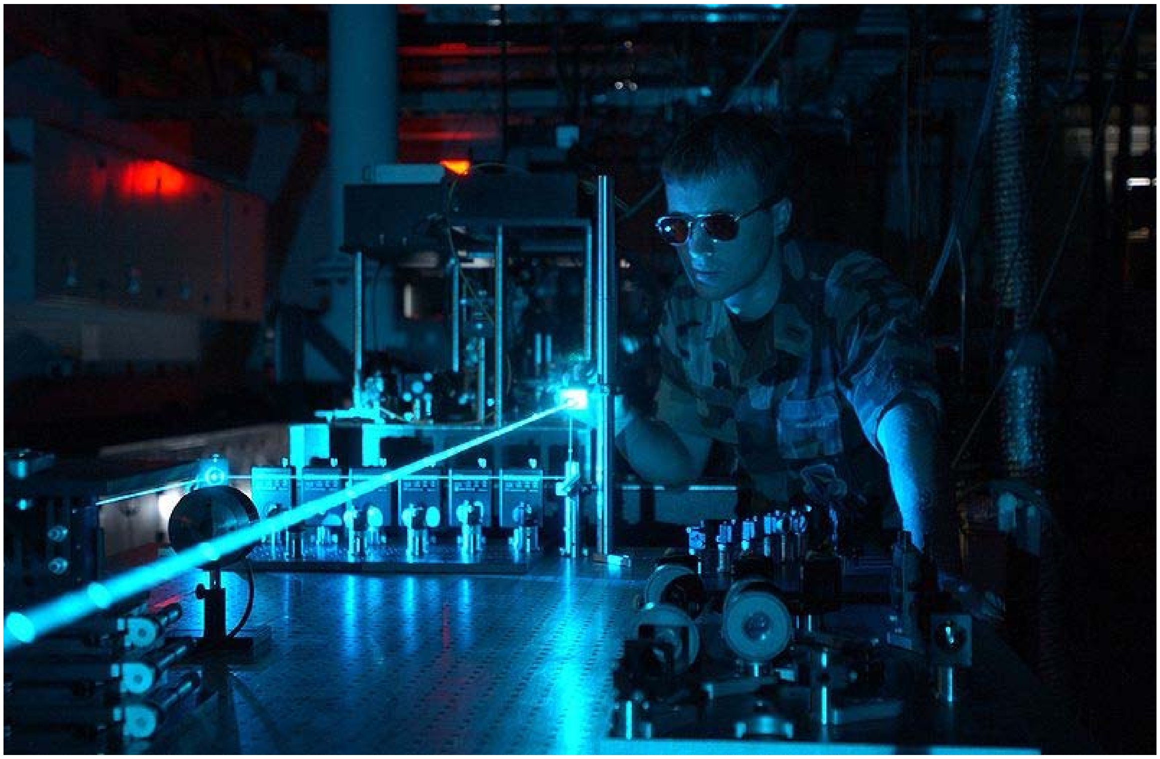
Selected applications

Optical coherence tomography is an established medical imaging technique. It is widely used, for example, to obtain high-resolution images of the retina and the anterior segment of the eye, which can, for example, provide a straightforward method of assessing axonal integrity in multiple sclerosis. Researchers are also seeking to develop a method that uses frequency domain OCT to image coronary arteries in order to detect vulnerable lipid-rich plaques.

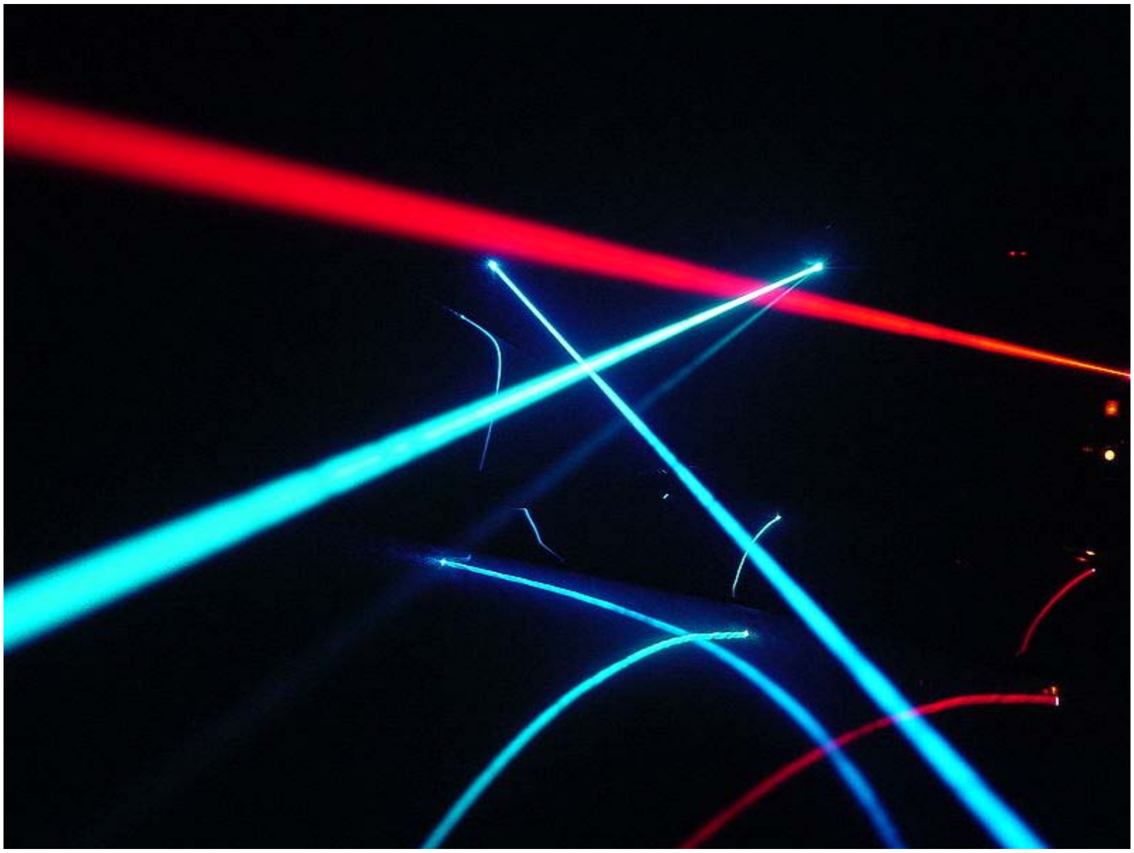
Optical coherence tomography is also applicable and increasingly used in industrial applications, such as Non Destructive Testing(NDT), material thickness measurements, surface roughness characterization, surface and cross-section imaging , and volume loss measurements. OCT systems with feedback can be used to control manufacturing processes. With high speed data acquisition and sub-micron resolution, OCT is adaptable to perform both inline and off-line. Fiber-based OCT systems are particularly adaptable to industrial environments. These can access and scan interiors of hard-to-reach spaces , and are able to operate in hostile environments - whether radioactive, cryogenic or very hot .

Chapter 7

Laser



United States Air Force laser experiment



Laser beams in fog, reflected on a car windshield

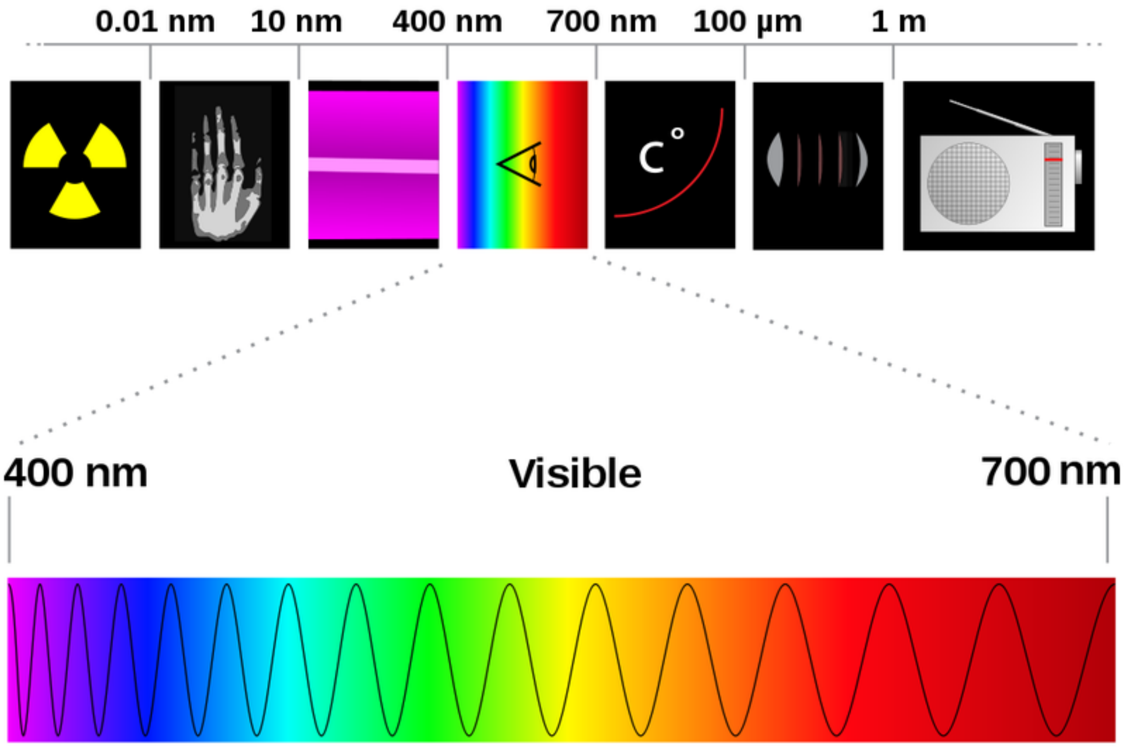
A **laser** is a device that emits light (electromagnetic radiation) through a process of optical amplification based on the stimulated emission of photons. The term "laser" originated as an acronym for *Light Amplification by Stimulated Emission of Radiation*. The emitted laser light is notable for its high degree of spatial and temporal coherence, unattainable using other technologies.

Spatial coherence typically is expressed through the output being a narrow beam which is diffraction-limited, often a so-called "pencil beam." Laser beams can be focused to very tiny spots, achieving a very high irradiance. Or they can be launched into a beam of very low divergence in order to concentrate their power at a large distance.

Temporal (or longitudinal) coherence implies a polarized wave at a single frequency whose phase is correlated over a relatively large distance (the coherence length) along the beam. A beam produced by a thermal or other incoherent light source has an instantaneous amplitude and phase which vary randomly with respect to time and position, and thus a very short coherence length.

Most so-called "single wavelength" lasers actually produce radiation in several *modes* having slightly different frequencies (wavelengths), often not in a single polarization. And although temporal coherence implies monochromaticity, there are even lasers that emit a broad spectrum of light, or emit different wavelengths of light simultaneously. There are some lasers which are not single spatial mode and consequently their light beams diverge more than required by the diffraction limit. However all such devices are classified as "lasers" based on their method of producing that light: stimulated emission. Lasers are employed in applications where light of the required spatial or temporal coherence could not be produced using simpler technologies.

Terminology



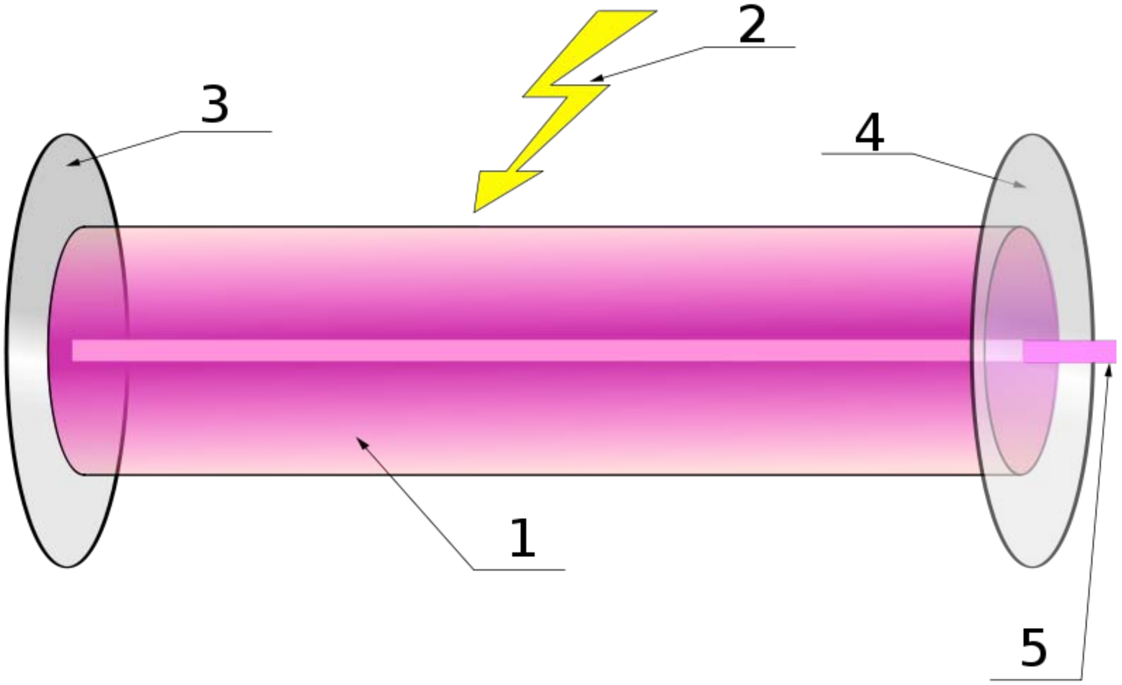
From left to right: gamma rays, X-rays, ultraviolet rays, visible spectrum, infrared, microwaves, radio waves. Bottom: enlargement of visible spectrum from violet (400nm) to red (700nm).

The word *laser* started as an acronym for "light amplification by stimulated emission of radiation"; in modern usage "light" broadly denotes electromagnetic radiation of any frequency, not only visible light, hence *infrared laser*, *ultraviolet laser*, *X-ray laser*, and so on. Because the microwave predecessor of the laser, the maser, was developed first, devices of this sort operating at microwave and radio frequencies are referred to as "masers" rather than "microwave lasers" or "radio lasers". In the early technical literature, especially at Bell Telephone Laboratories, the laser was called an **optical maser**; this term is now obsolete.

A laser which produces light by itself is technically an optical oscillator rather than an optical amplifier as suggested by the acronym. It has been humorously noted that the acronym LOSER, for "light oscillation by stimulated emission of radiation," would have been more correct. With the widespread use of the original acronym as a common noun, actual optical amplifiers have come to be referred to as "laser amplifiers", notwithstanding the apparent redundancy in that designation.

The back-formed verb *to lase* is frequently used in the field, meaning "to produce laser light," especially in reference to the gain medium of a laser; when a laser is operating it is said to be "lasing." Further use of the words *laser* and *maser* in an extended sense, not referring to laser technology or devices, can be seen in usages such as *astrophysical maser* and *atom laser*.

Design



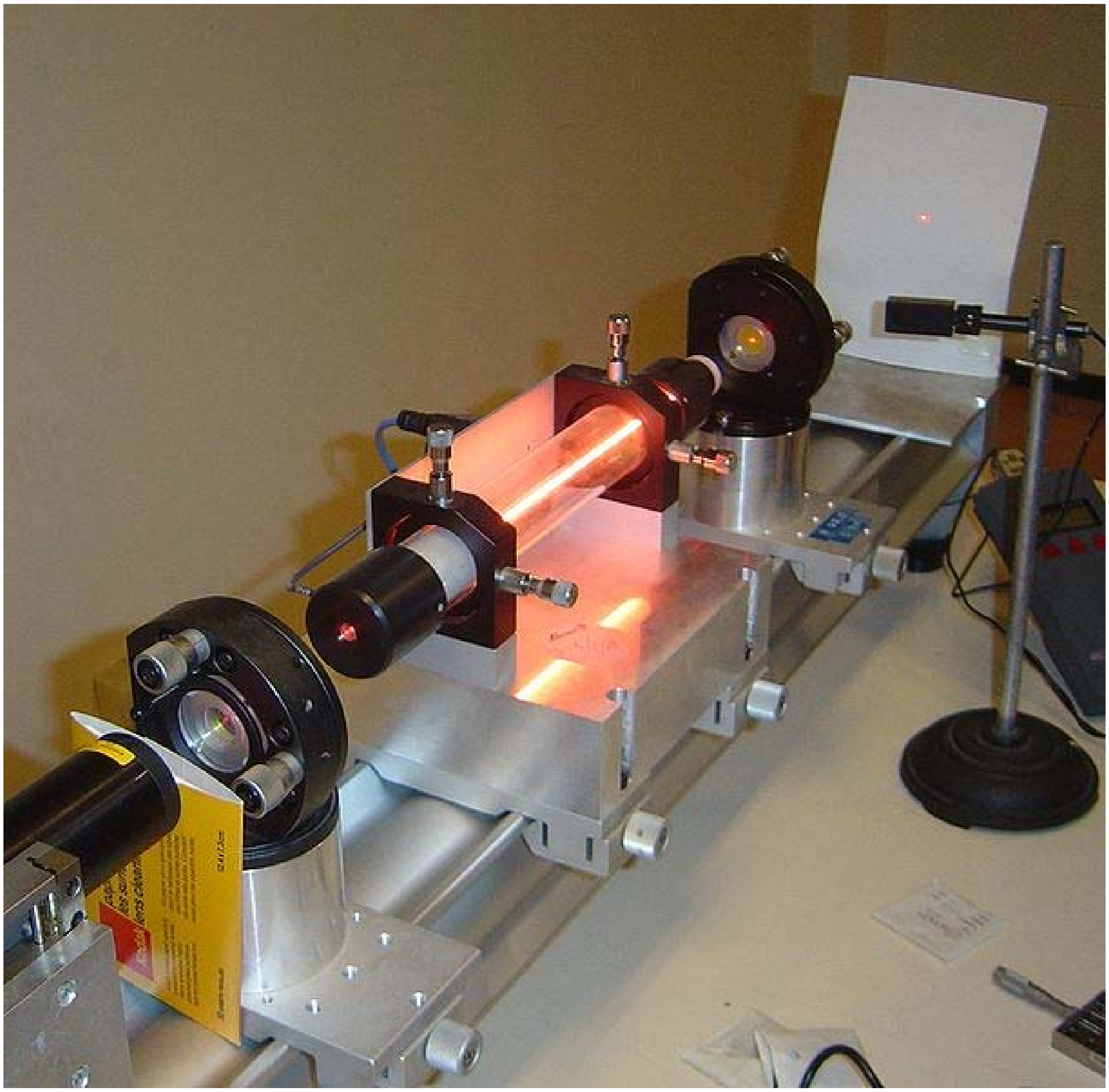
Principal components:
1. Gain medium
2. Laser pumping energy
3. High reflector
4. Output coupler
5. Laser beam

A laser consists of a gain medium inside a highly reflective optical cavity, as well as a means to supply energy to the gain medium. The gain medium is a material with properties that allow it to amplify light by stimulated emission. In its simplest form, a cavity consists of two mirrors arranged such that light bounces back and forth, each time passing through the gain medium. Typically one of the two mirrors, the output coupler, is partially transparent. The output laser beam is emitted through this mirror.

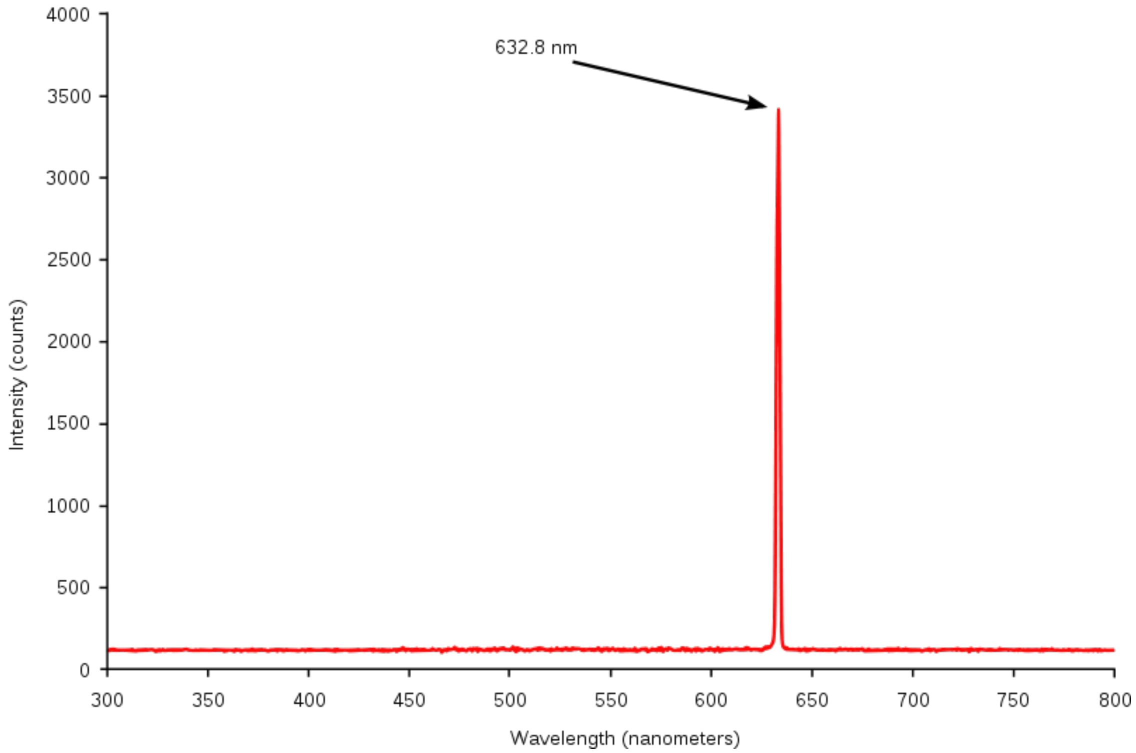
Light of a specific wavelength that passes through the gain medium is amplified (increases in power); the surrounding mirrors ensure that most of the light makes many passes through the gain medium, being amplified repeatedly. Part of the light that is between the mirrors (that is, within the cavity) passes through the partially transparent mirror and escapes as a beam of light.

The process of supplying the energy required for the amplification is called pumping. The energy is typically supplied as an electrical current or as light at a different wavelength. Such light may be provided by a flash lamp or perhaps another laser. Most practical lasers contain additional elements that affect properties such as the wavelength of the emitted light and the shape of the beam.

Laser physics



A helium-neon laser demonstration at the Kastler-Brossel Laboratory at Univ. Paris 6. The pink-orange glow running through the center of the tube is from the electric discharge which inadvertently produces incoherent light, just as in a neon tube. That glowing plasma however also acts as the gain medium through which the internal beam passes as it is reflected in between the two mirrors. Laser radiation output from the front mirror can be seen to produce a tiny (about 1mm in diameter) intense spot on the screen to the right. Although it is a deep and pure red color, spots of laser light are so intense that cameras are typically overexposed and distort their color, often appearing more white.



Spectrum of a helium neon laser illustrating its very high spectral purity (limited by the measuring apparatus). The .002 nm bandwidth of the lasing medium is well over 10,000 times narrower than the spectral width of a light-emitting diode (whose spectrum is shown **here** for comparison), with the bandwidth of a single longitudinal mode being much narrower still.

The gain medium of a laser is a material of controlled purity, size, concentration, and shape, which amplifies the beam by the process of stimulated emission. It can be of any state: gas, liquid, solid or plasma. The gain medium absorbs pump energy, which raises some electrons into higher-energy ("excited") quantum states. Particles can interact with light by either absorbing or emitting photons. Emission can be spontaneous or stimulated. In the latter case, the photon is emitted in the same direction as the light that is passing by. When the number of particles in one excited state exceeds the number of particles in some lower-energy state, population inversion is achieved and the amount of stimulated emission due to light that passes through is larger than the amount of absorption. Hence, the light is amplified. By itself, this makes an optical amplifier. When an optical amplifier is placed inside a resonant optical cavity, one obtains a laser.

The light generated by stimulated emission is very similar to the input signal in terms of wavelength, phase, and polarization. This gives laser light its characteristic coherence, and allows it to maintain the uniform polarization and often monochromaticity established by the optical cavity design.

The optical resonator is sometimes referred to as an "optical cavity", but this is a misnomer: lasers use open resonators as opposed to the literal cavity that would be employed at microwave frequencies in a maser. The resonator typically consists of two mirrors between which a coherent beam of light travels in both directions, reflecting back on itself so that an average photon will pass through the gain medium repeatedly before it is emitted from the output aperture or lost to diffraction or absorption. If the gain (amplification) in the medium is larger than the resonator losses, then the power of the recirculating light can rise exponentially. But each stimulated emission event returns an atom from its excited state to the ground state, reducing the gain of the medium. With increasing beam power the net gain (gain times loss) reduces to unity and the gain medium is said to be saturated. In a continuous wave (CW) laser, the balance of pump power against gain saturation and cavity losses produces an equilibrium value of the laser power inside the cavity; this equilibrium determines the operating point of the laser. If the applied pump power is too small, the gain will never be sufficient to overcome the resonator losses, and laser light will not be produced. The minimum pump power needed to begin laser action is called the *lasing threshold*. The gain medium will amplify any photons passing through it, regardless of direction; but only the photons in a spatial mode supported by the resonator will pass more than once through the medium and receive substantial amplification.

The beam in the cavity and the output beam of the laser, when travelling in free space (or a homogenous medium) rather than waveguides (as in an optical fiber laser), can be approximated as a Gaussian beam in most lasers; such beams exhibit the minimum divergence for a given diameter. However some high power lasers may be multimode, with the transverse modes often approximated using Hermite-Gaussian or Laguerre-Gaussian functions. It has been shown that unstable laser resonators (not used in most lasers) produce fractal shaped beams. Near the beam "waist" (or focal region) it is highly *collimated*: the wavefronts are planar, normal to the direction of propagation, with no beam divergence at that point. However due to diffraction, that can only remain true well within the Rayleigh range. The beam of a single transverse mode (gaussian beam) laser eventually diverges at an angle which varies inversely with the beam diameter, as required by diffraction theory. Thus, the "pencil beam" directly generated by a common helium-neon laser would spread out to a size of perhaps 500 kilometers when shone on the Moon (from the distance of the earth). On the other hand the light from a semiconductor laser typically exits the tiny crystal with a large divergence: up to 50°. However even such a divergent beam can be transformed into a similarly collimated beam by means of a lens system, as is always included, for instance, in a laser pointer whose light originates from a laser diode. That is possible due to the light being of a single spatial mode. This unique property of laser light, spatial coherence, cannot be replicated using standard light sources (except by discarding most of the light) as can be appreciated by comparing the beam from a flashlight (torch) or spotlight to that of almost any laser.

The mechanism of producing radiation in a laser relies on stimulated emission, where energy is extracted from a transition in an atom or molecule. This is a quantum phenomenon discovered by Einstein who derived the relationship between the A coefficient describing spontaneous emission and the B coefficient which applies to absorption and stimulated emission. However in the case of the free electron laser, atomic energy levels are not involved; it appears that the operation of this rather exotic device can be explained without reference to quantum mechanics.

Continuous and pulsed modes of operation

A laser can be classified as operating in either continuous or pulsed mode, depending on whether the power output is essentially continuous over time or whether its output takes the form of pulses of light on one or another time scale. Of course even a laser whose output is normally continuous can be intentionally turned on and off at some rate in order to create pulses of light. When the modulation rate is on time scales much slower than the cavity lifetime and the time period over which energy can be stored in the lasing medium or pumping mechanism, then it is still classified as a "modulated" or "pulsed" continuous wave laser. Most laser diodes used in communication systems fall in that category.

Continuous wave operation

Some applications of lasers depend on a beam whose output power is constant over time. Such a laser is known as *continuous wave* (*CW*). Many types of lasers can be made to operate in continuous wave mode to satisfy such an application. Many of these lasers actually lase in several longitudinal modes at the same time, and beats between the slightly different optical frequencies of those oscillations will in fact produce amplitude variations on time scales shorter than the round-trip time (the reciprocal of the frequency spacing between modes), typically a few nanoseconds or less. In most cases these lasers are still termed "continuous wave" as their output power is steady when averaged over any longer time periods, with the very high frequency power variations having little or no impact in the intended application. (However the term is not applied to mode locked lasers, where the *intention* is to create very short pulses at the rate of the round-trip time).

For continuous wave operation it is required for the population inversion of the gain medium to be continually replenished by a steady pump source. In some lasing media this is impossible. In some other lasers it would require pumping the laser at a very high continuous power level which would be impractical or destroy the laser by producing excessive heat. Such lasers cannot be run in CW mode.

Pulsed operation

Pulsed operation of lasers refers to any laser not classified as continuous wave, so that the optical power appears in pulses of some duration at some repetition rate. This encompasses a wide range of technologies addressing a number of different motivations. Some lasers are pulsed simply because they cannot be run in continuous mode.

In other cases the application requires the production of pulses having as large an energy as possible. Since the pulse energy is equal to the average power divided by the repetition rate, this goal can sometimes be satisfied by lowering the rate of pulses so that more energy can be built up in between pulses. In laser ablation for example, a small volume of material at the surface of a work piece can be evaporated if it is heated in a very short time, whereas supplying the energy gradually would allow for the heat to be absorbed into the bulk of the piece, never attaining a sufficiently high temperature at a particular point.

Other applications rely on the peak pulse power (rather than the energy in the pulse), especially in order to obtain nonlinear optical effects. For a given pulse energy, this requires creating pulses of the shortest possible duration utilizing techniques such as Q-switching.

The optical bandwidth of a pulse cannot be narrower than the reciprocal of the pulse width. In the case of extremely short pulses, that implies lasing over a considerable bandwidth, quite contrary to the very narrow bandwidths typical of CW lasers. The lasing medium in some *dye lasers* and *vibronic solid-state lasers* produces optical gain over a wide bandwidth, making a laser possible which can thus generate pulses of light as short as a few femtoseconds (10^{-15} s).

Q-switching

In a Q-switched laser, the population inversion is allowed to build up by introducing loss inside the resonator which exceeds the gain of the medium; this can also be described as a reduction of the quality factor or 'Q' of the cavity. Then, after the pump energy stored in the laser medium has approached the maximum possible level, the introduced loss mechanism (often an electro- or acousto-optical element) is rapidly removed (or that occurs by itself in a passive device), allowing lasing to begin which rapidly obtains the stored energy in the gain medium. This results in a short pulse incorporating that energy, and thus a high peak power.

Mode-locking

A mode-locked laser is capable of emitting extremely short pulses on the order of tens of picoseconds down to less than 10 femtoseconds. These pulses will repeat at the round trip time, that is, the time that it takes light to complete one round trip between the mirrors comprising the resonator. Due to the Fourier limit (also known as energy-time uncertainty), a pulse of such short temporal length has a spectrum spread over a considerable bandwidth. Thus such a gain medium must have a gain bandwidth sufficiently broad to amplify those frequencies. An example of a suitable material is titanium-doped, artificially grown sapphire (Ti:sapphire) which has a very wide gain bandwidth and can thus produce pulses of only a few femtoseconds duration.

Such mode-locked lasers are a most versatile tool for researching processes occurring on extremely short time scales (known as femtosecond physics, femtosecond chemistry and ultrafast science), for maximizing the effect of nonlinearity in optical materials (e.g. in second-harmonic generation, parametric down-conversion, optical parametric oscillators and the like) due to the large peak power, and in ablation applications. Again, because of the extremely short pulse duration, such a laser will produce pulses which achieve an extremely high peak power.

Pulsed pumping

Another method of achieving pulsed laser operation is to pump the laser material with a source that is itself pulsed, either through electronic charging in the case of flash lamps, or another laser which is already pulsed. Pulsed pumping was historically used with dye lasers where the inverted population lifetime of a dye molecule was so short that a high energy, fast pump was needed. The way to overcome this problem was to charge up large capacitors which are then switched to discharge through flashlamps, producing an intense flash. Pulsed pumping is also required for three-level lasers in which the lower energy level rapidly becomes highly populated preventing further lasing until those atoms relax to the ground state. These lasers, such as the excimer laser and the copper vapor laser, can never be operated in CW mode.

History

Foundations

In 1917, Albert Einstein established the theoretic foundations for the laser and the maser in the paper *Zur Quantentheorie der Strahlung* (On the Quantum Theory of Radiation); via a re-derivation of Max Planck’s law of radiation, conceptually based upon probability coefficients (Einstein coefficients) for the absorption, spontaneous emission, and stimulated emission of electromagnetic radiation; in 1928, Rudolf W. Ladenburg confirmed the existences of the phenomena of stimulated emission and negative absorption; in 1939, Valentin A. Fabrikant predicted the use of stimulated emission to amplify “short” waves; in 1947, Willis E. Lamb and R. C. Retherford found apparent stimulated emission in hydrogen spectra and effected the first demonstration of stimulated emission; in 1950, Alfred Kastler (Nobel Prize for Physics 1966) proposed the method of optical pumping, experimentally confirmed, two years later, by Brossel, Kastler, and Winter.

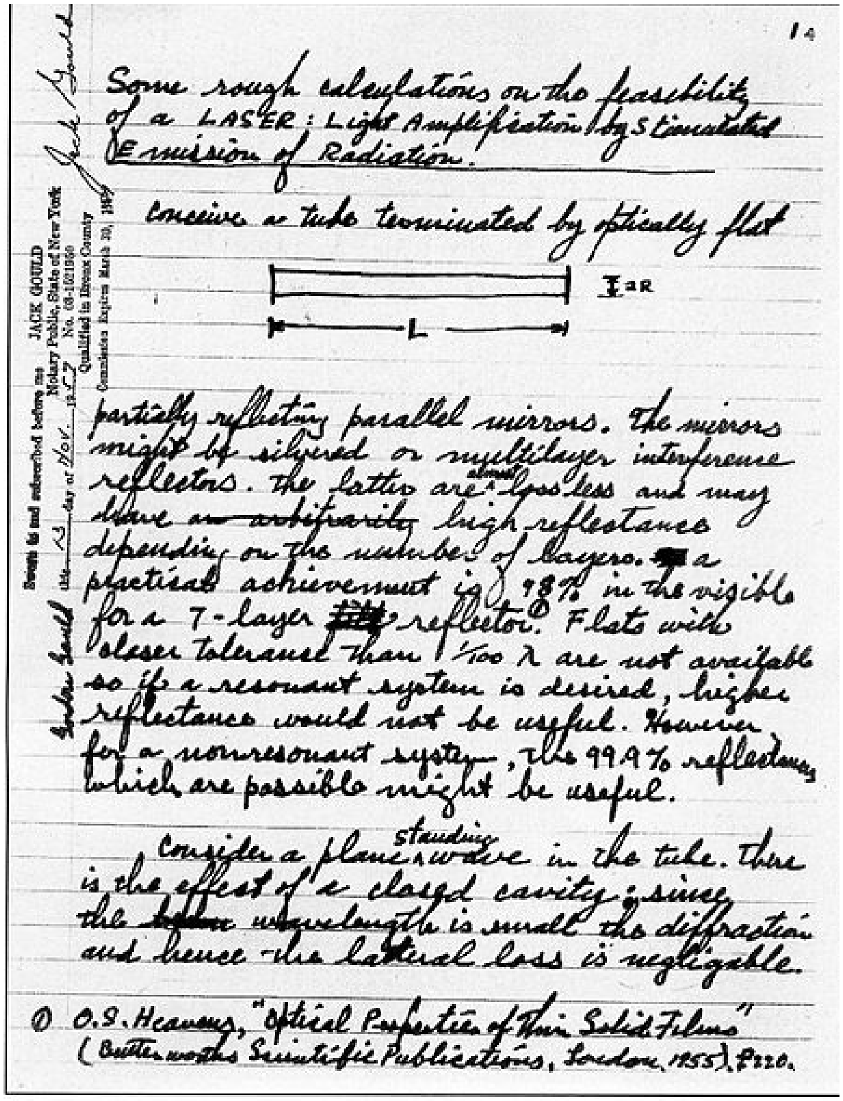
Maser

In 1953, Charles Hard Townes and graduate students James P. Gordon and Herbert J. Zeiger produced the first microwave amplifier, a device operating on similar principles to the laser, but amplifying microwave radiation rather than infrared or visible radiation. Townes's maser was incapable of continuous output. Meanwhile, in the Soviet Union, Nikolay Basov and Aleksandr Prokhorov were independently working on the quantum oscillator and solved the problem of continuous-output systems by using more than two energy levels. These gain media could release stimulated emissions between an excited state and a lower excited state, not the ground state, facilitating the maintenance of a population inversion. In 1955, Prokhorov and Basov suggested optical pumping of a multi-level system as a method for obtaining the population inversion, later a main method of laser pumping.

Townes reports that several eminent physicists — among them Niels Bohr, John von Neumann, Isidor Rabi, Polykarp Kusch, and Llewellyn Thomas — argued the maser violated Heisenberg's uncertainty principle and hence could not work. In 1964 Charles H. Townes, Nikolay Basov, and Aleksandr Prokhorov shared the Nobel Prize in Physics, “for fundamental work in the field of quantum electronics, which has led to the construction of oscillators and amplifiers based on the maser–laser principle”.

Laser

In 1957, Charles Hard Townes and Arthur Leonard Schawlow, then at Bell Labs, began a serious study of the infrared laser. As ideas developed, they abandoned infrared radiation to instead concentrate upon visible light. The concept originally was called an "optical maser". In 1958, Bell Labs filed a patent application for their proposed optical maser; and Schawlow and Townes submitted a manuscript of their theoretical calculations to the Physical Review.



14

Some rough calculations on the feasibility of a LASER: Light Amplification by Stimulated Emission of Radiation.

Conceive a tube terminated by optically flat

partially reflecting parallel mirrors. The mirrors might be silvered or multilayer interference reflectors. The latter are almost loss less and may ~~have an arbitrarily~~ high reflectance depending on the number of layers. a practical achievement is 98% in the visible for a 7-layer reflector.① Flats with closer tolerance than 1/100 λ are not available so if a resonant system is desired, higher reflectance would not be useful. However for a nonresonant system, the 99.9% reflectances which are possible might be useful.

Consider a plane standing wave in the tube. there is the effect of a closed cavity; since the ~~[illegible]~~ wavelength is small the diffraction and hence the lateral loss is negligable.

① O.S. Heavens, "Optical Properties of Thin Solid Films" (Butterworths Scientific Publications, London, 1955), p220.

Sworn to and subscribed before me this 13 day of Nov. 1957
JACK GOULD
Notary Public, State of New York
No. 03-1021950
Qualified in Bronx County
Commission Expires March 30, 1959

Jack Gould

Gordon Gould

LASER notebook: First page of the notebook wherein Gordon Gould coined the LASER acronym, and described the technologic elements for constructing the device.

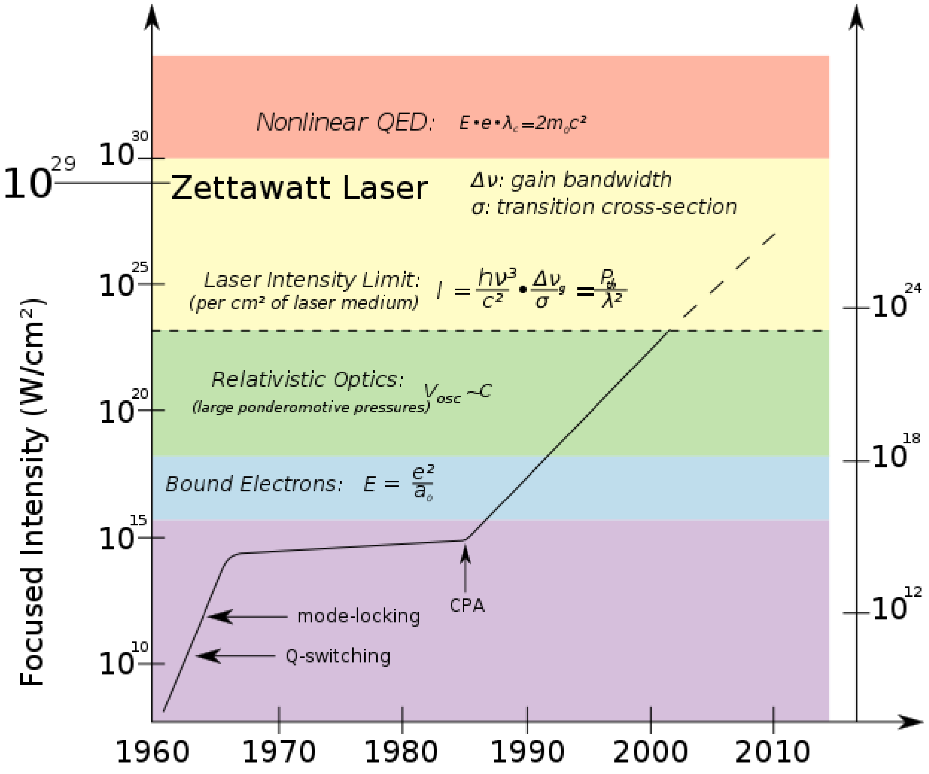
Simultaneously, at Columbia University, graduate student Gordon Gould was working on a doctoral thesis about the energy levels of excited thallium. When Gould and Townes met, they spoke of radiation emission, as a general subject; afterwards, in November 1957, Gould noted his ideas for a "laser", including using an open resonator (later an essential laser-device component). Moreover, in 1958, Prokhorov independently proposed using an open resonator, the first published appearance (the USSR) of this idea. Elsewhere, in the US, Schawlow and Townes had agreed to an open-resonator laser design — apparently unaware of Prokhorov's publications and Gould's unpublished laser work.

At a conference in 1959, Gordon Gould published the term LASER in the paper *The LASER, Light Amplification by Stimulated Emission of Radiation*. Gould's linguistic intention was using the "-aser" word particle as a suffix — to accurately denote the spectrum of the light emitted by the LASER device; thus x-rays: *xaser*, ultraviolet: *uvaser*, et cetera; none established itself as a discrete term, although "raser" was briefly popular for denoting radio-frequency-emitting devices.

Gould's notes included possible applications for a laser, such as spectrometry, interferometry, radar, and nuclear fusion. He continued developing the idea, and filed a patent application in April 1959. The U.S. Patent Office denied his application, and awarded a patent to Bell Labs, in 1960. That provoked a twenty-eight-year lawsuit, featuring scientific prestige and money as the stakes. Gould won his first minor patent in 1977, yet it was not until 1987 that he won the first significant patent lawsuit victory, when a Federal judge ordered the US Patent Office to issue patents to Gould for the optically pumped and the gas discharge laser devices.

In 1960, Theodore H. Maiman constructed the first functioning laser, at Hughes Research Laboratories, Malibu, California, ahead of several research teams, including those of Townes, at Columbia University, Arthur Schawlow, at Bell Labs, and Gould, at the TRG (Technical Research Group) company. Maiman's functional laser used a solid-state flashlamp-pumped synthetic ruby crystal to produce red laser light, at 694 nanometres wavelength; however, the device only was capable of pulsed operation, because of its three-level pumping design scheme. Later in 1960, the Iranian physicist Ali Javan, and William R. Bennett, and Donald Herriot, constructed the first gas laser, using helium and neon that was capable of continuous operation in the infrared (US Patent 3,149,290); later, Javan received the Albert Einstein Award in 1993. Basov and Javan proposed the semiconductor laser diode concept. In 1962, Robert N. Hall demonstrated the first *laser diode* device, made of gallium arsenide and emitted at 850 nm the near-infrared band of the spectrum. Later, in 1962, Nick Holonyak, Jr. demonstrated the first semiconductor laser with a visible emission. This first semiconductor laser could only be used in pulsed-beam operation, and when cooled to liquid nitrogen temperatures (77˚K). In 1970, Zhores Alferov, in the USSR, and Izuo Hayashi and Morton Panish of Bell Telephone Laboratories also independently developed room-temperature, continual-operation diode lasers, using the heterojunction structure.

Recent innovations



Graph showing the history of maximum laser pulse intensity throughout the past 40 years.

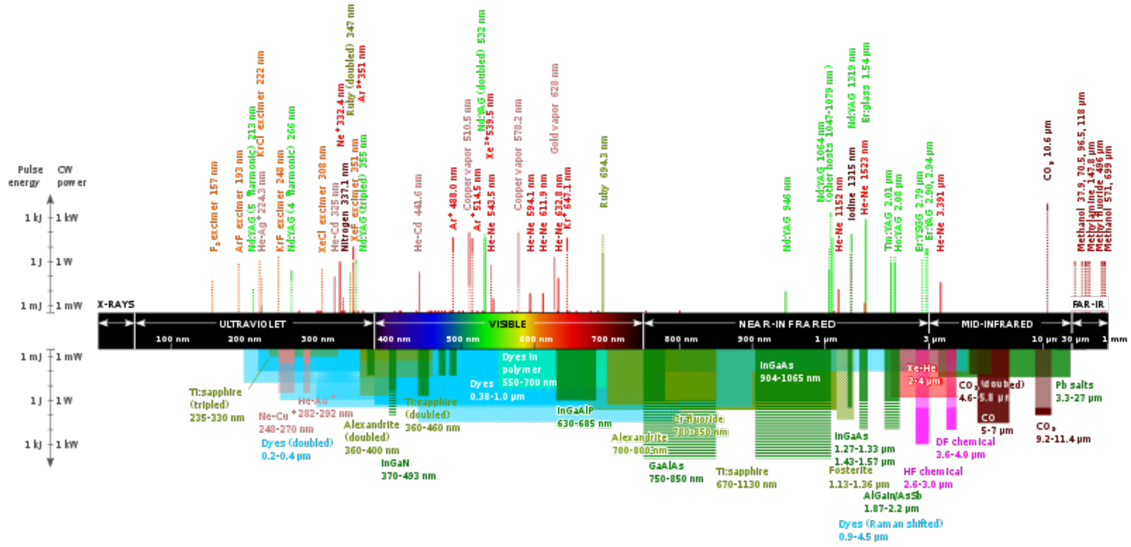
Since the early period of laser history, laser research has produced a variety of improved and specialized laser types, optimized for different performance goals, including:

- new wavelength bands
- maximum average output power
- maximum peak pulse energy
- maximum peak pulse power
- minimum output pulse duration
- maximum power efficiency
- minimum cost

and this research continues to this day.

Lasing without maintaining the medium excited into a population inversion was discovered in 1992 in sodium gas and again in 1995 in rubidium gas by various international teams. This was accomplished by using an external maser to induce "optical transparency" in the medium by introducing and destructively interfering the ground electron transitions between two paths, so that the likelihood for the ground electrons to absorb any energy has been cancelled.

Types and operating principles



Wavelengths of commercially available lasers. Laser types with distinct laser lines are shown above the wavelength bar, while below are shown lasers that can emit in a wavelength range.

Gas lasers

Following the invention of the HeNe gas laser, many other gas discharges have been found to amplify light coherently. Gas lasers using many different gases have been built and used for many purposes. The helium-neon laser (HeNe) is able to operate at a number of different wavelengths, however the vast majority are engineered to lase at 633 nm; these relatively low cost but highly coherent lasers are extremely common in optical research and educational laboratories. Commercial carbon dioxide (CO_2) lasers can emit many hundreds of watts in a single spatial mode which can be concentrated into a tiny spot. This emission is in the thermal infrared at 10.6 µm; such lasers are regularly used in industry for cutting and welding. The efficiency of a CO_2 laser is unusually high: over 10%. Argon-ion lasers can operate at a number of lasing transitions between 351 and 528.7 nm. Depending on the optical design one or more of these transitions can be lasing simultaneously; the most commonly used lines are 458 nm, 488 nm and 514.5 nm. A nitrogen transverse electrical discharge in gas at atmospheric pressure (TEA) laser is an inexpensive gas laser, often home-built by hobbyists, which produces rather incoherent UV light at 337.1 nm. Metal ion lasers are gas lasers that generate deep ultraviolet wavelengths. Helium-silver (HeAg) 224 nm and neon-copper (NeCu) 248 nm are two examples. Like all low-pressure gas lasers, the gain media of these lasers have quite narrow oscillation linewidths, less than 3 GHz (0.5 picometers), making them candidates for use in fluorescence suppressed Raman spectroscopy.

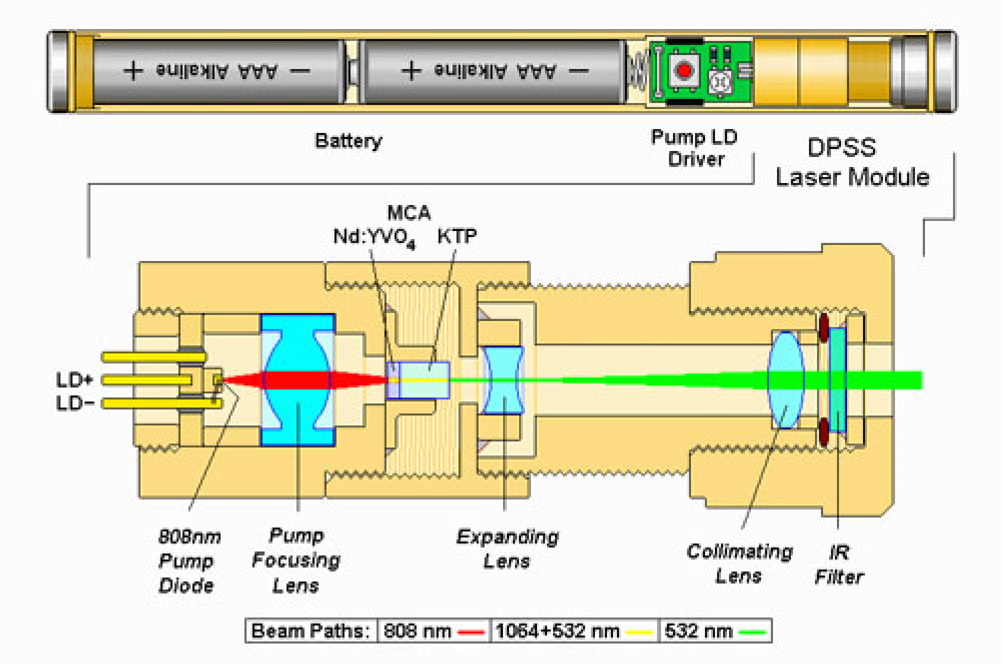
Chemical lasers

Chemical lasers are powered by a chemical reaction permitting a large amount of energy to be released quickly. Such very high power lasers are especially of interest to the military, however continuous wave chemical lasers at very high power levels, fed by streams of gasses, have been developed and have some industrial applications. As examples, in the Hydrogen fluoride laser (2700-2900 nm) and the Deuterium fluoride laser (3800 nm) the reaction is the combination of hydrogen or deuterium gas with combustion products of ethylene in nitrogen trifluoride.

Excimer lasers

Excimer lasers are a special sort of gas laser powered by an electric discharge in which the lasing medium is an excimer, or more precisely an exciplex in existing designs. These are molecules which can only exist with one atom in an excited electronic state. Once the molecule transfers its excitation energy to a photon, therefore, its atoms are no longer bound to each other and the molecule disintegrates. This drastically reduces the population of the lower energy state thus greatly facilitating a population inversion. Excimers currently used are all noble gas compounds; noble gasses are chemically inert and can only form compounds while in an excited state. Excimer lasers typically operate at ultraviolet wavelengths with major applicatons including semiconductor photolithography and LASIK eye surgery. Commonly used excimer molecules include ArF (emission at 193 nm), KrCl (222 nm), KrF (248 nm), XeCl (308 nm), and XeF (351 nm). The molecular fluorine laser, emitting at 157 nm in the vacuum ultraviolet is sometimes referred to as an excimer laser, however this appears to be a misnomer inasmuch as F_2 is a stable compound.

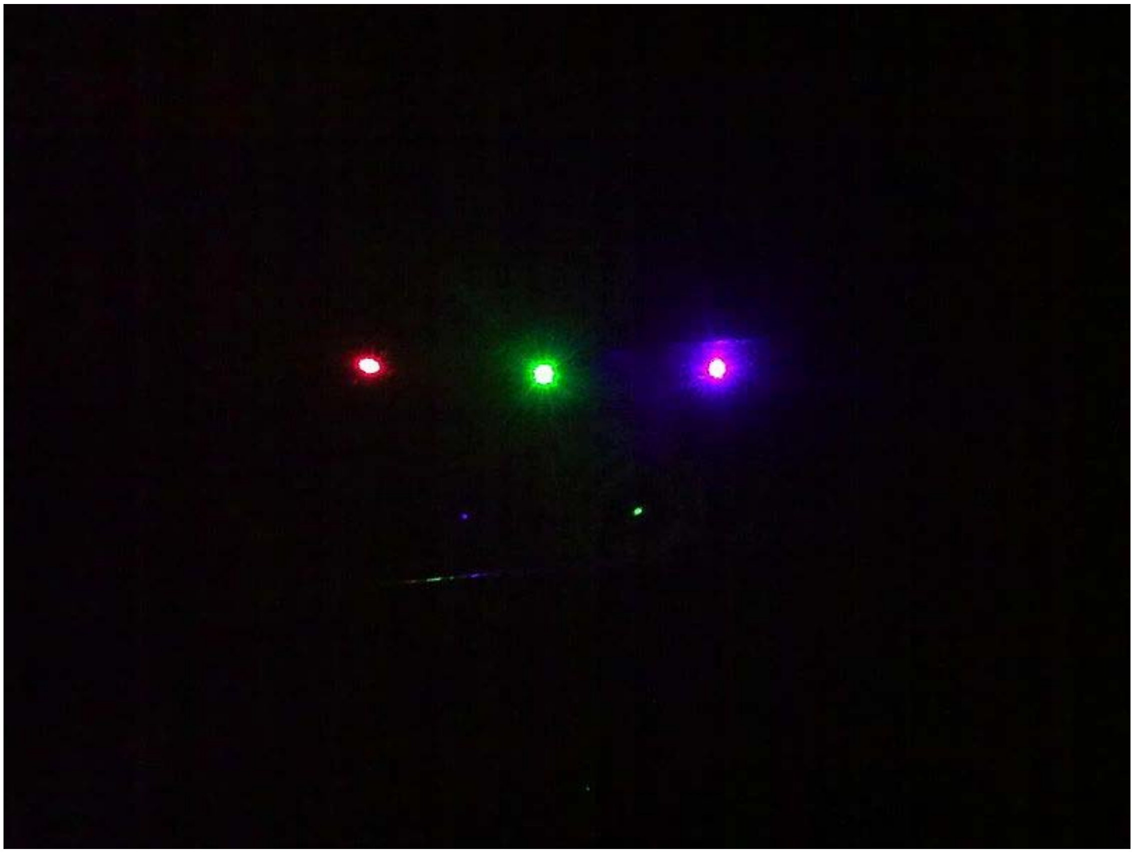
Solid-state lasers



A frequency-doubled green laser pointer, showing internal construction. Two AAA cells and electronics power the laser module (lower diagram) This contains a powerful 808 nm IR diode laser that optically pumps a Nd:YVO_4 crystal inside a laser cavity. That laser produces 1064 nm (infrared) light which is mainly confined inside the resonator. Also inside the laser cavity, however, is a non-linear KTP crystal which causes frequency doubling, resulting in green light at 532 nm. The front mirror is transparent to this visible wavelength which is then expanded and collimated using two lenses (in this particular design).

Solid-state lasers use a crystalline or glass rod which is "doped" with ions that provide the required energy states. For example, the first working laser was a ruby laser, made from ruby (chromium-doped corundum). The population inversion is actually maintained in the "dopant", such as chromium or neodymium. These materials are pumped optically using a shorter wavelength than the lasing wavelength, often from a flashtube or from another laser.

It should be noted that "solid-state" in this sense refers to a crystal or glass, but this usage is distinct from the designation of "solid-state electronics" in referring to semiconductors. Semiconductor lasers (laser diodes) are pumped electrically and are thus *not* referred to as solid-state lasers. The class of solid-state lasers would, however, properly include fiber lasers in which dopants in the glass lase under optical pumping. But in practice these are simply referred to as "fiber lasers" with "solid-state" reserved for lasers using a solid rod of such a material.



Laser spots (650, 532, 405 nm)

Neodymium is a common "dopant" in various solid-state laser crystals, including yttrium orthovanadate (Nd:YVO$_4$), yttrium lithium fluoride (Nd:YLF) and yttrium aluminium garnet (Nd:YAG). All these lasers can produce high powers in the infrared spectrum at 1064 nm. They are used for cutting, welding and marking of metals and other materials, and also in spectroscopy and for pumping dye lasers.

These lasers are also commonly frequency doubled, tripled or quadrupled, in so-called "diode pumped solid state" or DPSS lasers. Under second, third, or fourth harmonic generation these produce 532 nm (green, visible), 355 nm and 266 nm (Ultraviolet|UV]]) beams. This is the technology behind the bright laser pointers particularly at green (532 nm) and other short visible wavelengths.

Ytterbium, holmium, thulium, and erbium are other common "dopants" in solid-state lasers. Ytterbium is used in crystals such as Yb:YAG, Yb:KGW, Yb:KYW, Yb:SYS, Yb:BOYS, Yb:CaF2, typically operating around 1020-1050 nm. They are potentially very efficient and high powered due to a small quantum defect. Extremely high powers in ultrashort pulses can be achieved with Yb:YAG. Holmium-doped YAG crystals emit at 2097 nm and form an efficient laser operating at infrared wavelengths strongly absorbed by water-bearing tissues. The Ho-YAG is usually operated in a pulsed mode, and passed through optical fiber surgical devices to resurface joints, remove rot from teeth, vaporize cancers, and pulverize kidney and gall stones.

Titanium-doped sapphire (Ti:sapphire) produces a highly tunable infrared laser, commonly used for spectroscopy. It is also notable for use as a mode-locked laser producing ultrashort pulses of extremely high peak power.

Thermal limitations in solid-state lasers arise from unconverted pump power that manifests itself as heat. This heat, when coupled with a high thermo-optic coefficient (dn/dT) can give rise to thermal lensing as well as reduced quantum efficiency. These types of issues can be overcome by another novel diode-pumped solid-state laser, the diode-pumped thin disk laser. The thermal limitations in this laser type are mitigated by using a laser medium geometry in which the thickness is much smaller than the diameter of the pump beam. This allows for a more even thermal gradient in the material. Thin disk lasers have been shown to produce up to kilowatt levels of power.

Fiber lasers

Solid-state lasers or laser amplifiers where the light is guided due to the total internal reflection in a single mode optical fiber are instead called fiber lasers. Guiding of light allows extremely long gain regions providing good cooling conditions; fibers have high surface area to volume ratio which allows efficient cooling. In addition, the fiber's waveguiding properties tend to reduce thermal distortion of the beam. Erbium and ytterbium ions are common active species in such lasers.

Quite often, the fiber laser is designed as a double-clad fiber. This type of fiber consists of a fiber core, an inner cladding and an outer cladding. The index of the three concentric layers is chosen so that the fiber core acts as a single-mode fiber for the laser emission while the outer cladding acts as a highly multimode core for the pump laser. This lets the pump propagate a large amount of power into and through the active inner core region, while still having a high numerical aperture (NA) to have easy launching conditions.

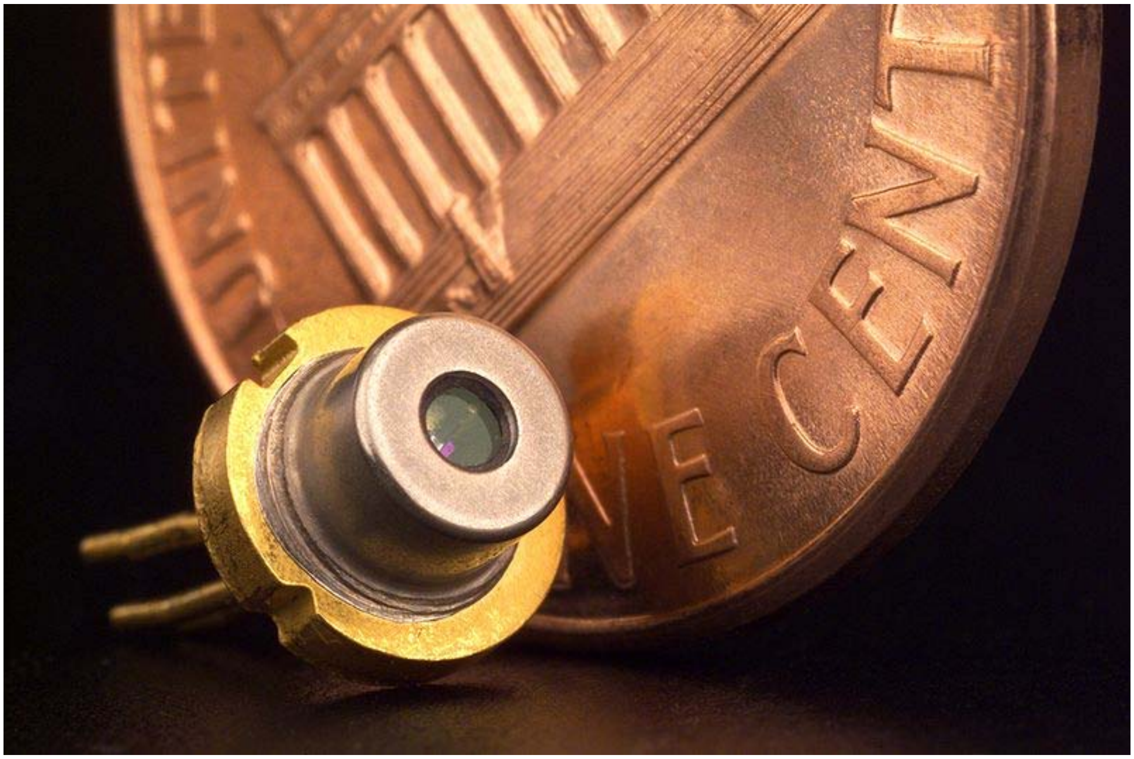
Pump light can be used more efficiently by creating a fiber disk laser, or a stack of such lasers.

Fiber lasers have a fundamental limit in that the intensity of the light in the fiber cannot be so high that optical nonlinearities induced by the local electric field strength can become dominant and prevent laser operation and/or lead to the material destruction of the fiber. This effect is called photodarkening. In bulk laser materials, the cooling is not so efficient, and it is difficult to separate the effects of photodarkening from the thermal effects, but the experiments in fibers show that the photodarkening can be attributed to the formation of long-living color centers.

Photonic crystal lasers

Photonic crystal lasers are lasers based on nano-structures that provide the mode confinement and the density of optical states (DOS) structure required for the feedback to take place. They are typical micrometre-sized and tunable on the bands of the photonic crystals.

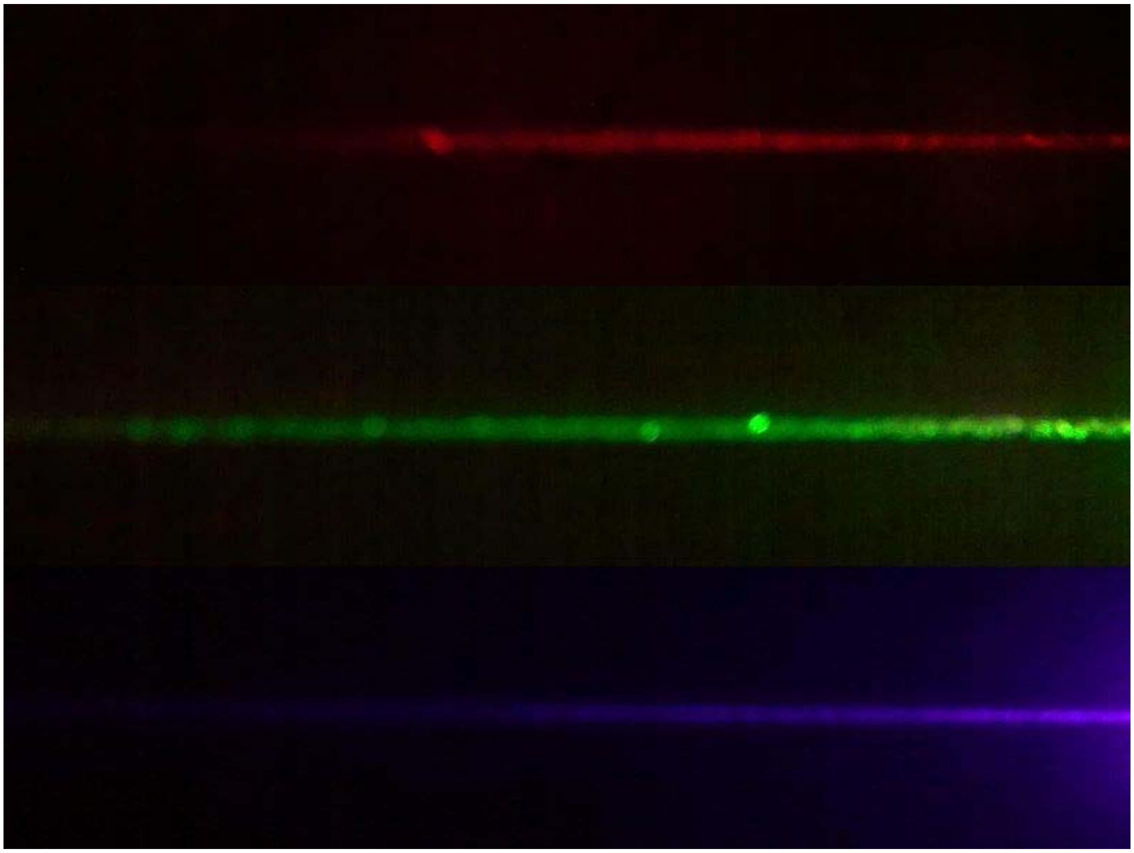
Semiconductor lasers



A 5.6 mm 'closed can' commercial laser diode, probably from a CD or DVD player

Semiconductor lasers are diodes which are electrically pumped. Recombination of electrons and holes created by the applied current introduces optical gain. Reflection from the ends of the crystal form an optical resonator, although the resonator can be external to the semiconductor in some designs.

Commercial laser diodes emit at wavelengths from 375 nm to 1800 nm, and wavelengths of over 3 μm have been demonstrated. Low to medium power laser diodes are used in laser printers and CD/DVD players. Laser diodes are also frequently used to optically pump other lasers with high efficiency. The highest power industrial laser diodes, with power up to 10 kW (70dBm), are used in industry for cutting and welding. External-cavity semiconductor lasers have a semiconductor active medium in a larger cavity. These devices can generate high power outputs with good beam quality, wavelength-tunable narrow-linewidth radiation, or ultrashort laser pulses.



Laser beams (red, green, violet)

Vertical cavity surface-emitting lasers (VCSELs) are semiconductor lasers whose emission direction is perpendicular to the surface of the wafer. VCSEL devices typically have a more circular output beam than conventional laser diodes, and potentially could be much cheaper to manufacture. As of 2005, only 850 nm VCSELs are widely available, with 1300 nm VCSELs beginning to be commercialized, and 1550 nm devices an area of research. VECSELs are external-cavity VCSELs. Quantum cascade lasers are semiconductor lasers that have an active transition between energy *sub-bands* of an electron in a structure containing several quantum wells.

The development of a silicon laser is important in the field of optical computing. Silicon is the material of choice for integrated circuits, and so electronic and silicon photonic components (such as optical interconnects) could be fabricated on the same chip. Unfortunately, silicon is a difficult lasing material to deal with, since it has certain properties which block lasing. However, recently teams have produced silicon lasers through methods such as fabricating the lasing material from silicon and other semiconductor materials, such as indium(III) phosphide or gallium(III) arsenide, materials which allow coherent light to be produced from silicon. These are called hybrid silicon laser. Another type is a Raman laser, which takes advantage of Raman scattering to produce a laser from materials such as silicon.

Dye lasers

Dye lasers use an organic dye as the gain medium. The wide gain spectrum of available dyes, or mixtures of dyes, allows these lasers to be highly tunable, or to produce very short-duration pulses (on the order of a few femtoseconds). Although these tunable lasers are mainly known in their liquid form, researchers have also demonstrated narrow-linewidth tunable emission in dispersive oscillator configurations incorporating solid-state dye gain media. In their most prevalent form these solid state dye lasers use dye-doped polymers as laser media.

Free electron lasers

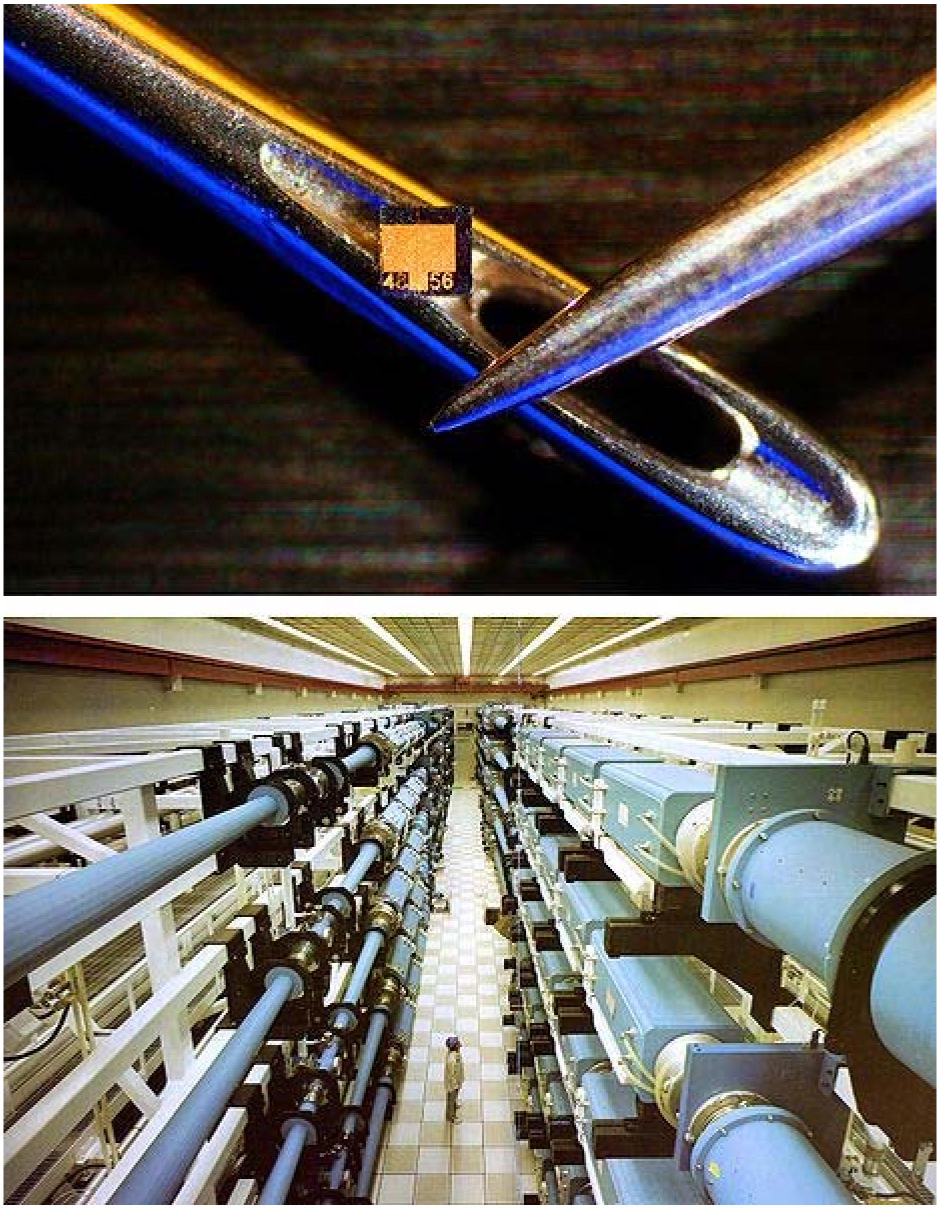
Free electron lasers, or FELs, generate coherent, high power radiation, that is widely tunable, currently ranging in wavelength from microwaves, through terahertz radiation and infrared, to the visible spectrum, to soft X-rays. They have the widest frequency range of any laser type. While FEL beams share the same optical traits as other lasers, such as coherent radiation, FEL operation is quite different. Unlike gas, liquid, or solid-state lasers, which rely on bound atomic or molecular states, FELs use a relativistic electron beam as the lasing medium, hence the term *free electron*.

Exotic laser media

In September 2007, the BBC News reported that there was speculation about the possibility of using positronium annihilation to drive a very powerful gamma ray laser. Dr. David Cassidy of the University of California, Riverside proposed that a single such laser could be used to ignite a nuclear fusion reaction, replacing the banks of hundreds of lasers currently employed in inertial confinement fusion experiments.

Space-based X-ray lasers pumped by a nuclear explosion have also been proposed as antimissile weapons. Such devices would be one-shot weapons.

Uses



Lasers range in size from microscopic diode lasers (top) with numerous applications, to football field sized neodymium glass lasers (bottom) used for inertial confinement fusion, nuclear weapons research and other high energy density physics experiments.

When lasers were invented in 1960, they were called "a solution looking for a problem". Since then, they have become ubiquitous, finding utility in thousands of highly varied applications in every section of modern society, including consumer electronics, information technology, science, medicine, industry, law enforcement, entertainment, and the military.

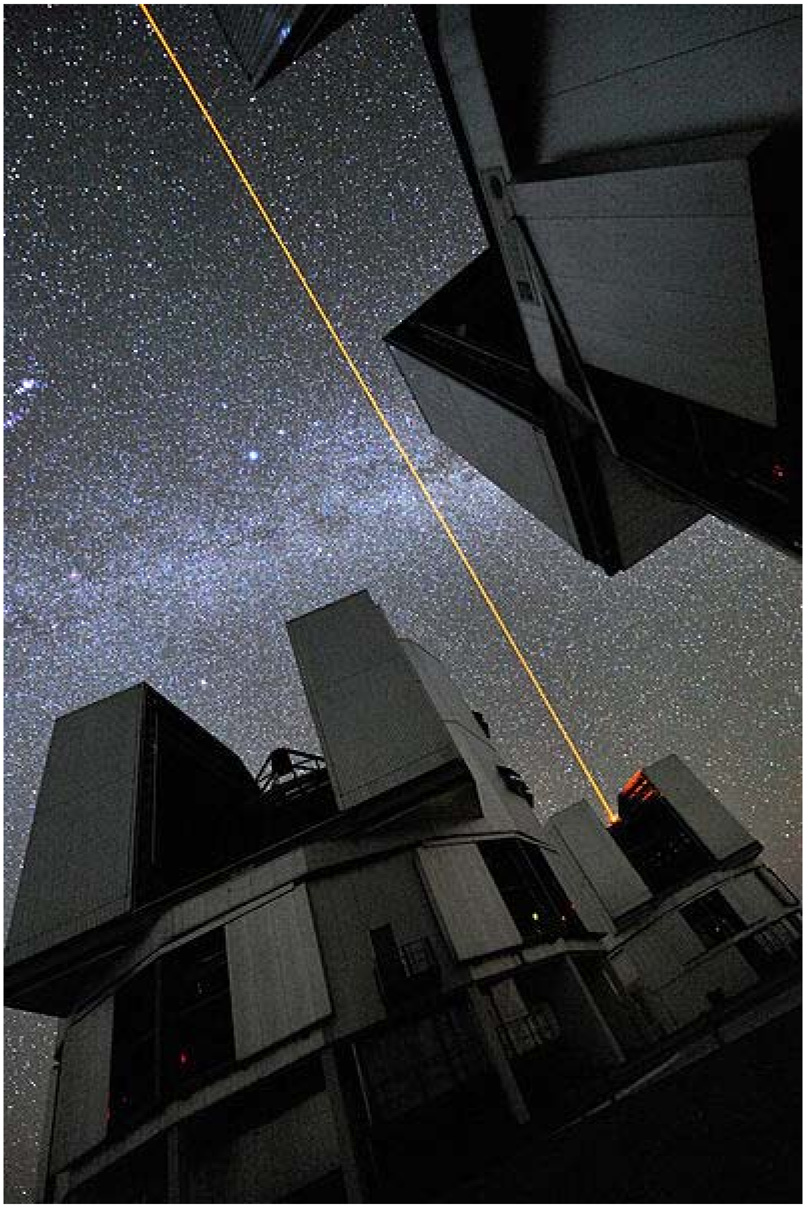
The first use of lasers in the daily lives of the general population was the supermarket barcode scanner, introduced in 1974. The laserdisc player, introduced in 1978, was the first successful consumer product to include a laser but the compact disc player was the first laser-equipped device to become common, beginning in 1982 followed shortly by laser printers.

Some other uses are:

- Medicine: Bloodless surgery, laser healing, surgical treatment, kidney stone treatment, eye treatment, dentistry
- Industry: Cutting, welding, material heat treatment, marking parts, non-contact measurement of parts
- Military: Marking targets, guiding munitions, missile defence, electro-optical countermeasures (EOCM), alternative to radar, blinding troops.
- Law enforcement: used for latent fingerprint detection in the forensic identification field
- Research: Spectroscopy, laser ablation, laser annealing, laser scattering, laser interferometry, LIDAR, laser capture microdissection, fluorescence microscopy
- Product development/commercial: laser printers, optical discs (e.g. CDs and the like), barcode scanners, thermometers, laser pointers, holograms, bubblegrams.
- Laser lighting displays: Laser light shows
- Cosmetic skin treatments: acne treatment, cellulite and striae reduction, and hair removal.

In 2004, excluding diode lasers, approximately 131,000 lasers were sold with a value of US$2.19 billion. In the same year, approximately 733 million diode lasers, valued at $3.20 billion, were sold.

Examples by power



Laser application in astronomical adaptive optics imaging

Different applications need lasers with different output powers. Lasers that produce a continuous beam or a series of short pulses can be compared on the basis of their average power. Lasers that produce pulses can also be characterized based on the *peak* power of each pulse. The peak power of a pulsed laser is many orders of magnitude greater than its average power. The average output power is always less than the power consumed.

The continuous or average power required for some uses:

- 1-5 mW – laser pointers

- 5 mW – CD-ROM drive
- 5–10 mW – DVD player or DVD-ROM drive
- 100 mW – High-speed CD-RW burner
- 250 mW – Consumer DVD-R burner
- 1 W – green laser in current Holographic Versatile Disc prototype development
- 1–20 W – output of the majority of commercially available solid-state lasers used for micro machining
- 30–100 W – typical sealed CO_2 surgical lasers
- 100–3000 W – typical sealed CO_2 lasers used in industrial laser cutting
- 1 kW – Output power expected to be achieved by a prototype 1 cm diode laser bar
- 100 kW – Claimed output of a CO_2 laser being developed by Northrop Grumman for military (weapon) applications

Examples of pulsed systems with high peak power:

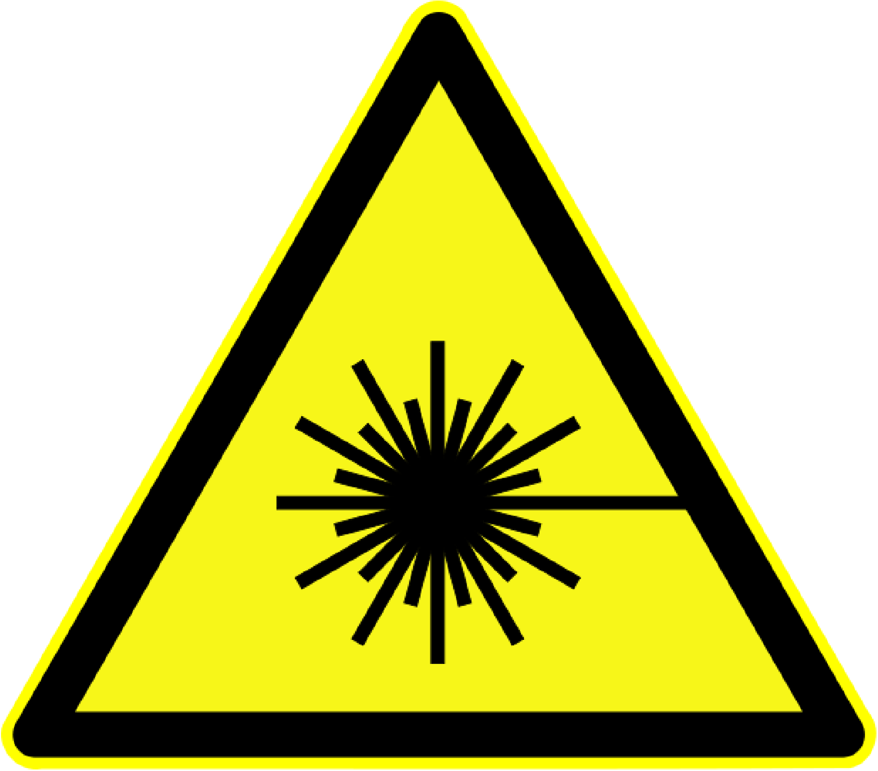
- 700 TW (700×10^{12} W) – National Ignition Facility, a 192-beam, 1.8-megajoule laser system adjoining a 10-meter-diameter target chamber.
- 1.3 PW (1.3×10^{15} W) – world's most powerful laser as of 1998, located at the Lawrence Livermore Laboratory

Hobby uses

In recent years, some hobbyists have taken interests in lasers. Lasers used by hobbyists are generally of class IIIa or IIIb, although some have made their own class IV types. However, compared to other hobbyists, laser hobbyists are far less common, due to the cost and potential dangers involved. Due to the cost of lasers, some hobbyists use inexpensive means to obtain lasers, such as salvaging laser diodes from broken DVD players (red), Blu-ray players (violet), or even higher power laser diodes from CD or DVD burners.

Hobbyists also have been taking surplus pulsed lasers from retired military applications and modifying them for pulsed holography. Pulsed Ruby and pulsed YAG lasers have been used.

Safety



Warning symbol for lasers.

Even the first laser was recognized as being potentially dangerous. Theodore Maiman characterized the first laser as having a power of one "Gillette" as it could burn through one Gillette razor blade. Today, it is accepted that even low-power lasers with only a few milliwatts of output power can be hazardous to human eyesight, when the beam from such a laser hits the eye directly or after reflection from a shiny surface. At wavelengths which the cornea and the lens can focus well, the coherence and low divergence of laser light means that it can be focused by the eye into an extremely small spot on the retina, resulting in localized burning and permanent damage in seconds or even less time.

Lasers are usually labeled with a safety class number, which identifies how dangerous the laser is:

- Class I/1 is inherently safe, usually because the light is contained in an enclosure, for example in CD players.
- Class II/2 is safe during normal use; the blink reflex of the eye will prevent damage. Usually up to 1 mW power, for example laser pointers.
- Class IIIa/3R lasers are usually up to 5 mW and involve a small risk of eye damage within the time of the blink reflex. Staring into such a beam for several seconds is likely to cause damage to a spot on the retina.
- Class IIIb/3B can cause immediate eye damage upon exposure.

- Class IV/4 lasers can burn skin, and in some cases, even scattered light can cause eye and/or skin damage. Many industrial and scientific lasers are in this class.

The indicated powers are for visible-light, continuous-wave lasers. For pulsed lasers and invisible wavelengths, other power limits apply. People working with class 3B and class 4 lasers can protect their eyes with safety goggles which are designed to absorb light of a particular wavelength.

Certain infrared lasers with wavelengths beyond about 1.4 micrometres are often referred to as being "eye-safe". This is because the intrinsic molecular vibrations of water molecules very strongly absorb light in this part of the spectrum, and thus a laser beam at these wavelengths is attenuated so completely as it passes through the eye's cornea that no light remains to be focused by the lens onto the retina. The label "eye-safe" can be misleading, however, as it only applies to relatively low power continuous wave beams; any high power or Q-switched laser at these wavelengths can burn the cornea, causing severe eye damage.

As weapons

Laser beams are famously employed as weapon systems in science fiction, but actual laser weapons are still in the experimental stage. The general idea of laser-beam weaponry is to hit a target with a train of brief pulses of light. The rapid evaporation and expansion of the surface causes shockwaves that damage the target. The power needed to project a high-powered laser beam of this kind is beyond the limit of current mobile power technology thus favoring chemically powered gas dynamic lasers.

Lasers of all but the lowest powers can potentially be used as incapacitating weapons, through their ability to produce temporary or permanent vision loss in varying degrees when aimed at the eyes. The degree, character, and duration of vision impairment caused by eye exposure to laser light varies with the power of the laser, the wavelength(s), the collimation of the beam, the exact orientation of the beam, and the duration of exposure. Lasers of even a fraction of a watt in power can produce immediate, permanent vision loss under certain conditions, making such lasers potential non-lethal but incapacitating weapons. The extreme handicap that laser-induced blindness represents makes the use of lasers even as non-lethal weapons morally controversial, and weapons designed to cause blindness have been banned by the Protocol on Blinding Laser Weapons. The U.S. Air Force is currently working on the YAL-1 airborne laser, mounted in a Boeing 747, to shoot down enemy ballistic missiles over enemy territory.

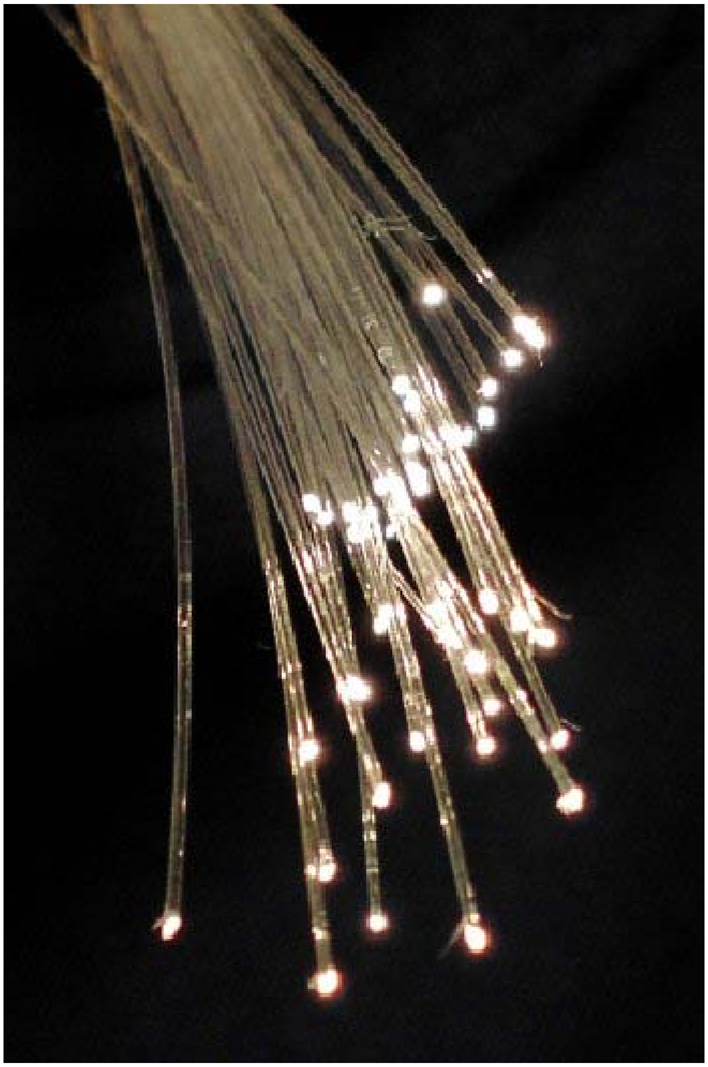
In the field of aviation, the hazards of exposure to ground-based lasers deliberately aimed at pilots have grown to the extent that aviation authorities have special procedures to deal with such hazards.

On March 18, 2009 Northrop Grumman claimed that its engineers in Redondo Beach had successfully built and tested an electrically powered CO_2 laser capable of producing a 100-kilowatt beam, powerful enough to destroy an airplane or a tank. According to Brian

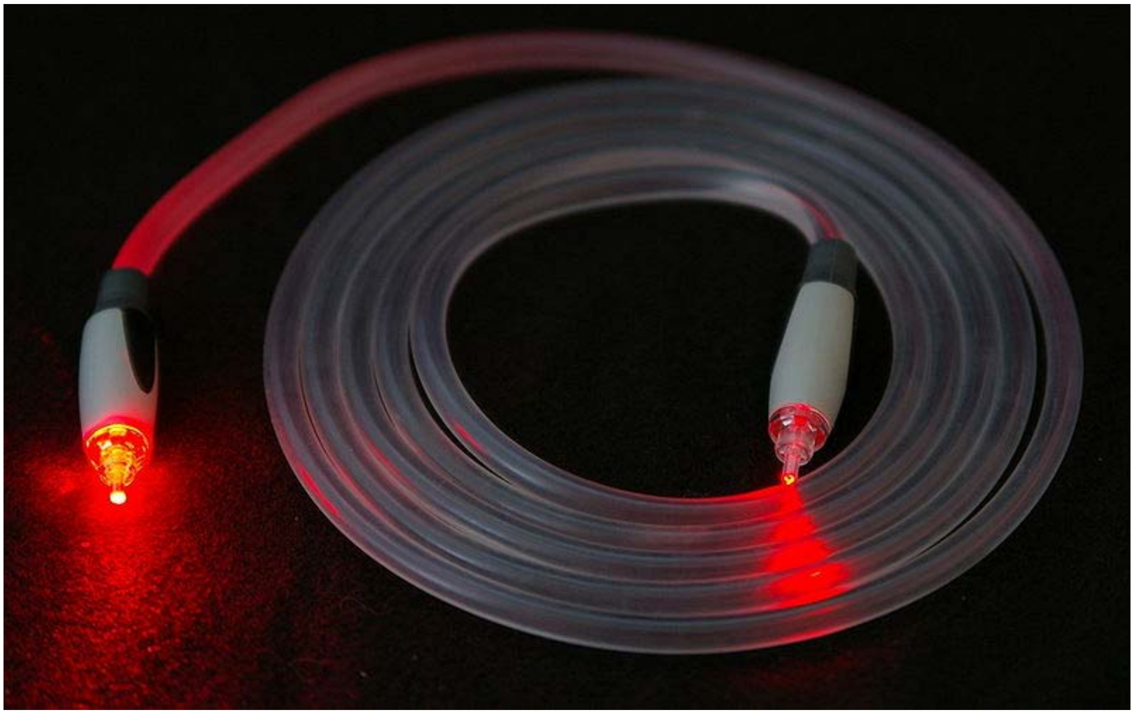
Strickland, manager for the United States Army's Joint High Power Solid State Laser program, an electrically powered laser is capable of being mounted in an aircraft, ship, or other vehicle because it requires much less space for its supporting equipment than a chemical laser. However the source of such a large electrical power in a mobile application remains unclear.

Chapter 8

Optical Fiber



A bundle of optical fibers



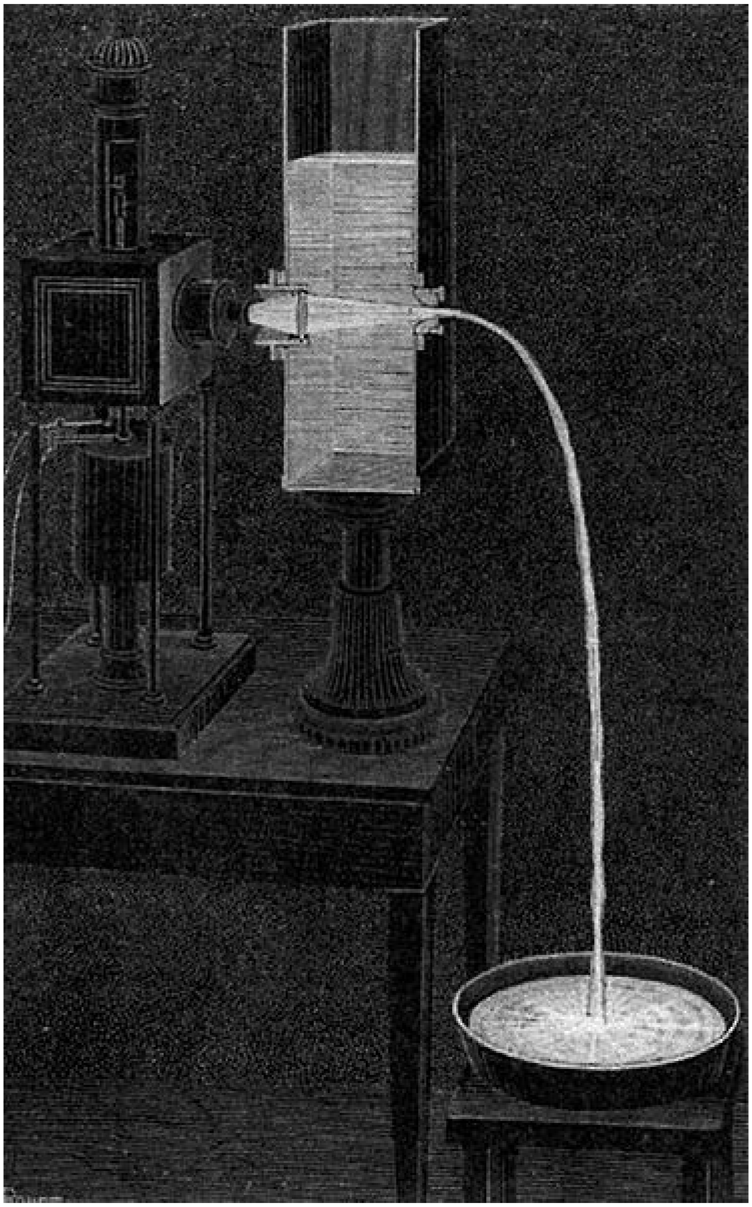
A TOSLINK fiber optic audio cable being illuminated at one end

An **optical fiber** or **optical fibre** is a thin, flexible, transparent fiber that acts as a waveguide, or "light pipe", to transmit light between the two ends of the fiber. The field of applied science and engineering concerned with the design and application of optical fibers is known as **fiber optics**. Optical fibers are widely used in fiber-optic communications, which permits transmission over longer distances and at higher bandwidths (data rates) than other forms of communication. Fibers are used instead of metal wires because signals travel along them with less loss and are also immune to electromagnetic interference. Fibers are also used for illumination, and are wrapped in bundles so they can be used to carry images, thus allowing viewing in tight spaces. Specially designed fibers are used for a variety of other applications, including sensors and fiber lasers.

Optical fiber typically consists of a transparent core surrounded by a transparent cladding material with a lower index of refraction. Light is kept in the core by total internal reflection. This causes the fiber to act as a waveguide. Fibers which support many propagation paths or transverse modes are called multi-mode fibers (MMF), while those which can only support a single mode are called single-mode fibers (SMF). Multi-mode fibers generally have a larger core diameter, and are used for short-distance communication links and for applications where high power must be transmitted. Single-mode fibers are used for most communication links longer than 1,050 meters (3,440 ft).

Joining lengths of optical fiber is more complex than joining electrical wire or cable. The ends of the fibers must be carefully cleaved, and then spliced together either mechanically or by fusing them together with heat. Special optical fiber connectors are used to make removable connections.

History



Daniel Colladon first described this "light fountain" or "light pipe" in an 1842 article titled *On the reflections of a ray of light inside a parabolic liquid stream*. This particular illustration comes from a later article by Colladon, in 1884.

Fiber optics, though used extensively in the modern world, is a fairly simple and old technology. Guiding of light by refraction, the principle that makes fiber optics possible, was first demonstrated by Daniel Colladon and Jacques Babinet in Paris in the early

1840s. John Tyndall included a demonstration of it in his public lectures in London a dozen years later. Tyndall also wrote about the property of total internal reflection in an introductory book about the nature of light in 1870: "When the light passes from air into water, the refracted ray is bent *towards* the perpendicular... When the ray passes from water to air it is bent *from* the perpendicular... If the angle which the ray in water encloses with the perpendicular to the surface be greater than 48 degrees, the ray will not quit the water at all: it will be *totally reflected* at the surface.... The angle which marks the limit where total reflection begins is called the limiting angle of the medium. For water this angle is 48°27', for flint glass it is 38°41', while for diamond it is 23°42'."

Practical applications, such as close internal illumination during dentistry, appeared early in the twentieth century. Image transmission through tubes was demonstrated independently by the radio experimenter Clarence Hansell and the television pioneer John Logie Baird in the 1920s. The principle was first used for internal medical examinations by Heinrich Lamm in the following decade. In 1952, physicist Narinder Singh Kapany conducted experiments that led to the invention of optical fiber. Modern optical fibers, where the glass fiber is coated with a transparent cladding to offer a more suitable refractive index, appeared later in the decade. Development then focused on fiber bundles for image transmission. The first fiber optic semi-flexible gastroscope was patented by Basil Hirschowitz, C. Wilbur Peters, and Lawrence E. Curtiss, researchers at the University of Michigan, in 1956. In the process of developing the gastroscope, Curtiss produced the first glass-clad fibers; previous optical fibers had relied on air or impractical oils and waxes as the low-index cladding material. A variety of other image transmission applications soon followed.

In the late 19th and early 20th centuries, light was guided through bent glass rods to illuminate body cavities. Alexander Graham Bell invented a 'Photophone' to transmit voice signals over an optical beam.

Jun-ichi Nishizawa, a Japanese scientist at Tohoku University, also proposed the use of optical fibers for communications in 1963, as stated in his book published in 2004 in India. Nishizawa invented other technologies which contributed to the development of optical fiber communications, such as the graded-index optical fiber as a channel for transmitting light from semiconductor lasers. Charles K. Kao and George A. Hockham of the British company Standard Telephones and Cables (STC) were the first to promote the idea that the attenuation in optical fibers could be reduced below 20 decibels per kilometer (dB/km), allowing fibers to be a practical medium for communication. They proposed that the attenuation in fibers available at the time was caused by impurities, which could be removed, rather than fundamental physical effects such as scattering. They correctly and systematically theorized the light-loss properties for optical fiber, and pointed out the right material to manufacture such fibers — silica glass with high purity. This discovery led to Kao being awarded the Nobel Prize in Physics in 2009.

NASA used fiber optics in the television cameras sent to the moon. At the time such use in the cameras was 'classified confidential' and only those with the right security clearance or those accompanied by someone with the right security clearance were permitted to handle the cameras.

The crucial attenuation limit of 20 dB/km was first achieved in 1970, by researchers Robert D. Maurer, Donald Keck, Peter C. Schultz, and Frank Zimar working for American glass maker Corning Glass Works, now Corning Incorporated. They demonstrated a fiber with 17 dB/km attenuation by doping silica glass with titanium. A few years later they produced a fiber with only 4 dB/km attenuation using germanium dioxide as the core dopant. Such low attenuation ushered in optical fiber telecommunication. In 1981, General Electric produced fused quartz ingots that could be drawn into fiber optic strands 25 miles (40 km) long.

Attenuation in modern optical cables is far less than in electrical copper cables, leading to long-haul fiber connections with repeater distances of 70–150 kilometers (43–93 mi). The erbium-doped fiber amplifier, which reduced the cost of long-distance fiber systems by reducing or eliminating optical-electrical-optical repeaters, was co-developed by teams led by David N. Payne of the University of Southampton and Emmanuel Desurvire at Bell Labs in 1986. Robust modern optical fiber uses glass for both core and sheath and is therefore less prone to aging processes. It was invented by Gerhard Bernsee of Schott Glass in Germany in 1973.

The emerging field of photonic crystals led to the development in 1991 of photonic-crystal fiber which guides light by diffraction from a periodic structure, rather than by total internal reflection. The first photonic crystal fibers became commercially available in 2000. Photonic crystal fibers can carry higher power than conventional fibers and their wavelength-dependent properties can be manipulated to improve performance.

Applications

Optical fiber communication

Optical fiber can be used as a medium for telecommunication and networking because it is flexible and can be bundled as cables. It is especially advantageous for long-distance communications, because light propagates through the fiber with little attenuation compared to electrical cables. This allows long distances to be spanned with few repeaters. Additionally, the per-channel light signals propagating in the fiber have been modulated at rates as high as 111 gigabits per second by NTT, although 10 or 40 Gbit/s is typical in deployed systems. Each fiber can carry many independent channels, each using a different wavelength of light (wavelength-division multiplexing (WDM)). The net data rate (data rate without overhead bytes) per fiber is the per-channel data rate reduced by the FEC overhead, multiplied by the number of channels (usually up to eighty in commercial dense WDM systems as of 2008). The current laboratory fiber optic data rate record, held by Bell Labs in Villarceaux, France, is multiplexing 155 channels, each carrying 100 Gbit/s over a 7000 km fiber. Nippon Telegraph and Telephone Corporation have also managed 69.1 Tbit/s over a single 240 km fiber (multiplexing 432 channels, equating to 171 Gbit/s per channel). Bell Labs also broke a 100 Petabit per second *kilometer* barrier (15.5 Tbit/s over a single 7000 km fiber).

For short distance applications, such as creating a network within an office building, fiber-optic cabling can be used to save space in cable ducts. This is because a single fiber can often carry much more data than many electrical cables, such as 4 pair Cat-5 Ethernet cabling. Fiber is also immune to electrical interference; there is no cross-talk between signals in different cables and no pickup of environmental noise. Non-armored fiber cables do not conduct electricity, which makes fiber a good solution for protecting communications equipment located in high voltage environments such as power generation facilities, or metal communication structures prone to lightning strikes. They can also be used in environments where explosive fumes are present, without danger of ignition. Wiretapping is more difficult compared to electrical connections, and there are concentric dual core fibers that are said to be tap-proof.

Fiber optic sensors

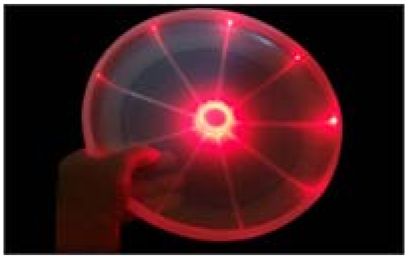
Fibers have many uses in remote sensing. In some applications, the sensor is itself an optical fiber. In other cases, fiber is used to connect a non-fiberoptic sensor to a measurement system. Depending on the application, fiber may be used because of its small size, or the fact that no electrical power is needed at the remote location, or because many sensors can be multiplexed along the length of a fiber by using different wavelengths of light for each sensor, or by sensing the time delay as light passes along the fiber through each sensor. Time delay can be determined using a device such as an optical time-domain reflectometer.

Optical fibers can be used as sensors to measure strain, temperature, pressure and other quantities by modifying a fiber so that the quantity to be measured modulates the intensity, phase, polarization, wavelength or transit time of light in the fiber. Sensors that vary the intensity of light are the simplest, since only a simple source and detector are required. A particularly useful feature of such fiber optic sensors is that they can, if required, provide distributed sensing over distances of up to one meter.

Extrinsic fiber optic sensors use an optical fiber cable, normally a multi-mode one, to transmit modulated light from either a non-fiber optical sensor, or an electronic sensor connected to an optical transmitter. A major benefit of extrinsic sensors is their ability to reach places which are otherwise inaccessible. An example is the measurement of temperature inside aircraft jet engines by using a fiber to transmit radiation into a radiation pyrometer located outside the engine. Extrinsic sensors can also be used in the same way to measure the internal temperature of electrical transformers, where the extreme electromagnetic fields present make other measurement techniques impossible. Extrinsic sensors are used to measure vibration, rotation, displacement, velocity, acceleration, torque, and twisting. A solid state version of the gyroscope using the interference of light has been developed. The fiber optic gyroscope (FOG) has no moving parts and exploits the Sagnac effect to detect mechanical rotation.

A common use for fiber optic sensors are in advanced intrusion detection security systems, where the light is transmitted along the fiber optic sensor cable, which is placed on a fence, pipeline or communication cabling, and the returned signal is monitored and analysed for disturbances. This return signal is digitally processed to identify if there is a disturbance, and if an intrusion has occurred an alarm is triggered by the fiber optic security system.

Other uses of optical fibers



A frisbee illuminated by fiber optics

Light reflected from optical fiber illuminates exhibited model

Fiber optic front sight on a hand gun

Fibers are widely used in illumination applications. They are used as light guides in medical and other applications where bright light needs to be shone on a target without a clear line-of-sight path. In some buildings, optical fibers are used to route sunlight from the roof to other parts of the building. Optical fiber illumination is also used for decorative applications, including signs, art, and artificial Christmas trees. Swarovski boutiques use optical fibers to illuminate their crystal showcases from many different angles while only employing one light source. Optical fiber is an intrinsic part of the light-transmitting concrete building product, LiTraCon.

Optical fiber is also used in imaging optics. A coherent bundle of fibers is used, sometimes along with lenses, for a long, thin imaging device called an endoscope, which is used to view objects through a small hole. Medical endoscopes are used for minimally invasive exploratory or surgical procedures (endoscopy). Industrial endoscopes are used for inspecting anything hard to reach, such as jet engine interiors.

In spectroscopy, optical fiber bundles are used to transmit light from a spectrometer to a substance which cannot be placed inside the spectrometer itself, in order to analyze its composition. A spectrometer analyzes substances by bouncing light off of and through them. By using fibers, a spectrometer can be used to study objects that are too large to fit inside, or gasses, or reactions which occur in pressure vessels.

An optical fiber doped with certain rare earth elements such as erbium can be used as the gain medium of a laser or optical amplifier. Rare-earth doped optical fibers can be used to provide signal amplification by splicing a short section of doped fiber into a regular (undoped) optical fiber line. The doped fiber is optically pumped with a second laser wavelength that is coupled into the line in addition to the signal wave. Both wavelengths of light are transmitted through the doped fiber, which transfers energy from the second pump wavelength to the signal wave. The process that causes the amplification is stimulated emission.

Optical fibers doped with a wavelength shifter are used to collect scintillation light in physics experiments.

Optical fiber can be used to supply a low level of power (around one watt) to electronics situated in a difficult electrical environment. Examples of this are electronics in high-powered antenna elements and measurement devices used in high voltage transmission equipment.

A growing trend in iron sights for arms, is the use of short pieces of optical fiber for contrast enhancement dots, made in such a way that ambient light falling on the length of the fiber is concentrated at the tip, making the dots slightly brighter than the surroundings. This method is most commonly used in front sights, but many makers offer sights that use fiber optics on front and rear sights. Fiber optic sights can now be found on handguns, rifles, and shotguns, both as aftermarket accessories and a growing number of factory guns.

Principle of operation

An optical fiber is a cylindrical dielectric waveguide (nonconducting waveguide) that transmits light along its axis, by the process of total internal reflection. The fiber consists of a *core* surrounded by a cladding layer, both of which are made of dielectric materials. To confine the optical signal in the core, the refractive index of the core must be greater than that of the cladding. The boundary between the core and cladding may either be abrupt, in *step-index fiber*, or gradual, in *graded-index fiber*.

Index of refraction

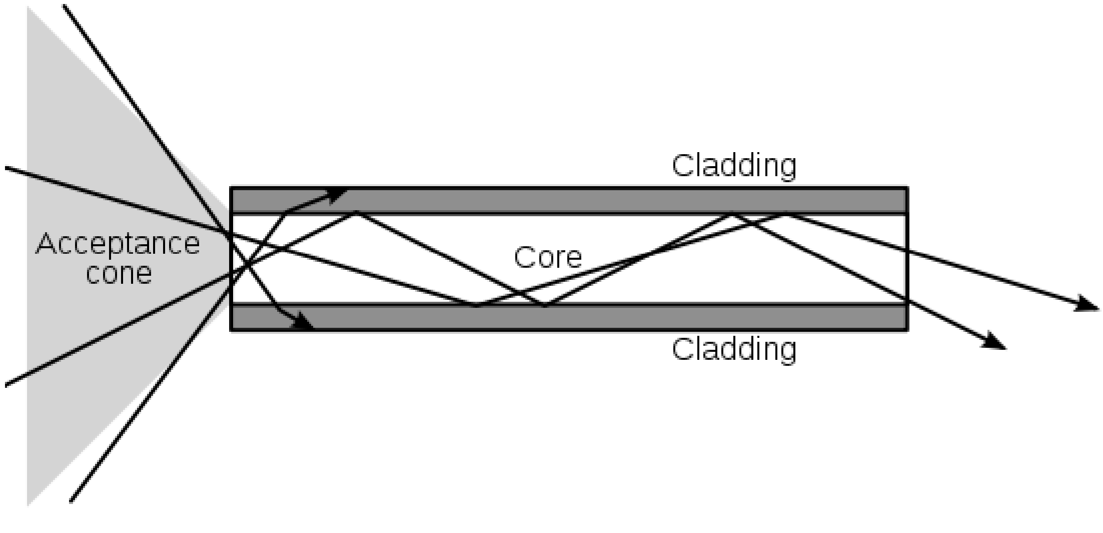
The index of refraction is a way of measuring the speed of light in a material. Light travels fastest in a vacuum, such as outer space. The speed of light in a vacuum is about 300,000 kilometres (186 thousand miles) per second. Index of refraction is calculated by dividing the speed of light in a vacuum by the speed of light in some other medium. The index of refraction of a vacuum is therefore 1, by definition. The typical value for the cladding of an optical fiber is 1.46. The core value is typically 1.48. The larger the index of refraction, the slower light travels in that medium. From this information, a good rule of thumb is that signal using optical fiber for communication will travel at around 200 million meters per second. Or to put it another way, to travel 1000 kilometers in fiber, the signal will take 5 milliseconds to propagate. Thus a phone call carried by fiber between Sydney and New York, a 12000 kilometer distance, means that there is an absolute minimum delay of 60 milliseconds (or around 1/16th of a second) between when one caller speaks to when the other hears. (Of course the fiber in this case will probably travel a longer route, and there will be additional delays due to communication equipment switching and the process of encoding and decoding the voice onto the fiber).

Total internal reflection

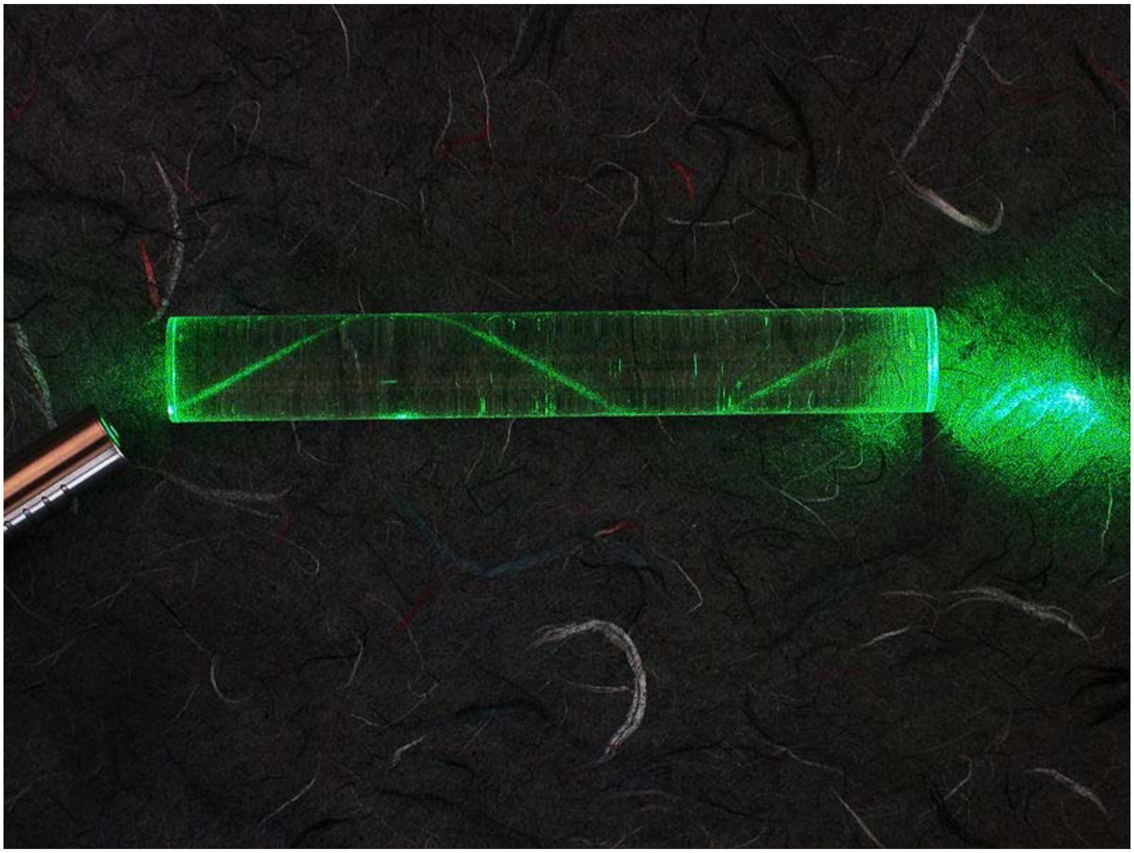
When light traveling in a dense medium hits a boundary at a steep angle (larger than the "critical angle" for the boundary), the light will be completely reflected. This effect is used in optical fibers to confine light in the core. Light travels along the fiber bouncing back and forth off of the boundary. Because the light must strike the boundary with an angle greater than the critical angle, only light that enters the fiber within a certain range of angles can travel down the fiber without leaking out. This range of angles is called the acceptance cone of the fiber. The size of this acceptance cone is a function of the refractive index difference between the fiber's core and cladding.

In simpler terms, there is a maximum angle from the fiber axis at which light may enter the fiber so that it will propagate, or travel, in the core of the fiber. The sine of this maximum angle is the numerical aperture (NA) of the fiber. Fiber with a larger NA requires less precision to splice and work with than fiber with a smaller NA. Single-mode fiber has a small NA.

Multi-mode fiber

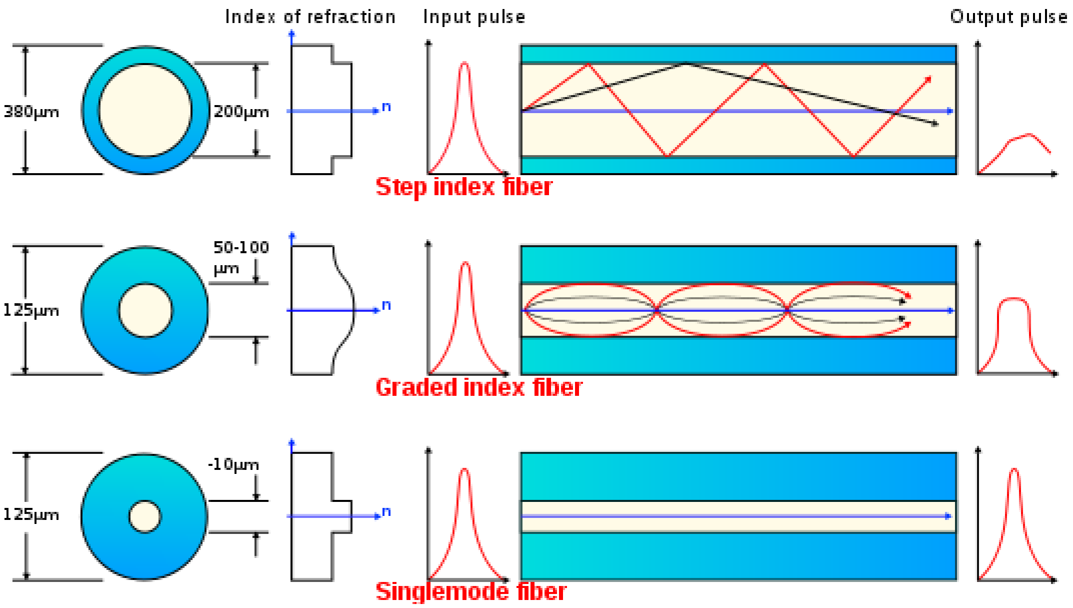


The propagation of light through a multi-mode optical fiber.



A laser bouncing down an acrylic rod, illustrating the total internal reflection of light in a multi-mode optical fiber.

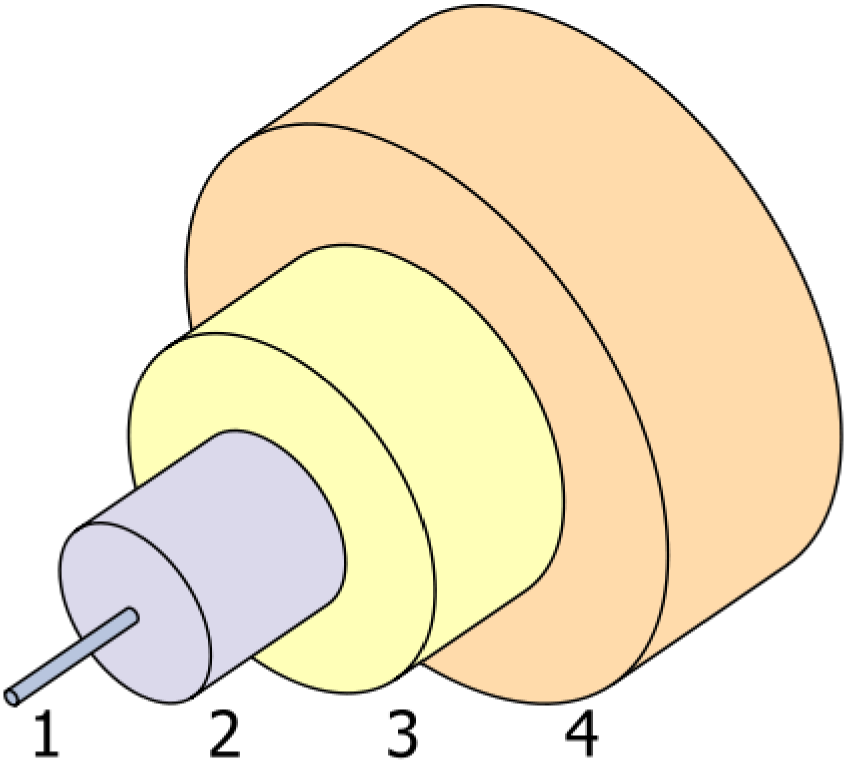
Fiber with large core diameter (greater than 10 micrometers) may be analyzed by geometrical optics. Such fiber is called *multi-mode fiber*, from the electromagnetic analysis (see below). In a step-index multi-mode fiber, rays of light are guided along the fiber core by total internal reflection. Rays that meet the core-cladding boundary at a high angle (measured relative to a line normal to the boundary), greater than the critical angle for this boundary, are completely reflected. The critical angle (minimum angle for total internal reflection) is determined by the difference in index of refraction between the core and cladding materials. Rays that meet the boundary at a low angle are refracted from the core into the cladding, and do not convey light and hence information along the fiber. The critical angle determines the acceptance angle of the fiber, often reported as a numerical aperture. A high numerical aperture allows light to propagate down the fiber in rays both close to the axis and at various angles, allowing efficient coupling of light into the fiber. However, this high numerical aperture increases the amount of dispersion as rays at different angles have different path lengths and therefore take different times to traverse the fiber.



Optical fiber types.

In graded-index fiber, the index of refraction in the core decreases continuously between the axis and the cladding. This causes light rays to bend smoothly as they approach the cladding, rather than reflecting abruptly from the core-cladding boundary. The resulting curved paths reduce multi-path dispersion because high angle rays pass more through the lower-index periphery of the core, rather than the high-index center. The index profile is chosen to minimize the difference in axial propagation speeds of the various rays in the fiber. This ideal index profile is very close to a parabolic relationship between the index and the distance from the axis.

Single-mode fiber



The structure of a typical single-mode fiber.
1. Core: 8 µm diameter
2. Cladding: 125 µm dia.
3. Buffer: 250 µm dia.
4. Jacket: 400 µm dia.

Fiber with a core diameter less than about ten times the wavelength of the propagating light cannot be modeled using geometric optics. Instead, it must be analyzed as an electromagnetic structure, by solution of Maxwell's equations as reduced to the electromagnetic wave equation. The electromagnetic analysis may also be required to understand behaviors such as speckle that occur when coherent light propagates in multi-mode fiber. As an optical waveguide, the fiber supports one or more confined transverse modes by which light can propagate along the fiber. Fiber supporting only one mode is called *single-mode* or *mono-mode fiber*. The behavior of larger-core multi-mode fiber can also be modeled using the wave equation, which shows that such fiber supports more than one mode of propagation (hence the name). The results of such modeling of multi-mode fiber approximately agree with the predictions of geometric optics, if the fiber core is large enough to support more than a few modes.

The waveguide analysis shows that the light energy in the fiber is not completely confined in the core. Instead, especially in single-mode fibers, a significant fraction of the energy in the bound mode travels in the cladding as an evanescent wave.

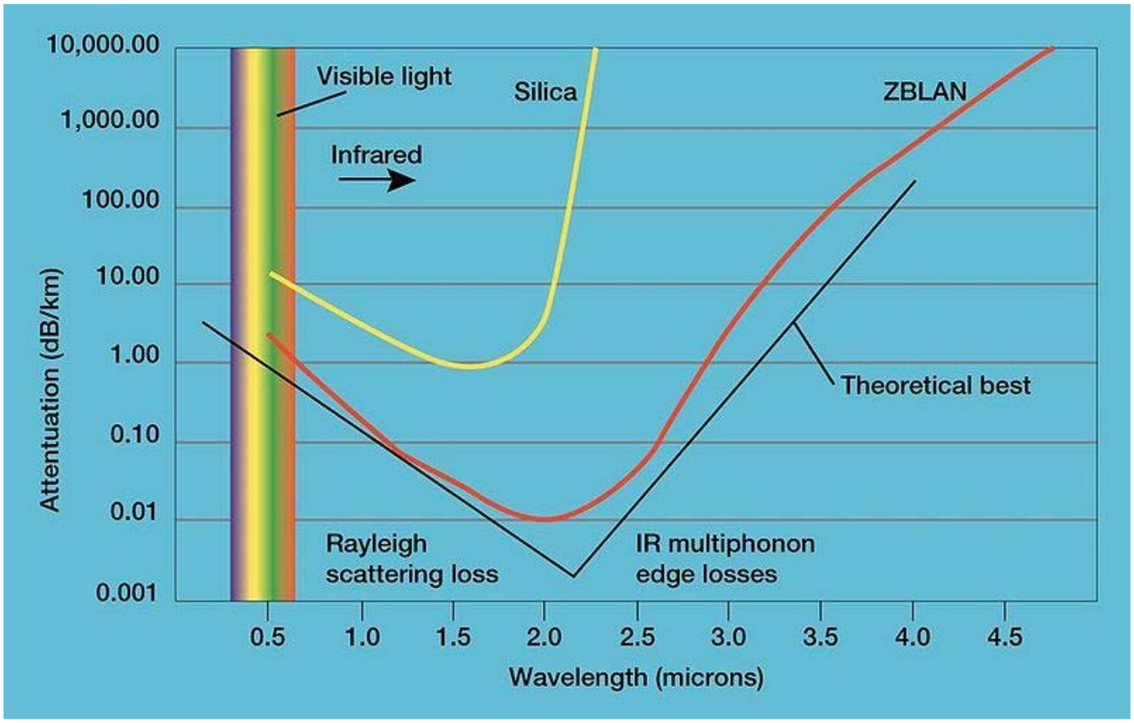
The most common type of single-mode fiber has a core diameter of 8–10 micrometers and is designed for use in the near infrared. The mode structure depends on the wavelength of the light used, so that this fiber actually supports a small number of additional modes at visible wavelengths. Multi-mode fiber, by comparison, is manufactured with core diameters as small as 50 micrometers and as large as hundreds of micrometers. The normalized frequency *V* for this fiber should be less than the first zero of the Bessel function J_0 (approximately 2.405).

Special-purpose fiber

Some special-purpose optical fiber is constructed with a non-cylindrical core and/or cladding layer, usually with an elliptical or rectangular cross-section. These include polarization-maintaining fiber and fiber designed to suppress whispering gallery mode propagation.

Photonic-crystal fiber is made with a regular pattern of index variation (often in the form of cylindrical holes that run along the length of the fiber). Such fiber uses diffraction effects instead of or in addition to total internal reflection, to confine light to the fiber's core. The properties of the fiber can be tailored to a wide variety of applications.

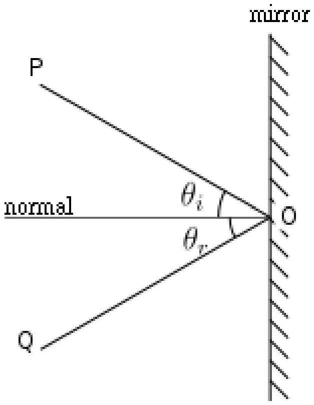
Mechanisms of attenuation



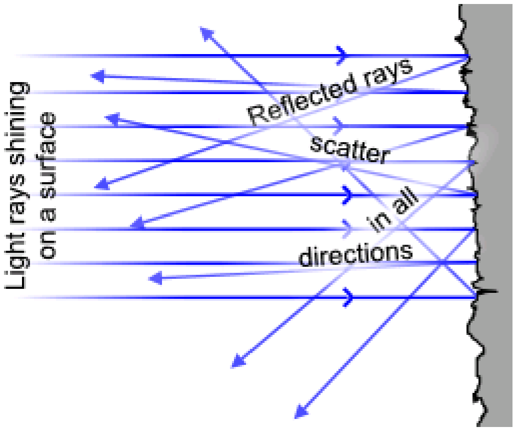
Light attenuation by ZBLAN and silica fibers

Attenuation in fiber optics, also known as transmission loss, is the reduction in intensity of the light beam (or signal) with respect to distance traveled through a transmission medium. Attenuation coefficients in fiber optics usually use units of dB/km through the medium due to the relatively high quality of transparency of modern optical transmission media. The medium is usually a fiber of silica glass that confines the incident light beam to the inside. Attenuation is an important factor limiting the transmission of a digital signal across large distances. Thus, much research has gone into both limiting the attenuation and maximizing the amplification of the optical signal. Empirical research has shown that attenuation in optical fiber is caused primarily by both scattering and absorption.

Light scattering



Specular reflection



Diffuse reflection

The propagation of light through the core of an optical fiber is based on total internal reflection of the lightwave. Rough and irregular surfaces, even at the molecular level, can cause light rays to be reflected in random directions. This is called diffuse reflection or scattering, and it is typically characterized by wide variety of reflection angles.

Light scattering depends on the wavelength of the light being scattered. Thus, limits to spatial scales of visibility arise, depending on the frequency of the incident light-wave and the physical dimension (or spatial scale) of the scattering center, which is typically in the form of some specific micro-structural feature. Since visible light has a wavelength of the order of one micrometre (one millionth of a meter) scattering centers will have dimensions on a similar spatial scale.

Thus, attenuation results from the incoherent scattering of light at internal surfaces and interfaces. In (poly)crystalline materials such as metals and ceramics, in addition to pores, most of the internal surfaces or interfaces are in the form of grain boundaries that separate tiny regions of crystalline order. It has recently been shown that when the size of the scattering center (or grain boundary) is reduced below the size of the wavelength of the light being scattered, the scattering no longer occurs to any significant extent. This phenomenon has given rise to the production of transparent ceramic materials.

Similarly, the scattering of light in optical quality glass fiber is caused by molecular level irregularities (compositional fluctuations) in the glass structure. Indeed, one emerging school of thought is that a glass is simply the limiting case of a polycrystalline solid. Within this framework, "domains" exhibiting various degrees of short-range order become the building blocks of both metals and alloys, as well as glasses and ceramics. Distributed both between and within these domains are micro-structural defects which will provide the most ideal locations for the occurrence of light scattering. This same phenomenon is seen as one of the limiting factors in the transparency of IR missile domes.

At high optical powers, scattering can also be caused by nonlinear optical processes in the fiber.

UV-Vis-IR absorption

In addition to light scattering, attenuation or signal loss can also occur due to selective absorption of specific wavelengths, in a manner similar to that responsible for the appearance of color. Primary material considerations include both electrons and molecules as follows:

1) At the electronic level, it depends on whether the electron orbitals are spaced (or "quantized") such that they can absorb a quantum of light (or photon) of a specific wavelength or frequency in the ultraviolet (UV) or visible ranges. This is what gives rise to color.

2) At the atomic or molecular level, it depends on the frequencies of atomic or molecular vibrations or chemical bonds, how close-packed its atoms or molecules are, and whether or not the atoms or molecules exhibit long-range order. These factors will determine the capacity of the material transmitting longer wavelengths in the infrared (IR), far IR, radio and microwave ranges.

The design of any optically transparent device requires the selection of materials based upon knowledge of its properties and limitations. The lattice absorption characteristics observed at the lower frequency regions (mid IR to far-infrared wavelength range) define the long-wavelength transparency limit of the material. They are the result of the interactive coupling between the motions of thermally induced vibrations of the constituent atoms and molecules of the solid lattice and the incident light wave radiation. Hence, all materials are bounded by limiting regions of absorption caused by atomic and molecular vibrations (bond-stretching)in the far-infrared (>10 μm).

Thus, multi-phonon absorption occurs when two or more phonons simultaneously interact to produce electric dipole moments with which the incident radiation may couple. These dipoles can absorb energy from the incident radiation, reaching a maximum coupling with the radiation when the frequency is equal to the fundamental vibrational mode of the molecular dipole (e.g. Si-O bond) in the far-infrared, or one of its harmonics.

The selective absorption of infrared (IR) light by a particular material occurs because the selected frequency of the light wave matches the frequency (or an integer multiple of the frequency) at which the particles of that material vibrate. Since different atoms and molecules have different natural frequencies of vibration, they will selectively absorb different frequencies (or portions of the spectrum) of infrared (IR) light.

Reflection and transmission of light waves occur because the frequencies of the light waves do not match the natural resonant frequencies of vibration of the objects. When IR light of these frequencies strikes an object, the energy is either reflected or transmitted.

Manufacturing

Materials

Glass optical fibers are almost always made from silica, but some other materials, such as fluorozirconate, fluoroaluminate, and chalcogenide glasses as well as crystalline materials like sapphire, are used for longer-wavelength infrared or other specialized applications. Silica and fluoride glasses usually have refractive indices of about 1.5, but some materials such as the chalcogenides can have indices as high as 3. Typically the index difference between core and cladding is less than one percent.

Plastic optical fibers (POF) are commonly step-index multi-mode fibers with a core diameter of 0.5 millimeters or larger. POF typically have higher attenuation coefficients than glass fibers, 1 dB/m or higher, and this high attenuation limits the range of POF-based systems.

Silica

Silica exhibits fairly good optical transmission over a wide range of wavelengths. In the near-infrared (near IR) portion of the spectrum, particularly around 1.5 μm, silica can have extremely low absorption and scattering losses of the order of 0.2 dB/km. A high transparency in the 1.4-μm region is achieved by maintaining a low concentration of hydroxyl groups (OH). Alternatively, a high OH concentration is better for transmission in the ultraviolet (UV) region.

Silica can be drawn into fibers at reasonably high temperatures, and has a fairly broad glass transformation range. One other advantage is that fusion splicing and cleaving of silica fibers is relatively effective. Silica fiber also has high mechanical strength against both pulling and even bending, provided that the fiber is not too thick and that the surfaces have been well prepared during processing. Even simple cleaving (breaking) of the ends of the fiber can provide nicely flat surfaces with acceptable optical quality. Silica is also relatively chemically inert. In particular, it is not hygroscopic (does not absorb water).

Silica glass can be doped with various materials. One purpose of doping is to raise the refractive index (e.g. with Germanium dioxide (GeO_2) or Aluminium oxide (Al_2O_3)) or to lower it (e.g. with fluorine or Boron trioxide (B_2O_3)). Doping is also possible with laser-active ions (for example, rare earth-doped fibers) in order to obtain active fibers to be used, for example, in fiber amplifiers or laser applications. Both the fiber core and cladding are typically doped, so that the entire assembly (core and cladding) is effectively the same compound (e.g. an aluminosilicate, germanosilicate, phosphosilicate or borosilicate glass).

Particularly for active fibers, pure silica is usually not a very suitable host glass, because it exhibits a low solubility for rare earth ions. This can lead to quenching effects due to clustering of dopant ions. Aluminosilicates are much more effective in this respect.

Silica fiber also exhibits a high threshold for optical damage. This property ensures a low tendency for laser-induced breakdown. This is important for fiber amplifiers when utilized for the amplification of short pulses.

Because of these properties silica fibers are the material of choice in many optical applications, such as communications (except for very short distances with plastic optical fiber), fiber lasers, fiber amplifiers, and fiber-optic sensors. The large efforts which have been put forth in the development of various types of silica fibers have further increased the performance of such fibers over other materials.

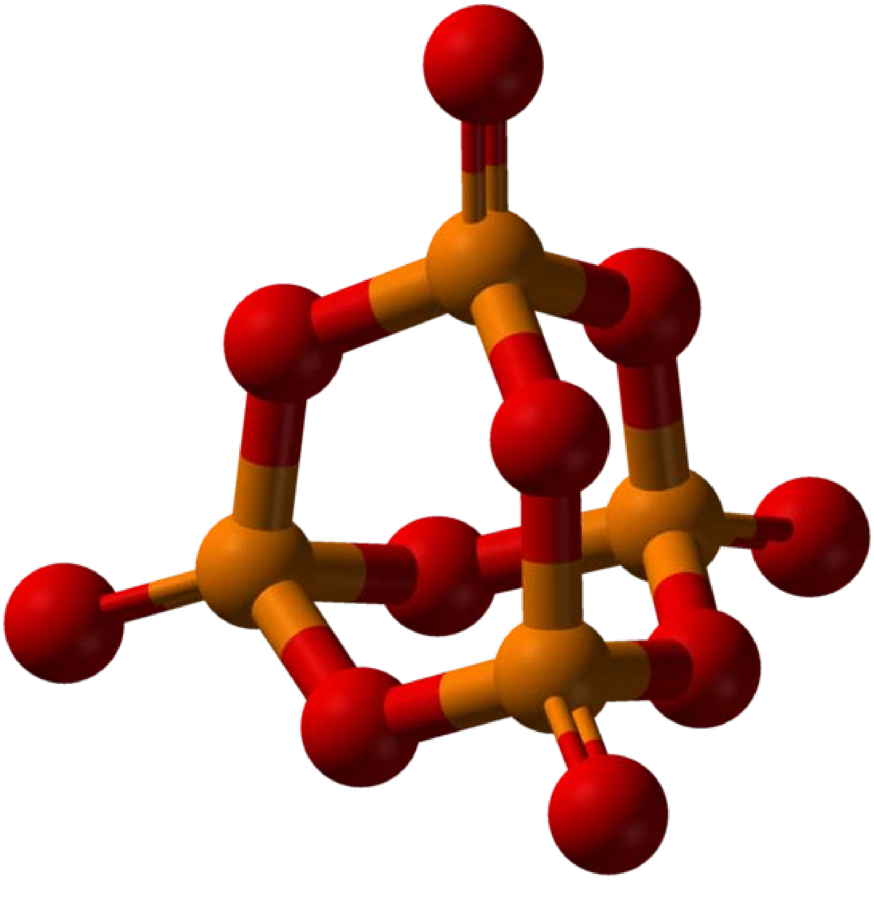
Fluorides

Fluoride glass is a class of non-oxide optical quality glasses composed of fluorides of various metals. Because of their low viscosity, it is very difficult to completely avoid crystallization while processing it through the glass transition (or drawing the fiber from the melt). Thus, although heavy metal fluoride glasses (HMFG) exhibit very low optical attenuation, they are not only difficult to manufacture, but are quite fragile, and have poor resistance to moisture and other environmental attacks. Their best attribute is that they lack the absorption band associated with the hydroxyl (OH) group (3200–3600 cm^{-1}), which is present in nearly all oxide-based glasses.

An example of a heavy metal fluoride glass is the ZBLAN glass group, composed of zirconium, barium, lanthanum, aluminium, and sodium fluorides. Their main technological application is as optical waveguides in both planar and fiber form. They are advantageous especially in the mid-infrared (2000–5000 nm) range.

HMFGs were initially slated for optical fiber applications, because the intrinsic losses of a mid-IR fiber could in principle be lower than those of silica fibers, which are transparent only up to about 2 μm. However, such low losses were never realized in practice, and the fragility and high cost of fluoride fibers made them less than ideal as primary candidates. Later, the utility of fluoride fibers for various other applications was discovered. These include mid-IR spectroscopy, fiber optic sensors, thermometry, and imaging. Also, fluoride fibers can be used for guided lightwave transmission in media such as YAG (yttria-alumina garnet) lasers at 2.9 μm, as required for medical applications (e.g. ophthalmology and dentistry).

Phosphates



The P_4O_{10} cagelike structure—the basic building block for phosphate glass.

Phosphate glass constitutes a class of optical glasses composed of metaphosphates of various metals. Instead of the SiO_4 tetrahedra observed in silicate glasses, the building block for this glass former is Phosphorus pentoxide (P_2O_5), which crystallizes in at least four different forms. The most familiar polymorph (see figure) comprises molecules of P_4O_{10}.

Phosphate glasses can be advantageous over silica glasses for optical fibers with a high concentration of doping rare earth ions. A mix of fluoride glass and phosphate glass is fluorophosphate glass.

Chalcogenides

The chalcogens—the elements in group 16 of the periodic table—particularly sulfur (S), selenium (Se) and tellurium (Te)—react with more electropositive elements, such as silver, to form chalcogenides. These are extremely versatile compounds, in that they can be crystalline or amorphous, metallic or semiconducting, and conductors of ions or electrons.

Process

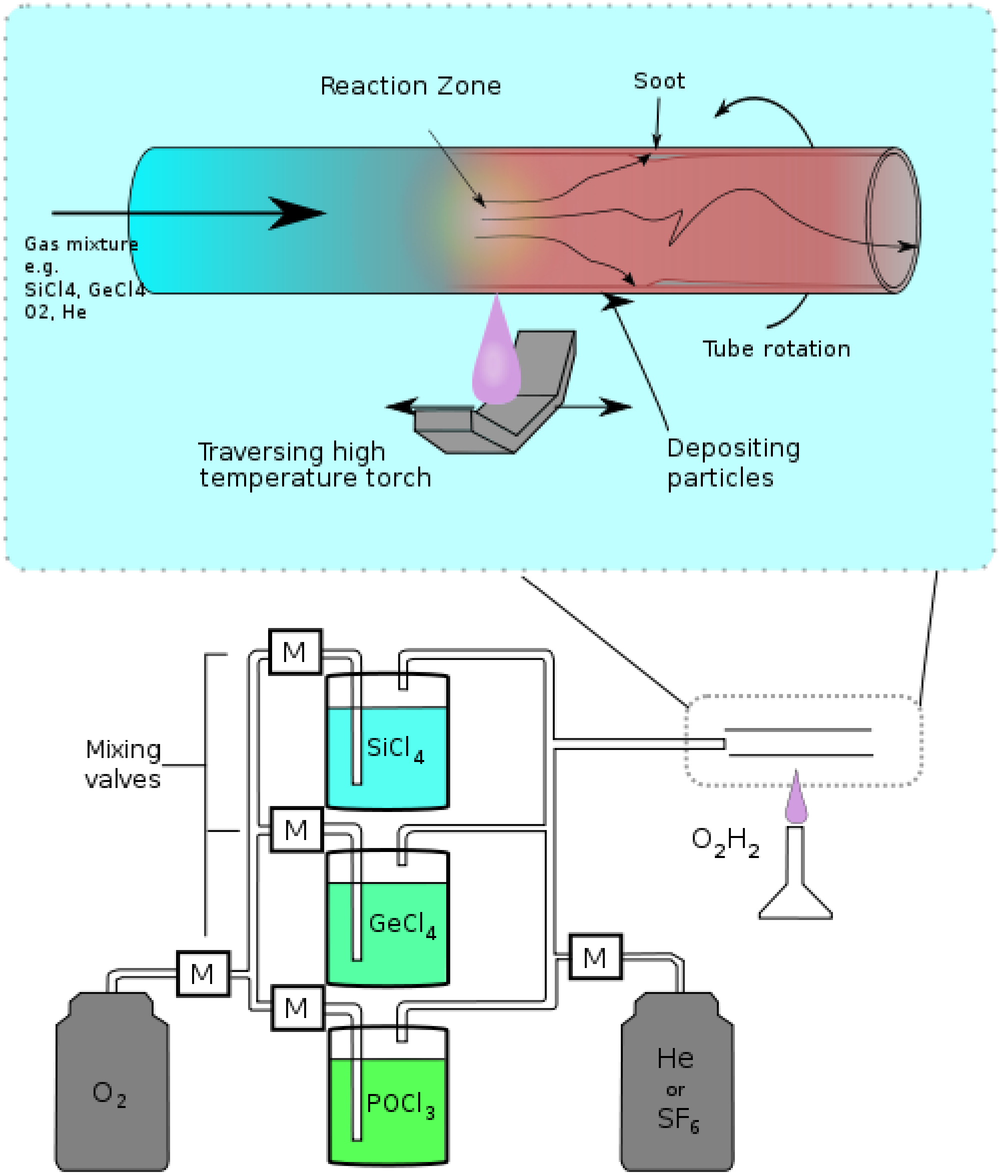


Illustration of the modified chemical vapor deposition (inside) process

Standard optical fibers are made by first constructing a large-diameter *preform*, with a carefully controlled refractive index profile, and then *pulling* the preform to form the long, thin optical fiber. The preform is commonly made by three chemical vapor deposition methods: *inside vapor deposition*, *outside vapor deposition*, and *vapor axial deposition*.

With *inside vapor deposition*, the preform starts as a hollow glass tube approximately 40 centimeters (16 in) long, which is placed horizontally and rotated slowly on a lathe. Gases such as silicon tetrachloride ($SiCl_4$) or germanium tetrachloride ($GeCl_4$) are injected with oxygen in the end of the tube. The gases are then heated by means of an external hydrogen burner, bringing the temperature of the gas up to 1900 K (1600 °C, 3000 °F), where the tetrachlorides react with oxygen to produce silica or germania (germanium dioxide) particles. When the reaction conditions are chosen to allow this reaction to occur in the gas phase throughout the tube volume, in contrast to earlier techniques where the reaction occurred only on the glass surface, this technique is called *modified chemical vapor deposition (MCVD)*.

The oxide particles then agglomerate to form large particle chains, which subsequently deposit on the walls of the tube as soot. The deposition is due to the large difference in temperature between the gas core and the wall causing the gas to push the particles outwards (this is known as thermophoresis). The torch is then traversed up and down the length of the tube to deposit the material evenly. After the torch has reached the end of the tube, it is then brought back to the beginning of the tube and the deposited particles are then melted to form a solid layer. This process is repeated until a sufficient amount of material has been deposited. For each layer the composition can be modified by varying the gas composition, resulting in precise control of the finished fiber's optical properties.

In outside vapor deposition or vapor axial deposition, the glass is formed by *flame hydrolysis*, a reaction in which silicon tetrachloride and germanium tetrachloride are oxidized by reaction with water (H_2O) in an oxyhydrogen flame. In outside vapor deposition the glass is deposited onto a solid rod, which is removed before further processing. In vapor axial deposition, a short *seed rod* is used, and a porous preform, whose length is not limited by the size of the source rod, is built up on its end. The porous preform is consolidated into a transparent, solid preform by heating to about 1800 K (1500 °C, 2800 °F).

The preform, however constructed, is then placed in a device known as a drawing tower, where the preform tip is heated and the optic fiber is pulled out as a string. By measuring the resultant fiber width, the tension on the fiber can be controlled to maintain the fiber thickness.

Coatings

The light is "guided" down the core of the fiber by an optical "cladding" with a lower refractive index that traps light in the core through "total internal reflection."

The cladding is coated by a "buffer" that protects it from moisture and physical damage. The buffer is what gets stripped off the fiber for termination or splicing. These coatings are UV-cured urethane acrylate composite materials applied to the outside of the fiber during the drawing process. The coatings protect the very delicate strands of glass fiber—about the size of a human hair—and allow it to survive the rigors of manufacturing, proof testing, cabling and installation.

Today’s glass optical fiber draw processes employ a dual-layer coating approach. An inner primary coating is designed to act as a shock absorber to minimize attenuation caused by microbending. An outer secondary coating protects the primary coating against mechanical damage and acts as a barrier to lateral forces. Sometimes a metallic armour layer is added to provide extra protection.

These fiber optic coating layers are applied during the fiber draw, at speeds approaching 100 kilometers per hour (60 mph). Fiber optic coatings are applied using one of two methods: wet-on-dry, in which the fiber passes through a primary coating application, which is then UV cured, then through the secondary coating application which is subsequently cured; and wet-on-wet, in which the fiber passes through both the primary and secondary coating applications and then goes to UV curing.

Fiber optic coatings are applied in concentric layers to prevent damage to the fiber during the drawing application and to maximize fiber strength and microbend resistance. Unevenly coated fiber will experience non-uniform forces when the coating expands or contracts, and is susceptible to greater signal attenuation. Under proper drawing and coating processes, the coatings are concentric around the fiber, continuous over the length of the application and have constant thickness.

Fiber optic coatings protect the glass fibers from scratches that could lead to strength degradation. The combination of moisture and scratches accelerates the aging and deterioration of fiber strength. When fiber is subjected to low stresses over a long period, fiber fatigue can occur. Over time or in extreme conditions, these factors combine to cause microscopic flaws in the glass fiber to propagate, which can ultimately result in fiber failure.

Three key characteristics of fiber optic waveguides can be affected by environmental conditions: strength, attenuation and resistance to losses caused by microbending. External fiber optic coatings protect glass optical fiber from environmental conditions that can affect the fiber’s performance and long-term durability. On the inside, coatings ensure the reliability of the signal being carried and help minimize attenuation due to microbending.

Practical issues

Optical fiber cables

An optical fiber cable

In practical fibers, the cladding is usually coated with a tough resin *buffer* layer, which may be further surrounded by a *jacket* layer, usually glass. These layers add strength to the fiber but do not contribute to its optical wave guide properties. Rigid fiber assemblies sometimes put light-absorbing ("dark") glass between the fibers, to prevent light that leaks out of one fiber from entering another. This reduces cross-talk between the fibers, or reduces flare in fiber bundle imaging applications.

Modern cables come in a wide variety of sheathings and armor, designed for applications such as direct burial in trenches, high voltage isolation, dual use as power lines, installation in conduit, lashing to aerial telephone poles, submarine installation, and insertion in paved streets. The cost of small fiber-count pole-mounted cables has greatly decreased due to the high demand for fiber to the home (FTTH) installations in Japan and South Korea.

Fiber cable can be very flexible, but traditional fiber's loss increases greatly if the fiber is bent with a radius smaller than around 30 mm. This creates a problem when the cable is bent around corners or wound around a spool, making FTTX installations more complicated. "Bendable fibers", targeted towards easier installation in home environments, have been standardized as ITU-T G.657. This type of fiber can be bent with a radius as low as 7.5 mm without adverse impact. Even more bendable fibers have been developed. Bendable fiber may also be resistant to fiber hacking, in which the signal in a fiber is surreptitiously monitored by bending the fiber and detecting the leakage.

Another important feature of cable is cable withstanding against the horizontally applied force. It is technically called max tensile strength defining how much force can applied to the cable during the installation period.

Telecom Anatolia fiber optic cable versions are reinforced with aramid yarns or glass yarns as intermediary strength member. In commercial terms, usage of the glass yarns are more cost effective while no loss in mechanical durability of the cable. Glass yarns also protect the cable core against rodents and termites.

Termination and splicing

ST connectors on multi-mode fiber.

Optical fibers are connected to terminal equipment by optical fiber connectors. These connectors are usually of a standard type such as *FC*, *SC*, *ST*, *LC*, or *MTRJ*.

Optical fibers may be connected to each other by connectors or by *splicing*, that is, joining two fibers together to form a continuous optical waveguide. The generally accepted splicing method is arc fusion splicing, which melts the fiber ends together with an electric arc. For quicker fastening jobs, a "mechanical splice" is used.

Fusion splicing is done with a specialized instrument that typically operates as follows: The two cable ends are fastened inside a splice enclosure that will protect the splices, and the fiber ends are stripped of their protective polymer coating (as well as the more sturdy outer jacket, if present). The ends are *cleaved* (cut) with a precision cleaver to make them perpendicular, and are placed into special holders in the splicer. The splice is usually inspected via a magnified viewing screen to check the cleaves before and after the splice. The splicer uses small motors to align the end faces together, and emits a small spark between electrodes at the gap to burn off dust and moisture. Then the splicer generates a larger spark that raises the temperature above the melting point of the glass, fusing the ends together permanently. The location and energy of the spark is carefully controlled so that the molten core and cladding do not mix, and this minimizes optical loss. A splice loss estimate is measured by the splicer, by directing light through the cladding on one side and measuring the light leaking from the cladding on the other side. A splice loss under 0.1 dB is typical. The complexity of this process makes fiber splicing much more difficult than splicing copper wire.

Mechanical fiber splices are designed to be quicker and easier to install, but there is still the need for stripping, careful cleaning and precision cleaving. The fiber ends are aligned and held together by a precision-made sleeve, often using a clear index-matching gel that enhances the transmission of light across the joint. Such joints typically have higher optical loss and are less robust than fusion splices, especially if the gel is used. All splicing techniques involve the use of an enclosure into which the splice is placed for protection afterward.

Fibers are terminated in connectors so that the fiber end is held at the end face precisely and securely. A fiber-optic connector is basically a rigid cylindrical barrel surrounded by a sleeve that holds the barrel in its mating socket. The mating mechanism can be "push and click", "turn and latch" ("bayonet"), or screw-in (threaded). A typical connector is installed by preparing the fiber end and inserting it into the rear of the connector body. Quick-set adhesive is usually used so the fiber is held securely, and a strain relief is secured to the rear. Once the adhesive has set, the fiber's end is polished to a mirror finish. Various polish profiles are used, depending on the type of fiber and the application. For single-mode fiber, the fiber ends are typically polished with a slight curvature, such that when the connectors are mated the fibers touch only at their cores. This is known as a "physical contact" (PC) polish. The curved surface may be polished at an angle, to make an "angled physical contact" (APC) connection. Such connections have higher loss than PC connections, but greatly reduced back reflection, because light that reflects from the angled surface leaks out of the fiber core; the resulting loss in signal strength is known as gap loss. APC fiber ends have low back reflection even when disconnected.

In the 1990s, terminating fiber optic cables was very labor intensive. The number of parts per connector, polishing of the fibers, and the need to oven-bake the epoxy in each connector made terminating fiber optic cables very difficult. Today, many different connectors are on the market and offer an easier, less labor intensive way of terminating the cables. Some of the most popular connectors have already been polished from the factory and include a gel inside the connector and those two steps help save money on labor especially on large projects. A cleave is made at a required length in order to get as close to the polished piece already inside the connector, with the gel surrounding the point where the two piece meet inside the connector very little light loss is exposed.

Free-space coupling

It is often necessary to align an optical fiber with another optical fiber, or with an optoelectronic device such as a light-emitting diode, a laser diode, or a modulator. This can involve either carefully aligning the fiber and placing it in contact with the device, or can use a lens to allow coupling over an air gap. In some cases the end of the fiber is polished into a curved form that is designed to allow it to act as a lens.

In a laboratory environment, a bare fiber end is coupled using a fiber launch system, which uses a microscope objective lens to focus the light down to a fine point. A precision translation stage (micro-positioning table) is used to move the lens, fiber, or device to allow the coupling efficiency to be optimized. Fibers with a connector on the end make this process much simpler: the connector is simply plugged into a pre-aligned fiberoptic collimator, which contains a lens that is either accurately positioned with respect to the fiber, or is adjustable. To achieve the best injection efficiency into single-mode fiber, the direction, position, size and divergence of the beam must all be optimized. With good beams, 70 to 90% coupling efficiency can be achieved.

With properly polished single-mode fibers, the emitted beam has an almost perfect Gaussian shape—even in the far field—if a good lens is used. The lens needs to be large enough to support the full numerical aperture of the fiber, and must not introduce aberrations in the beam. Aspheric lenses are typically used.

Fiber fuse

At high optical intensities, above 2 megawatts per square centimeter, when a fiber is subjected to a shock or is otherwise suddenly damaged, a *fiber fuse* can occur. The reflection from the damage vaporizes the fiber immediately before the break, and this new defect remains reflective so that the damage propagates back toward the transmitter at 1–3 meters per second (4−11 km/h, 2–8 mph). The open fiber control system, which ensures laser eye safety in the event of a broken fiber, can also effectively halt propagation of the fiber fuse. In situations, such as undersea cables, where high power levels might be used without the need for open fiber control, a "fiber fuse" protection device at the transmitter can break the circuit to prevent any damage.

Example

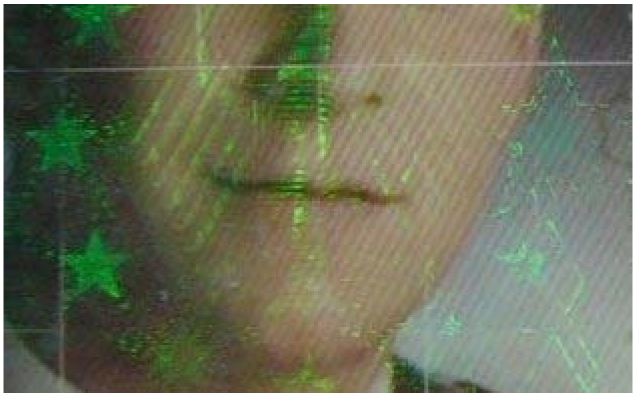
Fiber connections can be used for various types of connections. For example, most high definition televisions offer a digital audio optical connection. This allows the streaming of audio over light, using the TOSLink protocol.

Electric power transmission

Optical fiber can be used to transmit electricity. While the efficiency is not nearly that of traditional copper wire, it is especially useful in situations where it is desirable to not have a metallic conductor as in the case of use near MRI machines which produce strong magnetic currents.

Chapter 9

Holography



Identigram as a security element in a German identity card

Holography is a technique that allows the light scattered from an object to be recorded and later reconstructed so that it appears as if the object is in the same position relative to the recording medium as it was when recorded. The image changes as the position and orientation of the viewing system changes in exactly the same way as if the object were still present, thus making the recorded image (**hologram**) appear three-dimensional.

The technique of holography can also be used to store, retrieve, and process information optically. While it has been possible to create a 3-D holographic picture of a static object since the 1960s, it is only in the last few years that arbitrary scenes or videos can be shown on a holographic volumetric display.

Overview and history

Hologram artwork in MIT Museum

Holography was invented in 1947 by the Hungarian-British physicist Dennis Gabor (Hungarian name: Gábor Dénes), work for which he received the Nobel Prize in Physics in 1971. Pioneering work in the field of physics by other scientists including Mieczysław Wolfke resolved technical issues that previously had prevented advancement. The discovery was an unexpected result of research into improving electron microscopes at the British Thomson-Houston Company in Rugby, England, and the company filed a patent in December 1947 (patent GB685286). The technique as originally invented is still used in electron microscopy, where it is known as electron holography, but holography as a light-optical technique did not really advance until the development of the laser in 1960.

The first practical optical holograms that recorded 3D objects were made in 1962 by Yuri Denisyuk in the Soviet Union and by Emmett Leith and Juris Upatnieks at University of Michigan, USA. Advances in photochemical processing techniques to produce high-quality display holograms were achieved by Nicholas J. Phillips.

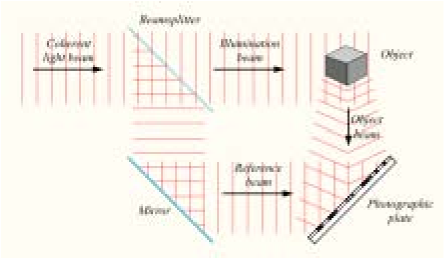
Several types of holograms can be made. Transmission holograms, such as those produced by Leith and Upatnieks, are viewed by shining laser light through them and looking at the reconstructed image from the side of the hologram opposite the source. A later refinement, the "rainbow transmission" hologram, allows more convenient illumination by white light rather than by lasers. Rainbow holograms are commonly seen today on credit cards as a security feature and on product packaging. These versions of the rainbow transmission hologram are commonly formed as surface relief patterns in a plastic film, and they incorporate a reflective aluminum coating that provides the light from "behind" to reconstruct their imagery.

Another kind of common hologram, the reflection or Denisyuk hologram, is capable of multicolour-image reproduction, using a white-light illumination source on the same side of the hologram as the viewer.

Specular holography is a related technique for making three-dimensional imagery by controlling the motion of specularities on a two-dimensional surface. It works by reflectively or refractively manipulating bundles of light rays, whereas Gabor-style holography works by diffractively reconstructing wavefronts.

One of the most promising recent advances in the short history of holography has been the mass production of low-cost solid-state lasers, such as those found in millions of DVD recorders and used in other common applications, which are sometimes also useful for holography. These cheap, compact, solid-state lasers can, under some circumstances, compete well with the large, expensive gas lasers previously required to make holograms and are already helping to make holography much more accessible to low-budget researchers, artists and dedicated hobbyists.

Theory



Holographic recording process

Though holography is often referred to as 3D photography, this is a misconception. A better analogy is sound recording where the sound field is encoded in such a way that it can later be reproduced. In holography, some of the light scattered from an object or a set of objects falls on the recording medium. A second light beam, known as the reference beam, also illuminates the recording medium, so that interference occurs between the two beams. The resulting light field generates a seemingly random pattern of varying intensity, which is recorded in the hologram. It can be shown that if the hologram is illuminated by the original reference beam, the reference beam is diffracted by the hologram to produce a diffracted light field which is identical to the light field which was scattered by the object or objects. Thus, someone looking into the hologram "sees" the objects even though they are no longer present. There are a variety of recording materials which can be used, including photographic film.

The first cameras used pinhole lenses. They consisted of a completely blacked-out box with a tiny pinhole on the side away from the film or screen. As a result, they only caught the scene before them from a single, tiny vantage point. The glass lenses that followed, were, in effect, simply giant pinholes, with all the light they collected being passed through a tiny point—a pinhole as it were—at the focal point of the glass lens before spreading out again before hitting the film or screen behind the lens.

The problem Dennis Gabor, the inventor of holography, set out to solve was how to take a picture of all the light passing through a large window, rather than just the light passing through one tiny pinhole. The person looking through this captured "window" would see the image in 3D by virtue of each of his or her eyes seeing the scene from a different viewpoint. Further, the person would be able to move his or her head around to the extent the window would allow to see the object from a variety of vantage points. (An early hologram from the 1960s featured an object with a glass magnifying lens mounted a few inches/centimeters in front of it.)

Dennis Gabor, in effect, needed a fast shutter, one so fast that it could "freeze" all the light waves at their current phase just as they were passing through the window, a shutter that moved at the speed of light. His successful approach worked in a way analogous to the way a strobe light is used to "freeze" the motion of rapidly moving mechanical equipment, such as engines: If the light comes on at exactly the same moment during every rotation of a piece of equipment, the rotating part(s) will appear to be standing still.

The function analogous to a strobe light, in holography, is performed by the "reference beam". In the illustration above, a portion of the light from the laser, the "illumination beam", is aimed at the scene, where it bounces off the objects in the scene directly onto the film—the window—with no pinhole or lens interposed. Another portion of the laser light, the "reference beam", instead of striking the object first, is split off from the original laser beam and directed straight onto the film.

To "play back" the scene, one resupplies the reference beam, shining it onto the developed film—the window. This reveals the originally captured phase of the light waves as they passed through the window on their way from the objects in the scene. In effect, as the illustration shows, one can now "look through" the window and see the original object behind it.

Dennis Gabor's invention was not nearly as simple as a strobe light. To go deeper into the theory of holography, it is next necessary to understand interference and diffraction.

Interference and diffraction

Interference occurs when one or more wavefronts are superimposed. Diffraction occurs whenever a wavefront encounters an object. The process of producing a holographic reconstruction is explained below purely in terms of interference and diffraction. It is somewhat simplified but is accurate enough to provide an understanding of how the holographic process works.

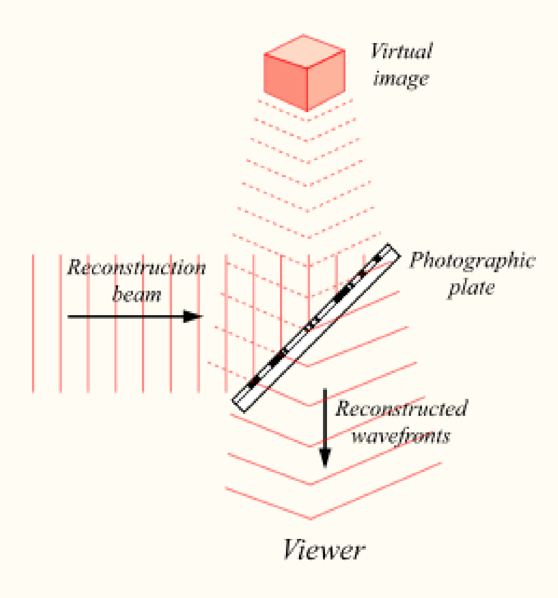
Plane wavefronts

A diffraction grating is a structure with a repeating pattern. A simple example is a metal plate with slits cut at regular intervals. Light rays travelling through it are bent at an angle determined by *λ*, the wavelength, and *d*, the distance between the slits, and is given by $\sin \theta = \lambda/d$.

A very simple hologram can be made by superimposing two plane waves from the same light source. One (the reference beam) hits the photographic plate normally, and the other one (the object beam) hits the plate at an angle, *θ*. The relative phase between the two beams varies across the photographic plate as $2\pi y \sin \theta/\lambda$, where *y* is the distance along the photographic plate. The two beams interfere with one another to form an interference pattern. The relative phase changes by 2π at intervals of $d = \lambda/\sin \theta$, so the spacing of the interference fringes is given by d. Thus, the relative phase of object and reference beam is encoded as the maxima and minima of the fringe pattern.

When the photographic plate is developed, the fringe pattern acts as a diffraction grating, and when the reference beam is incident upon the photographic plate, it is partly diffracted into the same angle *θ* at which the original object beam was incident. Thus, the object beam has been reconstructed. The diffraction grating created by the two waves interfering has *reconstructed* the "object beam", and it is therefore a hologram as defined above.

Point sources



Holographic reconstruction process

A slightly more complicated hologram can be made using a point source of light as object beam and a plane wave as reference beam to illuminate the photographic plate. An interference pattern is formed, which, in this case, is in the form of curves of decreasing separation with increasing distance from the centre (basically a sinusoidal zone plate).

The photographic plate is developed, giving a complicated pattern that can be considered to be made up of a diffraction pattern of varying spacing. When the plate is illuminated by the reference beam alone, it is diffracted by the grating into different angles, which depend on the local spacing of the pattern on the plate. It can be shown that the net effect of this is to reconstruct the object beam, so that it appears that light is coming from a point source behind the plate, even when the source has been removed. The light emerging from the photographic plate is identical to the light that emerged from the point source that used to be there. An observer looking into the plate from the other side will "see" a point source of light whether the original source of light is there or not.

This sort of hologram is effectively a concave lens, since it "converts" a plane wavefront into a divergent wavefront. It will also increase the divergence of any wave that is incident on it in exactly the same way as a normal lens does. Its focal length is the distance between the point source and the plate.

Complex objects

To record a hologram of a complex object, a laser beam is first split into two separate beams of light using a beam splitter of half-silvered glass or a birefringent material. One beam illuminates the object, reflecting its image onto the recording medium as it scatters the beam. The second (reference) beam illuminates the recording medium directly.

According to diffraction theory, each point in the object acts as a point source of light. Each of these point sources interferes with the reference beam, giving rise to an interference pattern. The resulting pattern is the sum of all *point source + reference beam* interference patterns.

When the object is no longer present, the holographic plate is illuminated by the reference beam. Each point-source diffraction grating will diffract part of the reference beam to reconstruct the wavefront from its point source. These individual wavefronts add together to reconstruct the whole of the object beam.

The viewer perceives a wavefront that is identical to the scattered wavefront of the object illuminated by the reference beam, so that it appears to him or her that the object is still in place. This image is known as a "virtual" image, as it is generated even though the object is no longer there. The direction of the light source seen illuminating the virtual image is that of the original illuminating beam.

Mathematical model

A light wave can be modeled by a complex number **U**, which represents the electric or magnetic field of the light wave. The amplitude and phase of the light are represented by the absolute value and angle of the complex number. The object and reference waves at any point in the holographic system are given by $\mathbf{U}_{\mathrm{O}}$ and $\mathbf{U}_{\mathrm{R}}$. The combined beam is given by $\mathbf{U}_{\mathrm{O}}$ + $\mathbf{U}_{\mathrm{R}}$. The energy of the combined beams is proportional to the square of magnitude of the electric wave:

$$|U_O + U_R|^2 = U_O U_R^* + |U_R|^2 + |U_O|^2 + U_O^* U_R$$

If a photographic plate is exposed to the two beams and then developed, its transmittance, **T**, is proportional to the light energy that was incident on the plate and is given by

$$T = k[U_O U_R^* + |U_R|^2 + |U_O|^2 + U_O^* U_R]$$

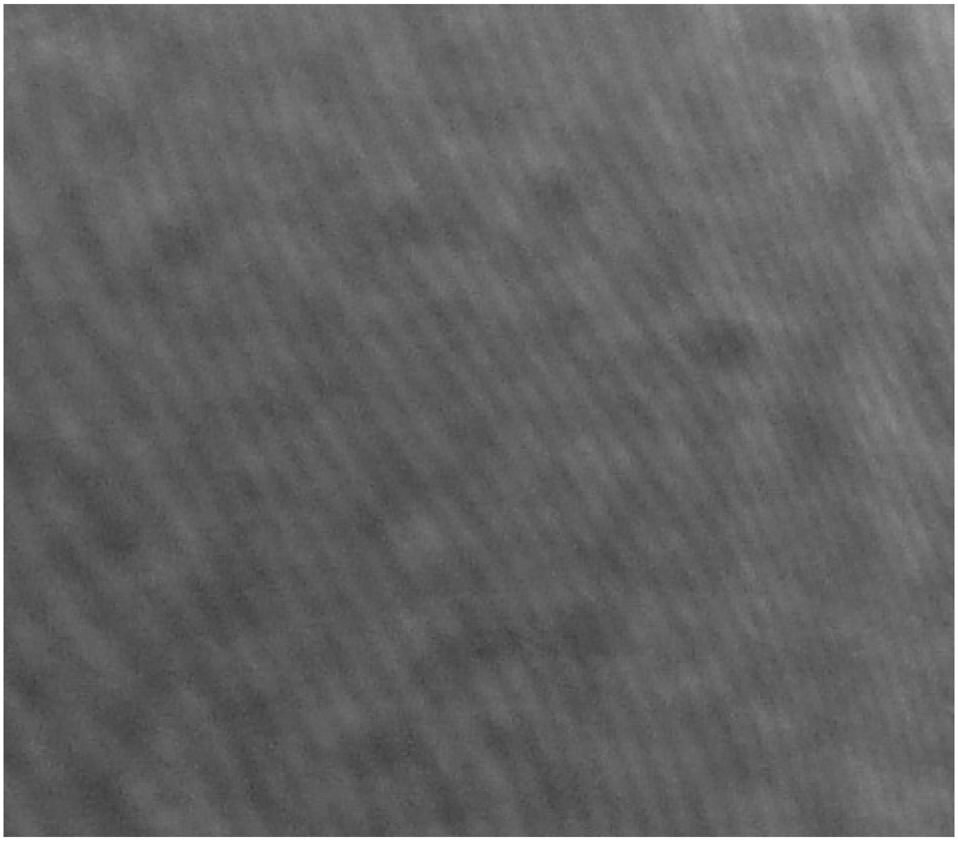
where *k* is a constant. When the developed plate is illuminated by the reference beam, the light transmitted through the plate, $\mathbf{U}_{\mathrm{H}}$ is

$$U_H = TU_R = k[U_O U_R^* + |U_R|^2 + |U_O|^2 + U_O^* U_R]U_R = k[U_O|U_R|^2 + |U_R|^2 U_R + |U_O|^2 U_R + U_O^* U_R^2]$$

It can be seen that $\mathbf{U}_{\mathrm{H}}$ has four terms. The first of these is proportional to $\mathbf{U}_{\mathrm{O}}$, and this is the reconstructed object beam. The second term represents the reference beam whose amplitude has been modified by $\mathbf{U}_{\mathrm{R}}^2$. The third also represents the reference beam, which has had its amplitude modified by $\mathbf{U}_{\mathrm{O}}^2$; this modification will cause the reference beam to be diffracted around its central direction. The fourth term is known as the "conjugate object beam". It has the reverse curvature to the object beam itself and forms a real image of the object in the space beyond the holographic plate.

Early holograms had both the object and reference beams illuminating the recording medium normally, which meant that all four beams emerging from the hologram were superimposed on one another. The off-axis hologram was developed by Leith and Upatnieks to overcome this problem. The object and reference beams are incident at well-separated angles onto the holographic recording medium, and the virtual, real and reference wavefronts all emerge at different angles, enabling the reconstructed object beam to be imaged clearly.

Viewing the hologram



Photograph of a hologram in front of a diffuse light background - 8×8 mm

The picture on the right is a photograph, taken against a diffuse light background, of a hologram recorded on photographic emulsion. The area shown is about 8 mm by 8 mm. The holographic recording is the random variation in intensity, which is an objective speckle pattern, and not the regular lines, which are likely to be due to interference arising from multiple reflections in the glass plate on which the photographic emulsion is mounted. It is no more possible to discern the subject of the hologram from this than it is to identify the music on a gramophone record by looking at the structure of the gramophone record surface. When this hologram is illuminated by a divergent laser beam, the viewer will see the object used to make it (in this case, a toy van), because the light is diffracted by the hologram to reconstruct the light that was scattered from the object.

When one looks at a scene, each eye captures a portion of the light scattered from the scene, and the lens of the eye forms an image of the scene on the retina, in which light from each angular position is focused to a specific angular position in the image plane. Since the hologram reconstructs the whole of the scattered light field that was incident on the hologram, the viewer sees the same image whether it is derived from the light field scattered from the object or the reconstructed light field produced by the hologram and is unable to tell whether he or she is looking at the real or the virtual object. If the viewer moves about, the object will appear to move in exactly the same way whether he or she is looking at the original light field or the reconstructed light field. If there are several objects in the scene, they will exhibit parallax. If the viewer is using both eyes (stereoscopic vision), he or she will get depth information when viewing the hologram in exactly the same way as when he or she is viewing the real scene.

A hologram is not a 3D photograph. A photograph records an image of the recorded scene from a single viewpoint, which is defined by the position of the camera lens. The hologram is not an image, but an encoding system which enables the scattered light field to be reconstructed. Images can then be formed from any point in the reconstructed beam either with a camera or by eye. It was very common in the early days of holography to use a chess board as the object and then take photographs at several different angles using the reconstructed light to show how the relative positions of the chess pieces appeared to change.

Since each point in the hologram contains light from the whole of the original scene, the whole scene can, in principle, be reconstructed from an arbitrarily small part of the hologram. To demonstrate this concept, the hologram can be broken into small pieces, and the entire object can still be seen from each small piece. If one envisions the hologram as a "window" on the object, then each small piece of hologram is just a part of the window from which it can still be viewed, even if the rest of the window is blocked off.

One does, however, lose resolution as the size of the hologram is decreased—the image becomes "fuzzier". This is a result of diffraction and arises in the same way as the resolution of an imaging system is ultimately limited by diffraction, where the resolution becomes coarser as the lens or lens-aperture diameter decreases.

Viewing and authoring

The object and the reference beams must be able to produce an interference pattern that is stable during the time in which the holographic recording is made. To do this, they must have the same frequency and the same relative phase during this time, that is, they must be mutually coherent. Many laser beams satisfy this condition, and lasers have been used to make holograms since their invention, though the first holograms by Gabor used "quasi-chromatic" light sources. In principle, two separate light sources could be used if the coherence condition could be satisfied, but in practice, a single laser is always used.

In addition, the medium used to record the fringe pattern must be able to resolve it, and some of the more common media used are listed below. The spacing of the fringes depends on the angle between the object and reference beams. For example, if this angle is 45° and the wavelength of the light is 0.5 μm, the fringe spacing is about 0.7 μm or 1400 lines/mm. A working hologram can be obtained even if not all the fringes are resolved, but the resolution of the image is reduced as the resolution of the recording medium decreases.

Mechanical stability is also very important when making a hologram. Any relative phase change between the object and reference beams due to vibration or air movement will cause the fringes on the recording medium to move, and if the phase change is greater than π, the fringe pattern is averaged out, and no holographic recording is obtained. Recording time can be several seconds or more, and given that a phase change of π is equivalent to a movement of $\lambda/2$, this is quite a stringent stability requirement.

Generally, the coherence length of the light determines the maximum depth in the scene of interest that can be recorded holographically. A good holography laser will typically have a coherence length of several meters, ample for a deep hologram. Certain pen laser pointers have been used to make small holograms. The size of these holograms is not restricted by the coherence length of the laser pointers (which can exceed several meters), but by their low power of below 5 mW.

The objects that form the scene must, in general, have optically rough surfaces so that they scatter light over a wide range of angles. A specularly reflecting (or shiny) surface reflects the light in only one direction at each point on its surface, so in general, most of the light will not be incident on the recording medium. The light scattered from objects with a rough surface forms an objective speckle pattern that has random amplitude and phase.

The reference beam is not normally a plane wavefront; it is usually a divergent wavefront that is formed by placing a convex lens in the path of the laser beam.

To reconstruct the object exactly from a transmission hologram, the reference beam must have the same wavelength and curvature, and must illuminate the hologram at the same angle, as the original reference beam (i.e., only the phase can be changed). Departure from any of these conditions will give a distorted reconstruction. While nearly all holograms are recorded using lasers, a narrow-band lamp or even sunlight is enough to recognize the reconstructed image.

The reconstructed hologram is enlarged if the light used to reconstruct the hologram has a higher wavelength. This initially generated some interest, since it seemed to be possible to use X-rays to make holograms of molecules and view them using visible light. However, X-ray holograms have not been created to date. This effect can be demonstrated using a light source that emits several different frequencies.

Exact reconstruction is achieved in holographic interferometry, where the holographically reconstructed wavefront interferes with the live wavefront, to map out any displacement of the live object, and gives a null fringe if the object has not moved.

Holographic recording media

The recording medium must be able to resolve the interference fringes as discussed above. It must also be sufficiently sensitive to record the fringe pattern in a time period short enough for the system to remain optically stable, i.e., any relative movement of the two beams must be significantly less than $\lambda/2$. It is possible to record holograms in certain materials using a high-power pulsed laser technique that uses only a couple of nanoseconds to record the holographic pattern.

The recording medium has to convert the interference pattern into an optical element that modifies either the amplitude or the phase of a light beam that is incident upon it. These are known as amplitude and phase holograms, respectively. In amplitude holograms, the modulation is in the varying absorption of the light by the hologram, as in a developed photographic emulsion that is less or more absorptive depending on the intensity of the light that illuminated it. In phase holograms, the optical distance (i.e., the refractive index or, in some cases, the thickness) in the material is modulated.

Most materials used for phase holograms reach the theoretical diffraction efficiency for holograms, which is 100% for thick holograms (Bragg diffraction regime) and 33.9% for thin holograms (Raman-Nath diffraction regime, holographic films typically some micrometers thick). Amplitude holograms have a lower efficiency than phase holograms and are therefore used more rarely.

The table below shows the principal materials for holographic recording. Note that these do not include the materials used in the mass replication of an existing hologram. The resolution limit given in the table indicates the maximal number of interference lines per millimeter of the gratings. The required exposure is for a long exposure. Short exposure times (less than 1/1000 of a second, such as with a pulsed laser) require a higher exposure due to reciprocity failure.

General properties of recording materials for holography. Source:

Material	**Reusable**	**Processing**	**Type of hologram**	**Max. efficiency**	**Required exposure [mJ/cm²]**	**Resolution limit [mm^{-1}]**
Photographic emulsions	No	Wet	Amplitude	6%	0.001–0.1	1,000–10,000
			Phase (bleached)	60%		
Dichromated gelatin	No	Wet	Phase	100%	10	10,000
Photoresists	No	Wet	Phase	33%	10	3,000
Photothermoplastics	Yes	Charge and heat	Phase	33%	0.01	500–1,200
Photopolymers	No	Post exposure	Phase	100%	1–1,000	2,000–5,000
Photochromics	Yes	None	Amplitude	2%	10–100	>5,000
Photorefractives	Yes	None	Phase	100%	0.1–50,000	2,000–10,000
Elastomers	No	None	Phase	--	300	--

Holoprinters

A holoprinter is a holographic printing device that can print out full-colour digital holograms from a rendered 3D model or a video series. The machine can cost up to half a million dollars and is about the size of a small room. It uses red, green and blue lasers to write a series of dots, or holopixels, across a holographic medium. The holopixel contains information about the whole image from its own unique perspective. The information for each holopixel is computed from a series of rendered images generated via computer graphics. The holographic medium is typically a polymer film. The film may require development after exposure. It is then laminated on to a hard plastic backing. Printing a digital hologram can take several hours, as each holopixel dot has to be written individually in three colours, where the colours overlap within the medium. The size of a holopixel is typically around a square millimeter.

There are only a few digital holoprinter manufacturers in the world, including Geola (Lithuania), View Holographics (UK) and Zebra Imaging (US).

Embossing and mass production

An existing hologram can be replicated, either in an optical way similar to holographic recording or, in the case of surface relief holograms, by embossing. Surface relief holograms are recorded in photoresists or photothermoplastics and allow cheap mass reproduction. Such embossed holograms are now widely used, for instance, as security features on credit cards or quality merchandise. The Royal Canadian Mint even produces holographic gold and silver coinage through a complex stamping process. The first book to feature a hologram on the front cover was *The Skook* (Warner Books, 1984) by JP Miller, featuring an illustration by Miller. That same year, "Telstar" by Ad Infinitum became the first record with a hologram cover and *National Geographic* published the first magazine with a hologram cover.

The first step in the embossing process is to make a stamper by electrodeposition of nickel on the relief image recorded on the photoresist or photothermoplastic. When the nickel layer is thick enough, it is separated from the master hologram and mounted on a metal backing plate. The material used to make embossed copies consists of a polyester base film, a resin separation layer and a thermoplastic film constituting the holographic layer.

The embossing process can be carried out with a simple heated press. The bottom layer of the duplicating film (the thermoplastic layer) is heated above its softening point and pressed against the stamper, so that it takes up its shape. This shape is retained when the film is cooled and removed from the press. In order to permit the viewing of embossed holograms in reflection, an additional reflecting layer of aluminum is usually added on the hologram recording layer.

It is possible to print holograms directly into steel using a sheet explosive charge to create the required surface relief.

Applications

Data storage

Holography can be put to a variety of uses other than recording images. Holographic data storage is a technique that can store information at high density inside crystals or photopolymers. The ability to store large amounts of information in some kind of media is of great importance, as many electronic products incorporate storage devices. As current storage techniques such as Blu-ray Disc reach the limit of possible data density (due to the diffraction-limited size of the writing beams), holographic storage has the potential to become the next generation of popular storage media. The advantage of this type of data storage is that the volume of the recording media is used instead of just the surface. Currently available SLMs can produce about 1000 different images a second at 1024×1024-bit resolution. With the right type of media (probably polymers rather than something like $LiNbO_3$), this would result in about one-gigabit-per-second writing speed. Read speeds can surpass this, and experts believe one-terabit-per-second readout is possible. In 2005, companies such as Optware and Maxell have produced a 120 mm disc that uses a holographic layer to store data to a potential 3.9&nbspTB, which they plan to market under the name Holographic Versatile Disc. Another company, InPhase Technologies, is developing a competing format. While many holographic data storage models have used "page-based" storage, where each recorded hologram holds a large amount of data, more recent research into using submicrometre-sized "microholograms" has resulted in several potential 3D optical data storage solutions. While this approach to data storage can not attain the high data rates of page-based storage, the tolerances, technological hurdles, and cost of producing a commercial product are significantly lower.

Security

UBS Kinebar gold bars use holograms as a security measure.

Security holograms are very difficult to forge, because they are replicated from a master hologram that requires expensive, specialized and technologically advanced equipment. They are used widely in many currencies, such as the Brazilian real 20 note, British pound 5/10/20 notes, Estonian kroon 25/50/100/500 notes, Canadian dollar 5/10/20/50/100 notes, Euro 5/10/20/50/100/200/500 notes, South Korean won 5000/10000/50000 notes, and Japanese yen 5000/10000 notes. They are also used in credit and bank cards as well as passports, ID cards, books, DVDs, and sports equipment.

Holography allows for different levels of security, depending on budget and intensity of security. The highest level of security in fully custom holography, this involves the design and creation of unique images in three dimensions, cost can range from $5,000 to $15,000. For a tightly budgeted project, there are two choices of hologram: overprint holographic diffraction foil or custom etched diffraction material, which aren't dimensional but diffract light into patterns of bright rainbow light. The best option for high security and value for small to medium sized projects is customized stock holography (HoloBank, 2002).

Art

Early on, artists saw the potential of holography as a medium and gained access to science laboratories to create their work. Holographic art is often the result of collaborations between scientists and artists, although some holographers would regard themselves as both an artist and a scientist.

Salvador Dalí claimed to have been the first to employ holography artistically. He was certainly the first and best-known surrealist to do so, but the 1972 New York exhibit of Dalí holograms had been preceded by the holographic art exhibition that was held at the Cranbrook Academy of Art in Michigan in 1968 and by the one at the Finch College gallery in New York in 1970, which attracted national media attention.

During the 1970s, a number of art studios and schools were established, each with their particular approach to holography. Notably, there was the San Francisco School of Holography established by Lloyd Cross, The Museum of Holography in New York founded by Rosemary (Possie) H. Jackson, the Royal College of Art in London and the Lake Forest College Symposiums organised by Tung Jeong (T.J.). None of these studios still exist; however, there is the Center for the Holographic Arts in New York and the HOLOcenter in Seoul, which offers artists a place to create and exhibit work.

During the 1980s, many artists who worked with holography helped the diffusion of this so-called "new medium" in the art world, such as Harriet Casdin-Silver of the USA, Dieter Jung of Germany, and Moysés Baumstein of Brazil, each one searching for a proper "language" to use with the three-dimensional work, avoiding the simple holographic reproduction of a scuplture or object. For instance, in Brazil, many concrete poets (Augusto de Campos, Décio Pignatari, Julio Plaza and José Wagner Garcia, associated with Moysés Baumstein) found in holography a way to express themselves and to renew the Concrete Poetry (or Shape Poetry).

A small but active group of artists still use holography as their main medium, and many more artists integrate holographic elements into their work. Some are associated with novel holographic techniques; for example, artist Matt Brand employed computational mirror design to eliminate image distortion from specular holography.

The MIT Museum and Jonathan Ross both have extensive collections of holography and on-line catalogues of art holograms.

Hobbyist use

Peace Within Reach, a Denisyuk DCG hologram by amateur Dave Battin

Since the beginning of holography, experimenters have explored its uses. Starting in 1971, Lloyd Cross started the San Francisco School of Holography and started to teach amateurs the methods of making holograms with inexpensive equipment. This method relied on the use of a large table of deep sand to hold the optics rigid and damp vibrations that would destroy the image.

Many of these holographers would go on to produce art holograms. In 1983, Fred Unterseher published the *Holography Handbook*, a remarkably easy-to-read description of making holograms at home. This brought in a new wave of holographers and gave simple methods to use the then-available AGFA silver halide recording materials.

In 2000, Frank DeFreitas published the *Shoebox Holography Book* and introduced using inexpensive laser pointers to countless hobbyists. This was a very important development for amateurs, as the cost for a 5 mW laser dropped from $1200 to $5 as semiconductor laser diodes reached mass market. Now, there are hundreds to thousands of amateur holographers worldwide.

In 2006, a large number of surplus Holography Quality Green Lasers (Coherent C315) became available and put Dichromated Gelatin (DCG) within the reach of the amateur holographer. The holography community was surprised at the amazing sensitivity of

DCG to green light. It had been assumed that the sensitivity would be non-existent. Jeff Blyth responded with the G307 formulation of DCG to increase the speed and sensitivity to these new lasers.

Many film suppliers have come and gone from the silver-halide market. While more film manufactures have filled in the voids, many amateurs are now making their own film. The favorite formulations are Dichromated Gelatin, Methylene Blue Sensitised Dichromated Gelatin and Diffusion Method Silver Halide preparations. Jeff Blyth has published very accurate methods for making film in a small lab or garage.

A small group of amateurs are even constructing their own pulsed lasers to make holograms of moving objects.

Holographic interferometry

Holographic interferometry (HI) is a technique that enables static and dynamic displacements of objects with optically rough surfaces to be measured to optical interferometric precision (i.e. to fractions of a wavelength of light). It can also be used to detect optical-path-length variations in transparent media, which enables, for example, fluid flow to be visualized and analyzed. It can also be used to generate contours representing the form of the surface.

It has been widely used to measure stress, strain, and vibration in engineering structures.

Interferometric microscopy

The hologram keeps the information on the amplitude and phase of the field. Several holograms may keep information about the same distribution of light, emitted to various directions. The numerical analysis of such holograms allows one to emulate large numerical aperture, which, in turn, enables enhancement of the resolution of optical microscopy. The corresponding technique is called interferometric microscopy. Recent achievements of interferometric microscopy allow one to approach the quarter-wavelength limit of resolution.

As sensors or biosensors

The hologram is made with a modified material that interacts with certain molecules generating a change in the fringe periodicity or refractive index, therefore, the color of the holographic reflection.

Dynamic holography

In static holography, recording, developing and reconstructing occur sequentially, and a permanent hologram is produced.

There also exist holographic materials that do not need the developing process and can record a hologram in a very short time. This allows one to use holography to perform some simple operations in an all-optical way. Examples of applications of such real-time holograms include phase-conjugate mirrors ("time-reversal" of light), optical cache memories, image processing (pattern recognition of time-varying images), and optical computing.

The amount of processed information can be very high (terabits/s), since the operation is performed in parallel on a whole image. This compensates for the fact that the recording time, which is in the order of a microsecond, is still very long compared to the processing time of an electronic computer. The optical processing performed by a dynamic hologram is also much less flexible than electronic processing. On one side, one has to perform the operation always on the whole image, and on the other side, the operation a hologram can perform is basically either a multiplication or a phase conjugation. In optics, addition and Fourier transform are already easily performed in linear materials, the latter simply by a lens. This enables some applications, such as a device that compares images in an optical way.

The search for novel nonlinear optical materials for dynamic holography is an active area of research. The most common materials are photorefractive crystals, but in semiconductors or semiconductor heterostructures (such as quantum wells), atomic vapors and gases, plasmas and even liquids, it was possible to generate holograms.

A particularly promising application is optical phase conjugation. It allows the removal of the wavefront distortions a light beam receives when passing through an aberrating medium, by sending it back through the same aberrating medium with a conjugated phase. This is useful, for example, in free-space optical communications to compensate for atmospheric turbulence (the phenomenon that gives rise to the twinkling of starlight).

Non-optical applications

In principle, it is possible to make a hologram for any wave.

Electron holography is the application of holography techniques to electron waves rather than light waves. Electron holography was invented by Dennis Gabor to improve the resolution and avoid the aberrations of the transmission electron microscope. Today it is commonly used to study electric and magnetic fields in thin films, as magnetic and electric fields can shift the phase of the interfering wave passing through the sample. The principle of electron holography can also be applied to interference lithography.

Acoustic holography is a method used to estimate the sound field near a source by measuring acoustic parameters away from the source via an array of pressure and/or particle velocity transducers. Measuring techniques included within acoustic holography are becoming increasingly popular in various fields, most notably those of transportation, vehicle and aircraft design, and NVH. The general idea of acoustic holography has led to different versions such as near-field acoustic holography (NAH) and statistically optimal near-field acoustic holography (SONAH). For audio rendition, the wave field synthesis is the most related procedure.

Atomic holography has evolved out of the development of the basic elements of atom optics. With the Fresnel diffraction lens and atomic mirrors atomic holography follows a natural step in the development of the physics (and applications) of atomic beams. Recent developments including atomic mirrors and especially ridged mirrors have provided the tools necessary for the creation of atomic holograms, although such holograms have not yet been commercialized.

Other applications

Holographic scanners are in use in post offices, larger shipping firms, and automated conveyor systems to determine the three-dimensional size of a package. They are often used in tandem with checkweighers to allow automated pre-packing of given volumes, such as a truck or pallet for bulk shipment of goods. Holograms produced in elastomers can be used as stress-strain reporters due to its elasticity and compressibility, the pressure and force applied are correlated to the reflected wavelength, therefore its color.