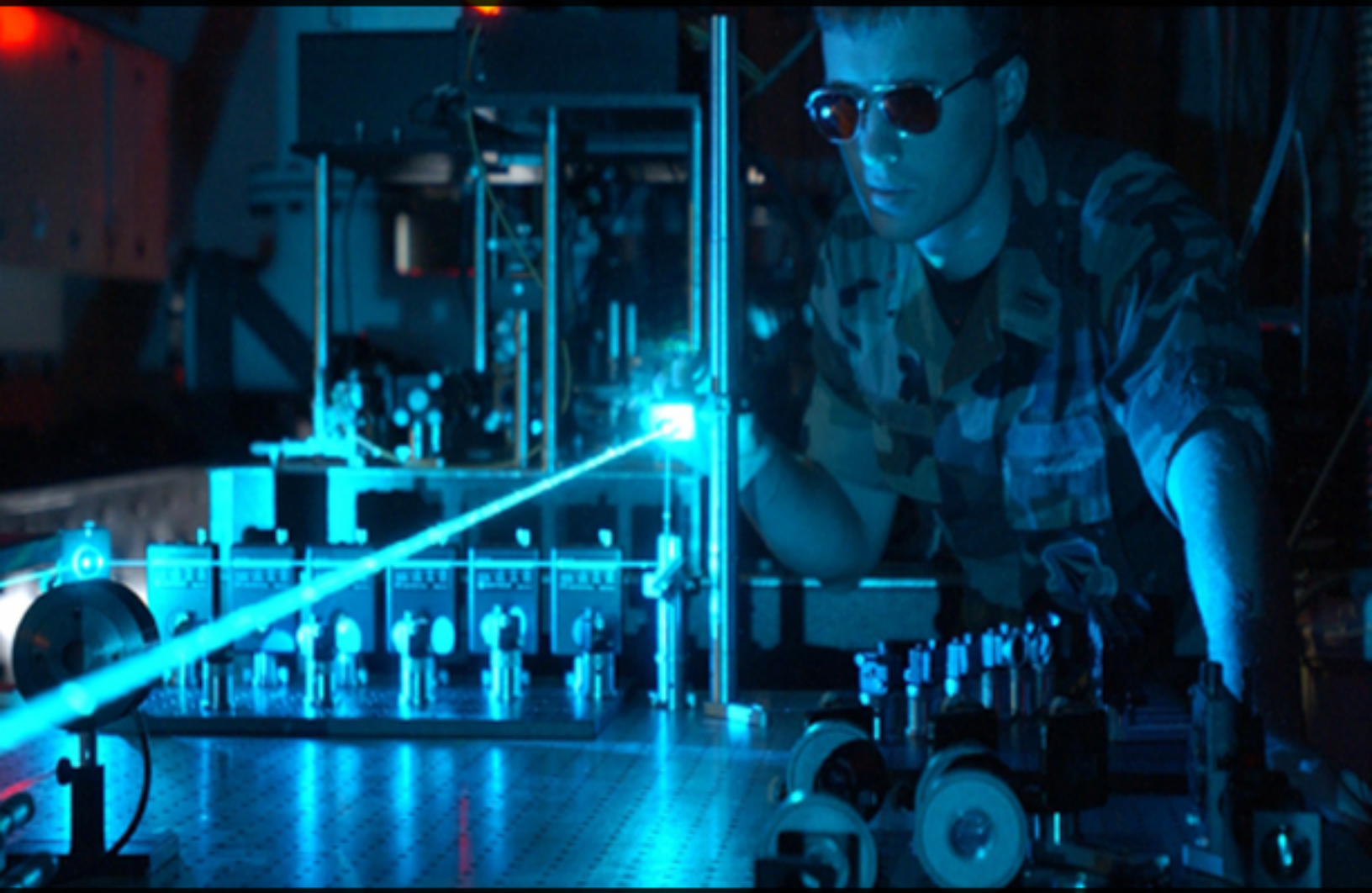
Optical Engineering & Optoelectronics



Lucille Samson

Kandi Coward

First Edition, 2012

ISBN 978-81-323-0963-5

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

Table of Contents

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.

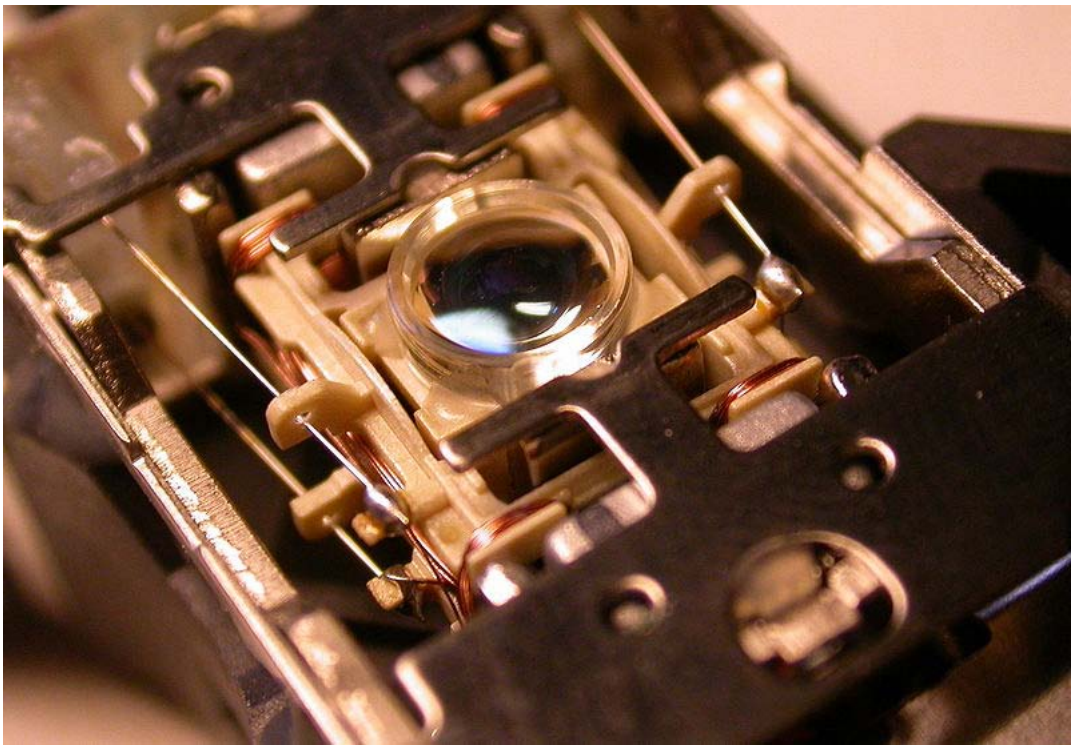
Modern desktop computers can now raytrace tens to hundreds of millions of rays per second through a lens, and perform hundreds to thousands of optimization cycles per second, rapidly exploring the n-dimensional design volume and even hill-climbing in and out of local minima in the search for the best 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 (Mbit/s)	Max (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 (total)	Diameter (cm)	Capacity (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.

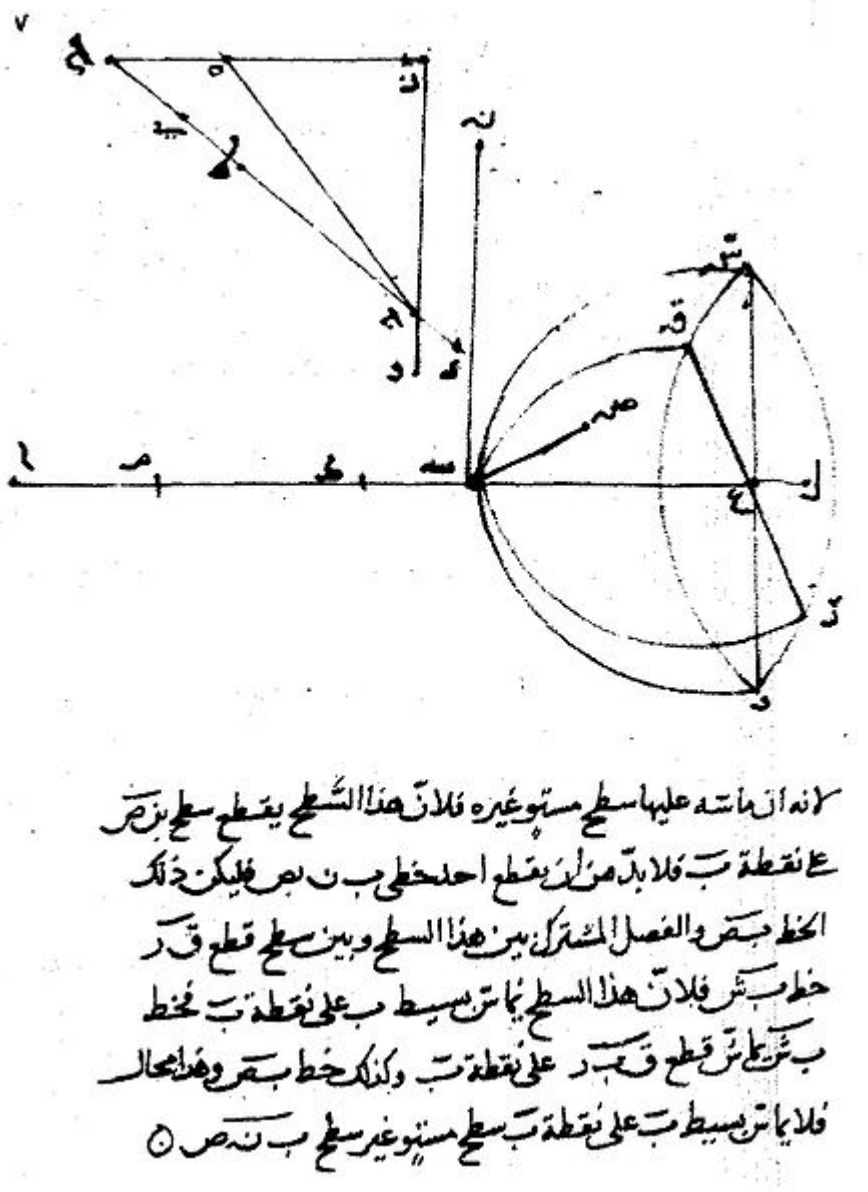
Most optical phenomena can be accounted for using the classical electromagnetic description of light. Complete electromagnetic descriptions of light are, however, often difficult to apply in practice. Practical optics is usually done using simplified models. The most common of these, geometric optics, 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.



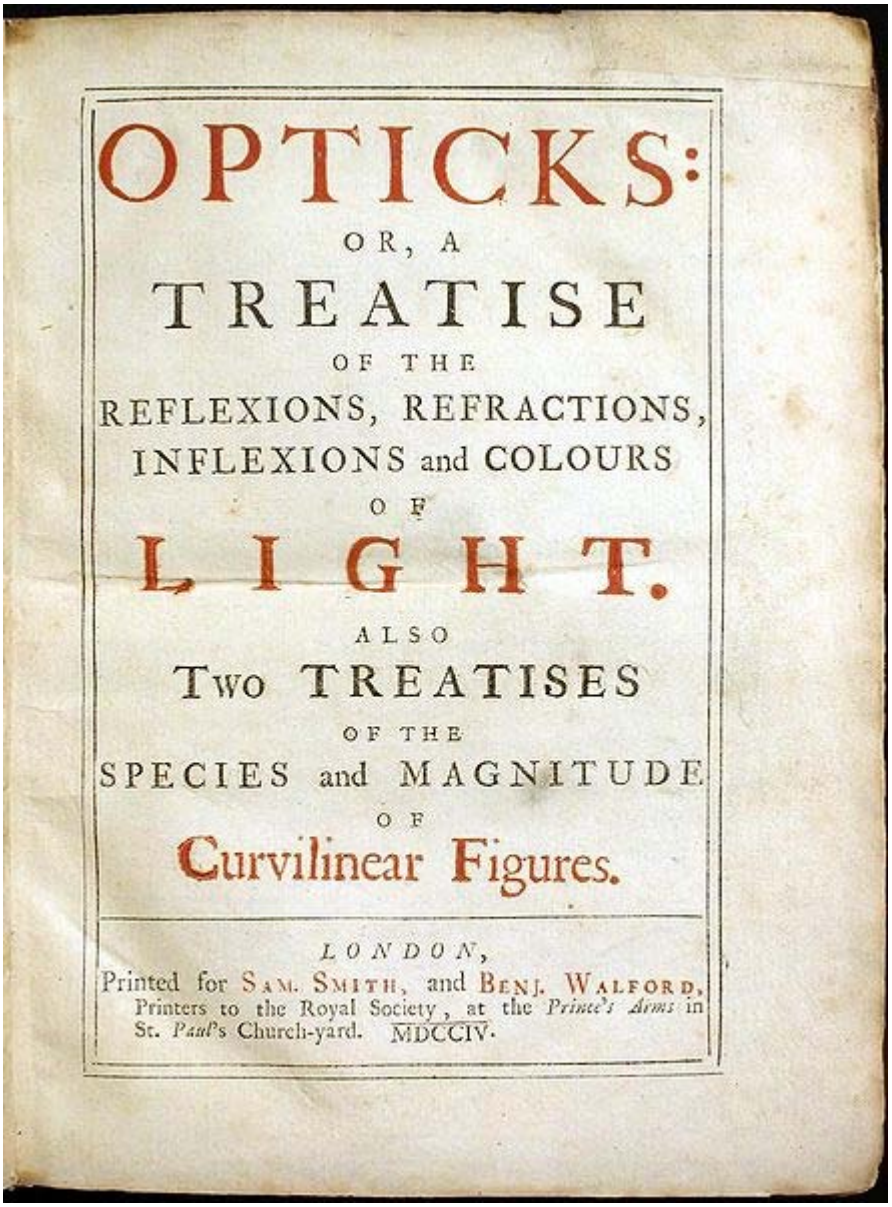
Reproduction of a page of Ibn Sahl's manuscript showing his knowledge of the law of refraction, now known as Snell's law.

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

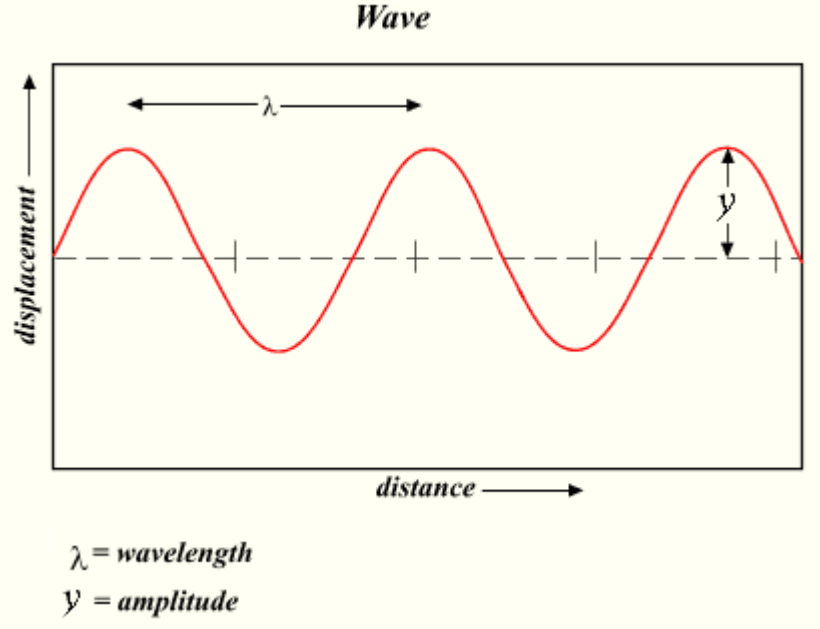
the time, partly because of his success in other areas of physics, he was generally considered to be the victor in the debate over the nature of light.

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:

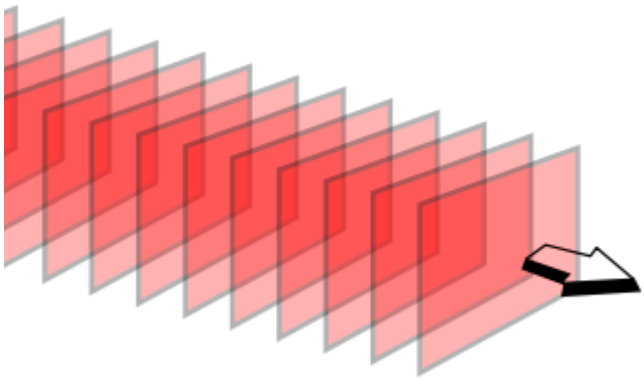
$$n = c / v.$$

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

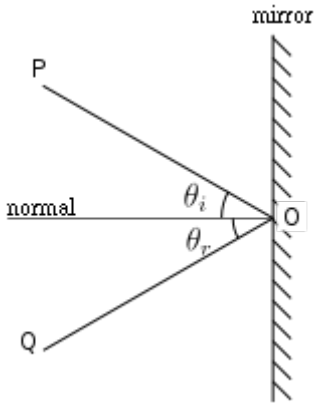


Diagram of specular reflection

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

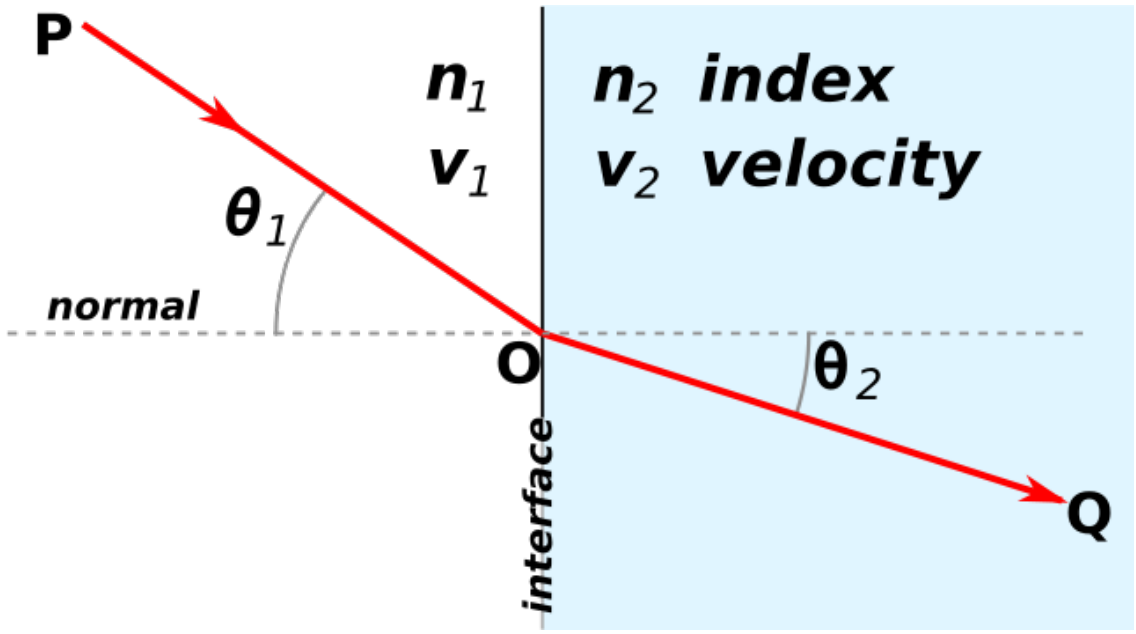


Illustration of Snell's Law for the case $n_1 < n_2$, such as air/water interface

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:

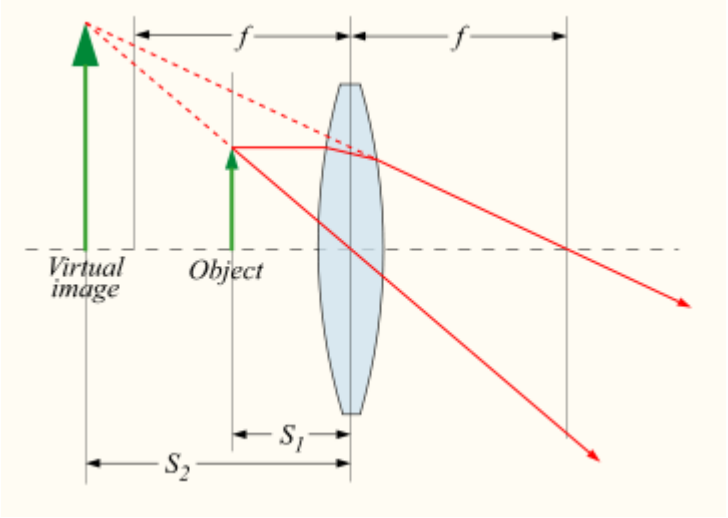
$$v_1 \sin\theta_2 = v_2 \sin\theta_1$$

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.

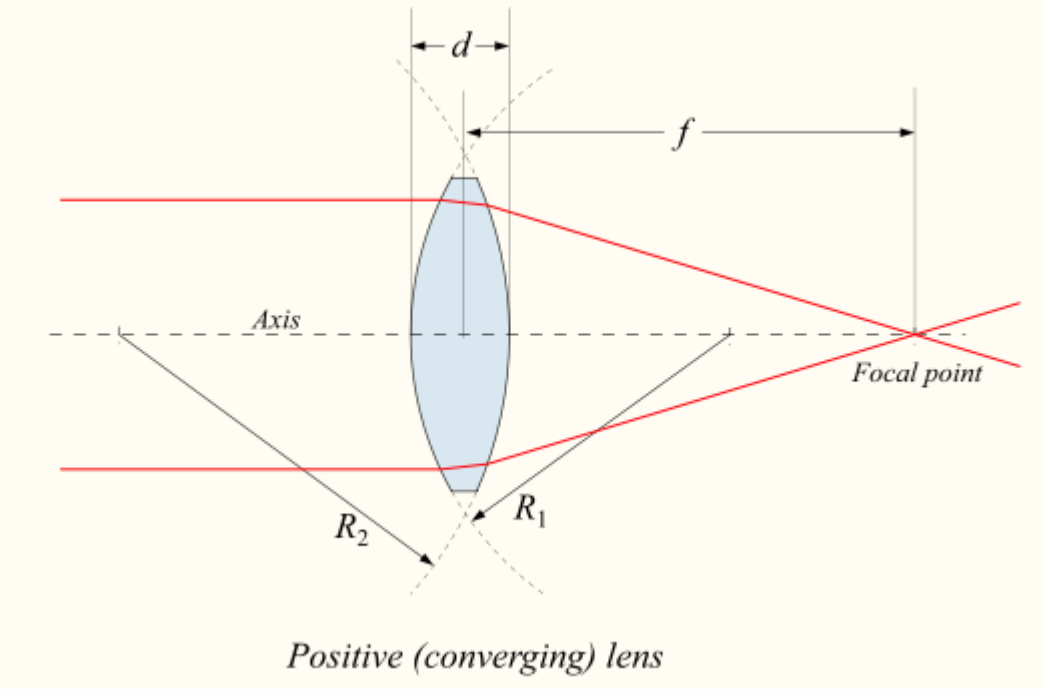


A ray tracing diagram for a converging lens.

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.

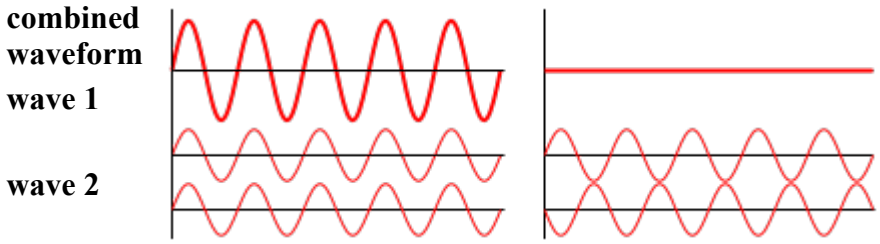
Using numerical modeling on a computer, optical scientists can simulate the propagation of light and account for most diffraction, interference, and polarization effects. Such simulations typically still rely on approximations, however, so this is not a full electromagnetic wave theory model of the propagation 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**

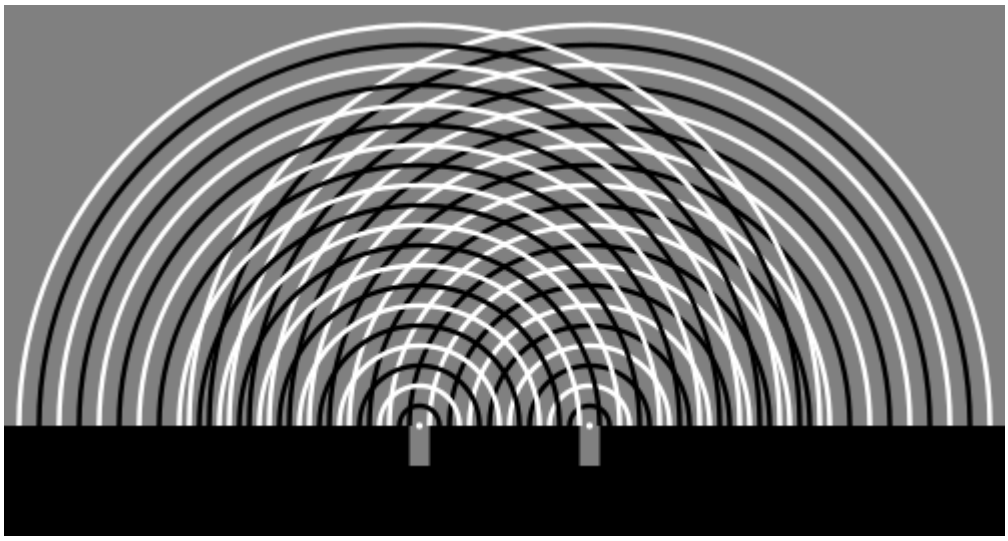
When oil or fuel is spilled, colorful patterns are formed by thin-film interference.

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.

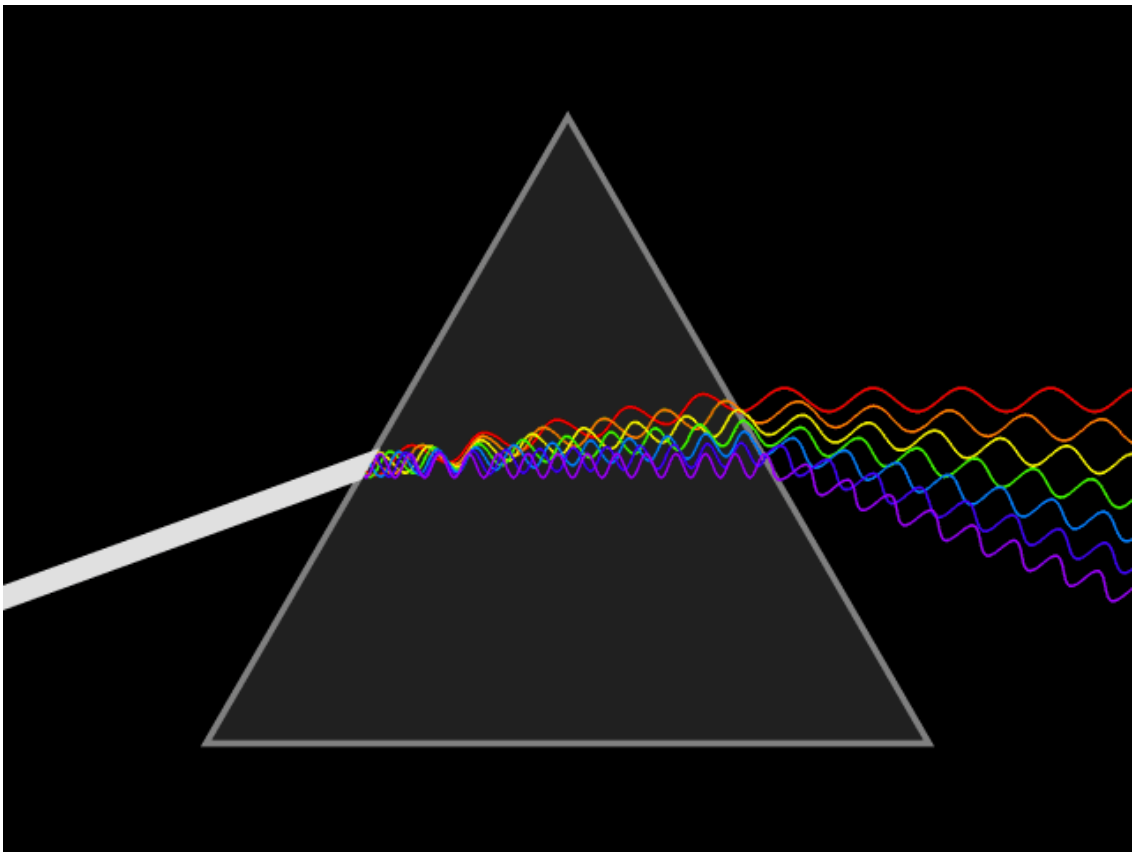
Diffraction effects limit the ability for an optical detector to optically resolve separate light sources. In general, light that is passing through an aperture will experience diffraction and the best images that can be created (as described in diffraction-limited optics) appear as a central spot with 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

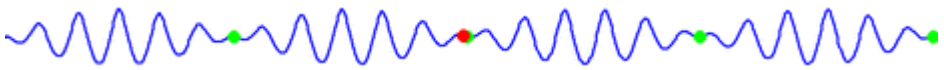


Conceptual animation of light dispersion through a prism. High frequency (blue) light is deflected the most, and low frequency (red) the least.

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 $\arcsin(\sin(\theta) / n)$. 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.

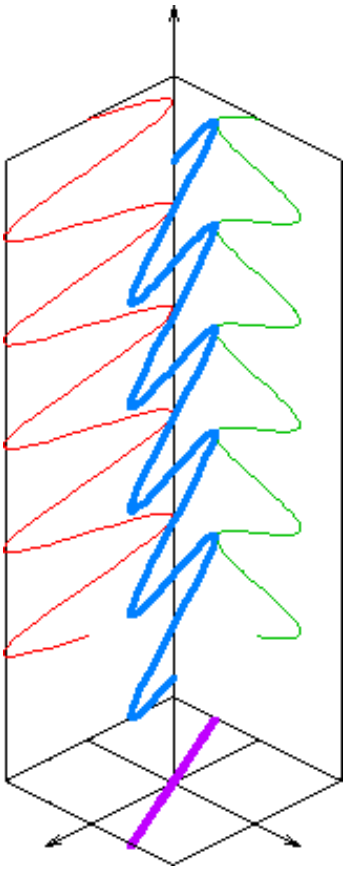
The result of group velocity dispersion, whether negative or positive, is ultimately temporal spreading of the pulse. This makes dispersion management extremely important in optical communications systems based on optical fibers, since if dispersion is too high, a group of pulses representing information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

Polarization

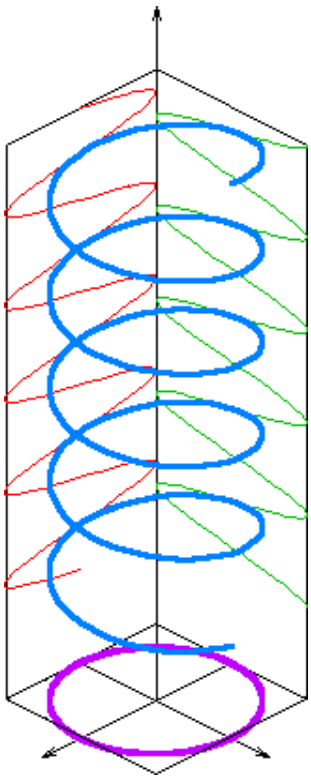
Polarization is a general property of waves that describes the orientation of their oscillations. For transverse waves such as many electromagnetic waves, it describes the orientation of the oscillations in the plane perpendicular to the wave's direction of travel. The oscillations may be oriented in 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

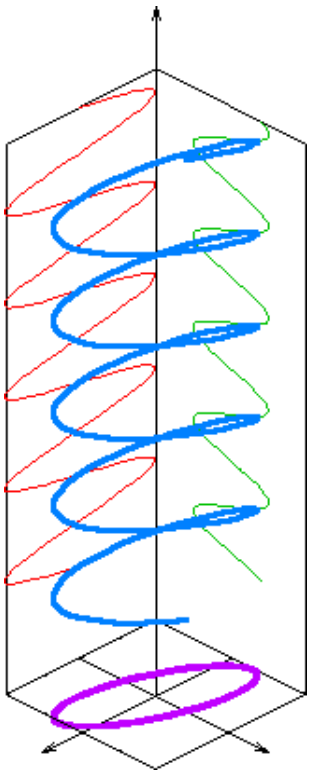
z indicating the direction of travel). The shape traced out in the x-y plane by the electric field vector is a Lissajous figure that describes the *polarization state*. The following figures show some examples of the evolution of the electric field vector (blue), with time (the vertical axes), at a particular 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



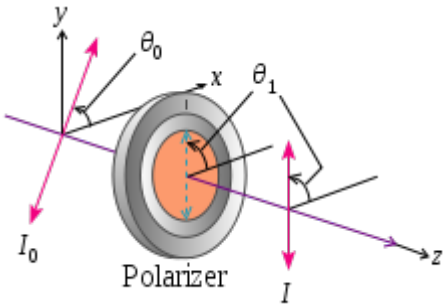
Elliptical polarization

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

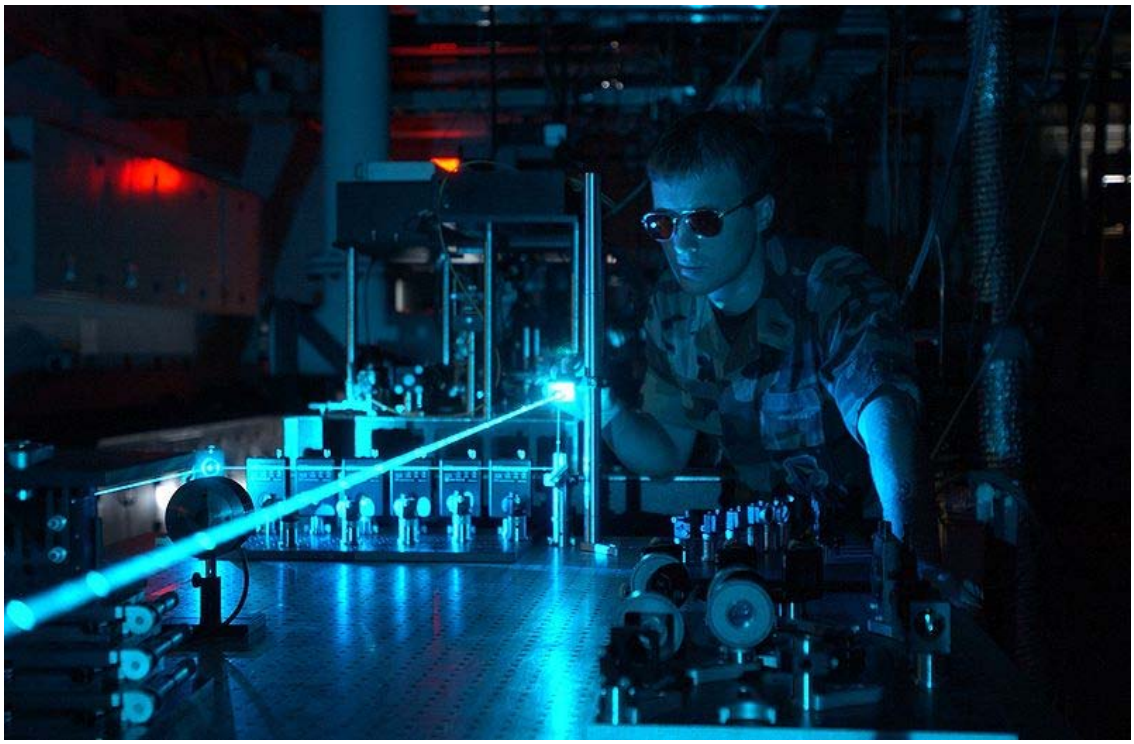
Modern optics encompasses the areas of optical science and engineering that became popular in the 20th century. These areas of optical science typically relate to the electromagnetic or quantum properties of light but do include other topics. A major subfield of modern optics, quantum optics, deals with 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.

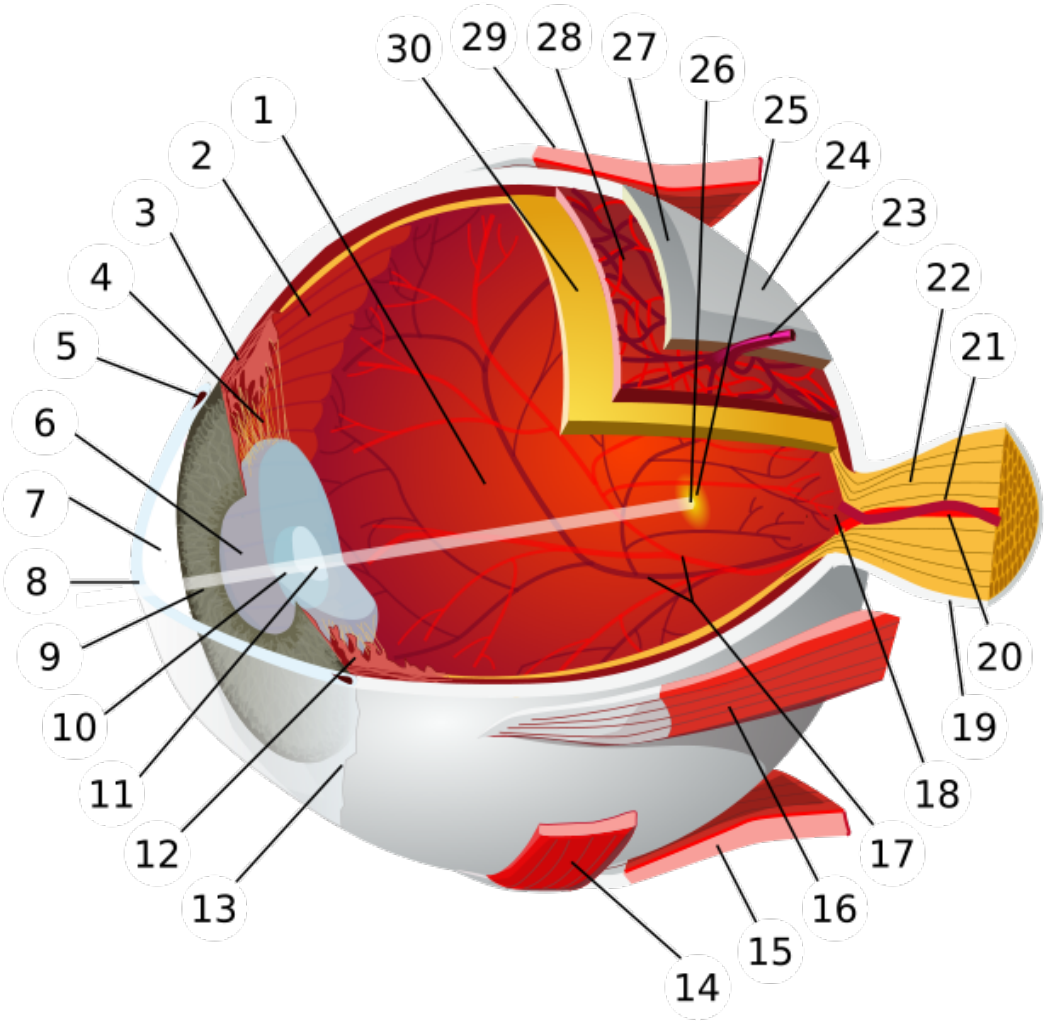
A laser is a device that emits light (electromagnetic radiation) through a process called *stimulated emission*. The term *laser* is an acronym for *Light Amplification by Stimulated Emission of Radiation*. Laser light is usually spatially coherent, which means that the light either is emitted in a narrow, 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.

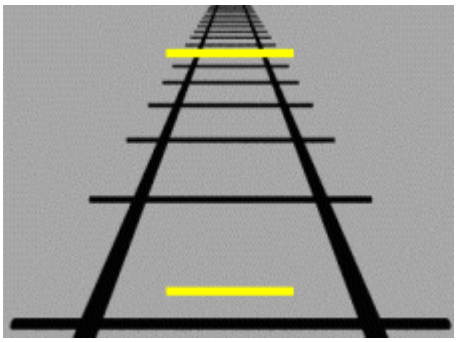
Defects in vision can be explained using optical principles. As people age, the lens becomes less flexible and the near point recedes from the eye, a condition known as presbyopia. Similarly, people suffering from hyperopia cannot decrease the focal length of their lens enough to allow for nearby objects 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



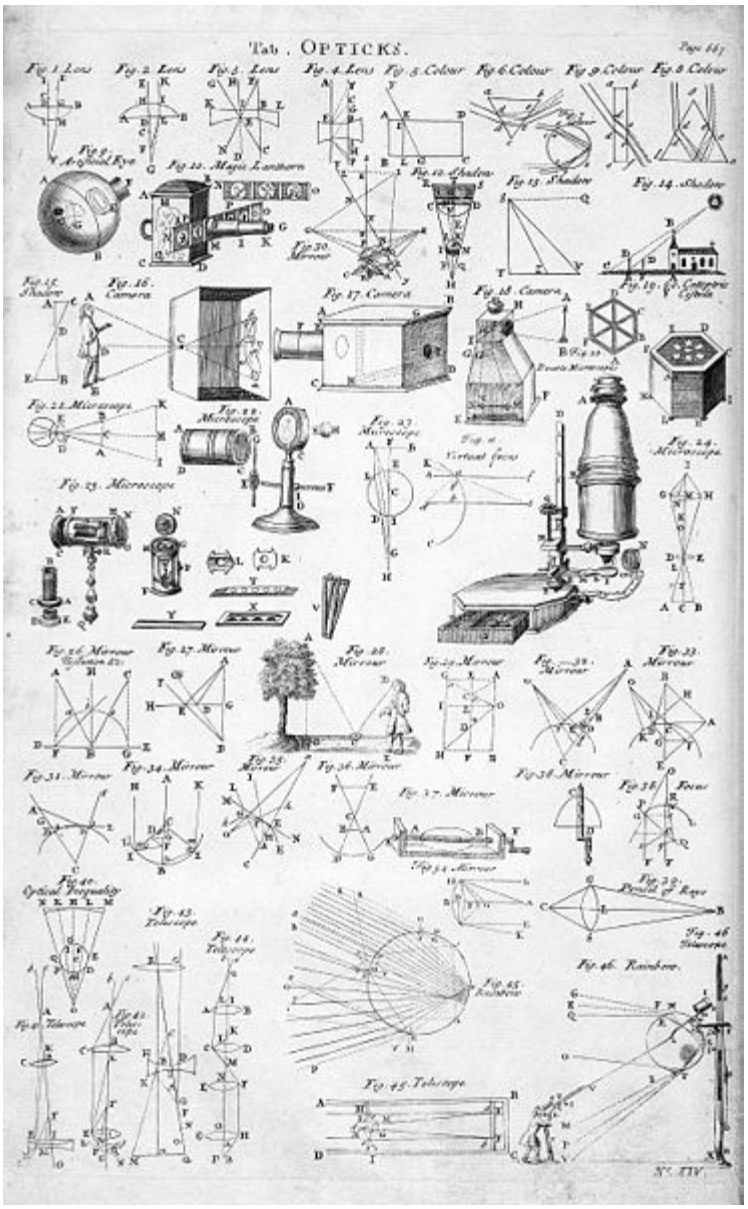
The Ponzo Illusion relies on the fact that parallel lines appear to converge as they approach infinity.

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)

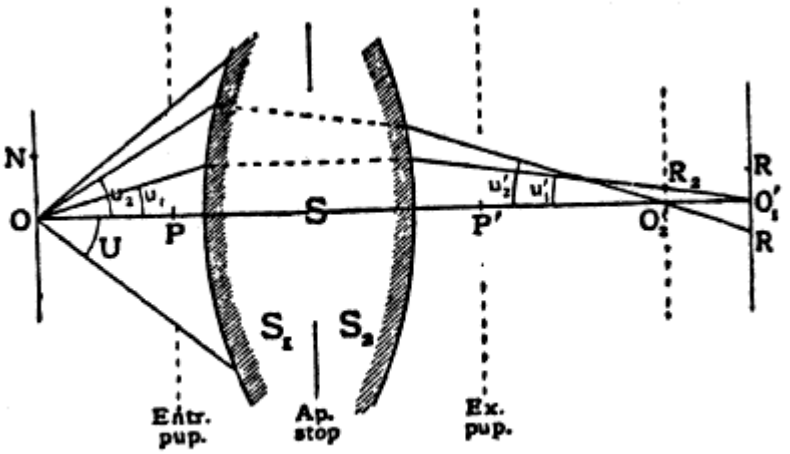


FIG. 5.

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.

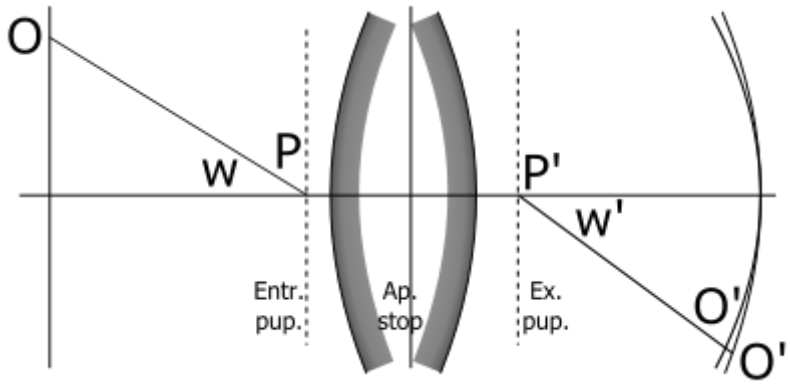


Figure 6

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

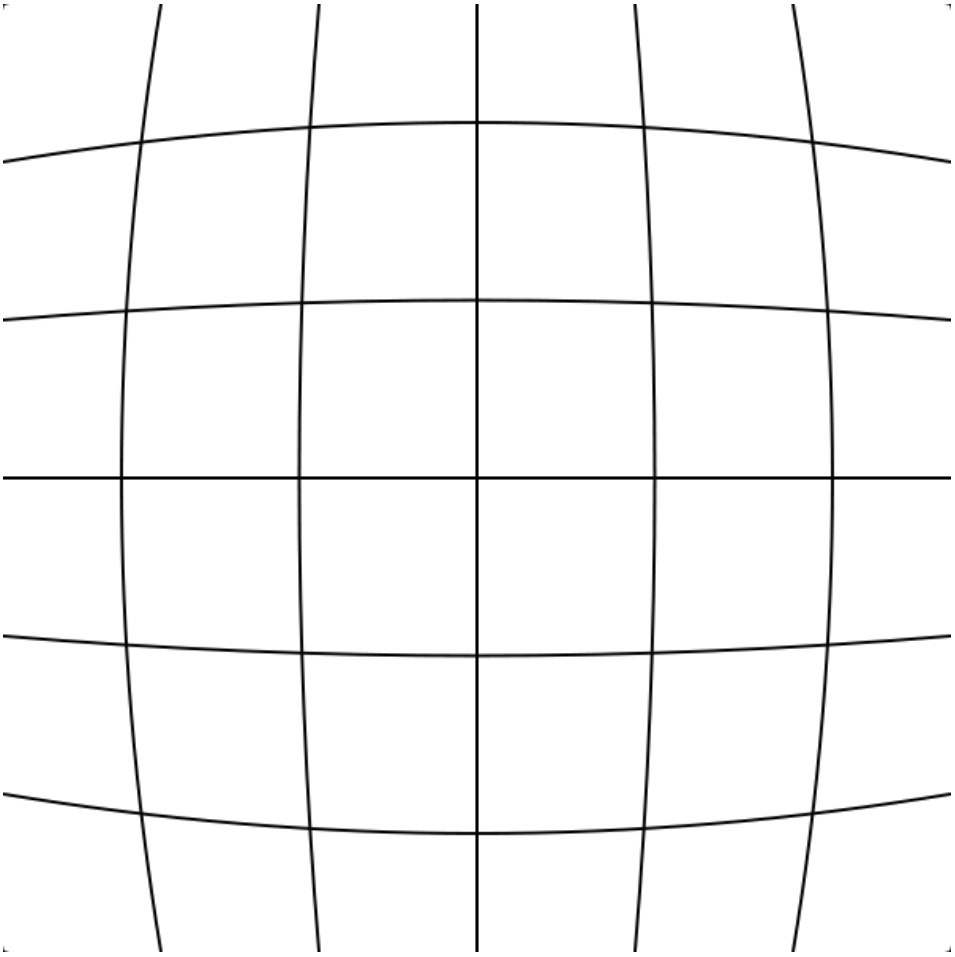


Fig. 7a: Barrel distortion

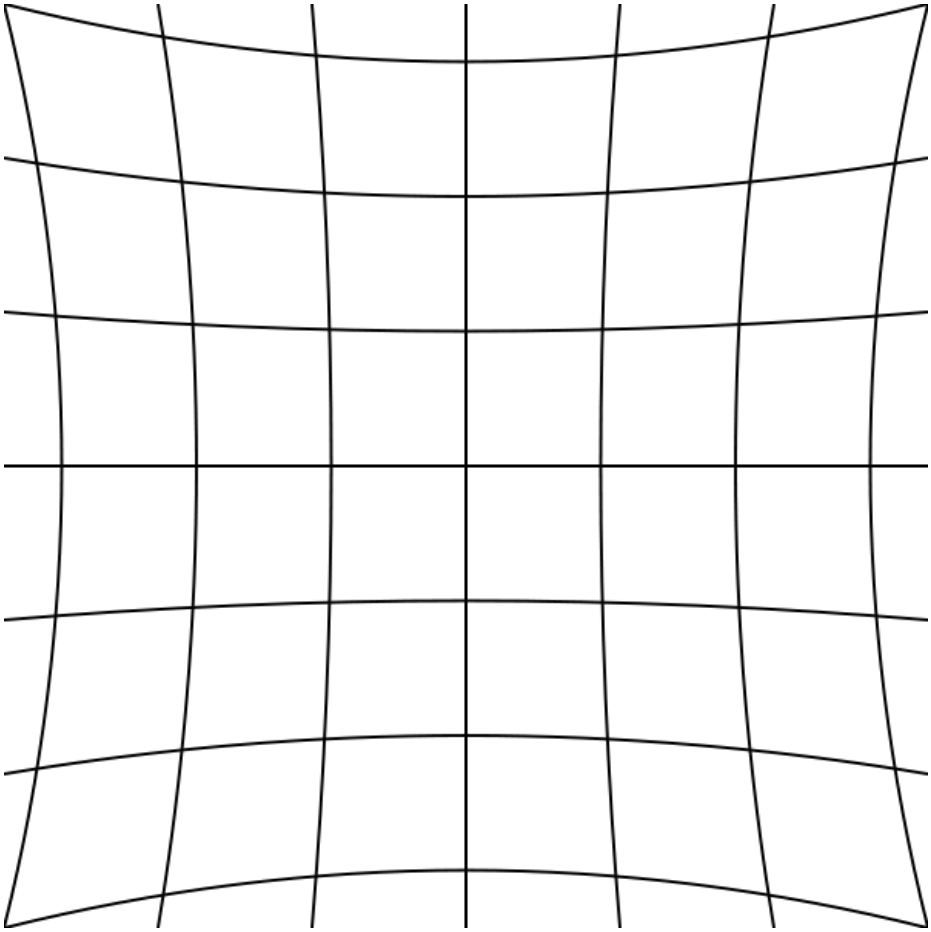


Fig. 7b: Pincushion distortion

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).

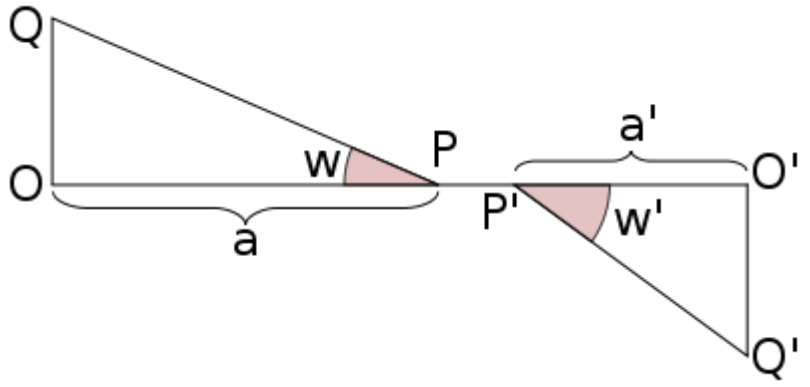


Figure 8

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 $a_0, \ldots, a_8$ 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.

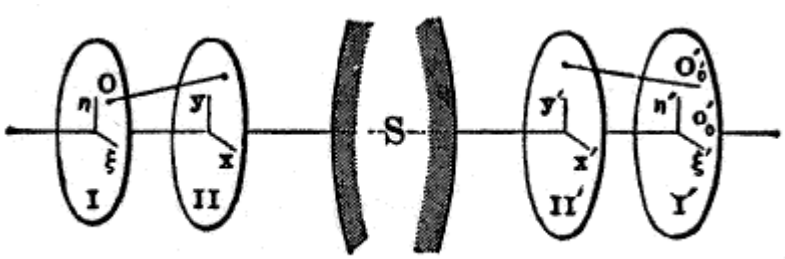


FIG. 9.

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 Dξ' and Dη' 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 Dξ' and Dη' 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) $df = k_1 dn_1 + k_2 dn_2$. 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).

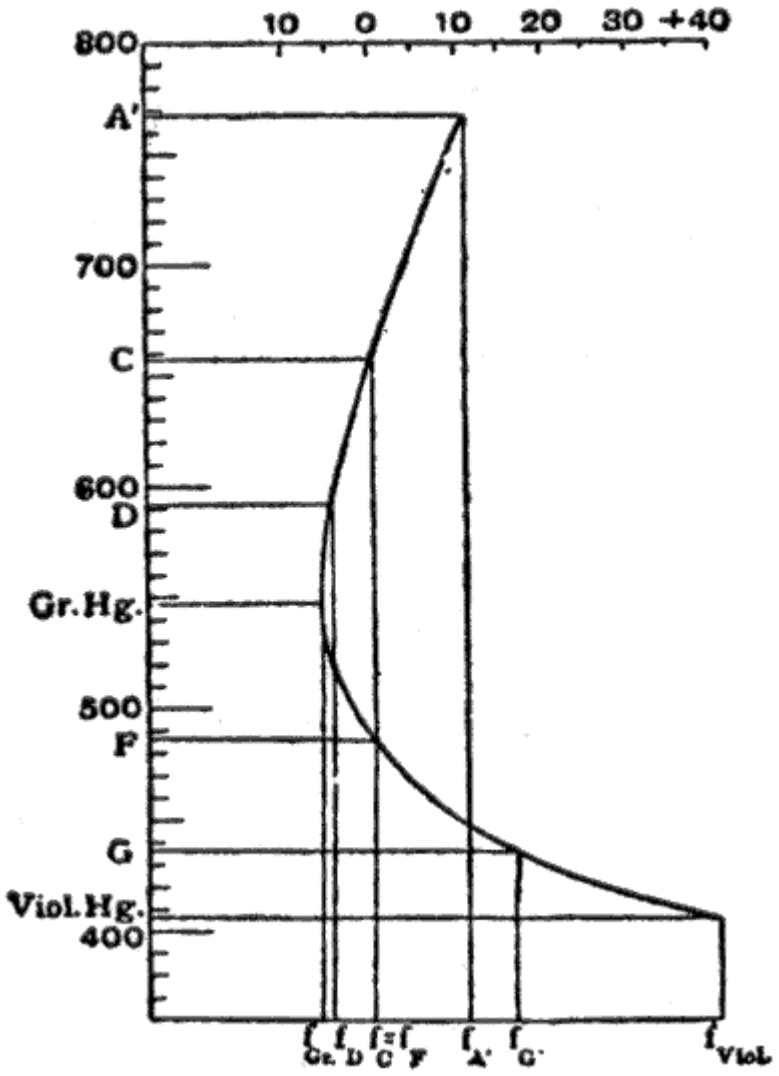
Instead of making df vanish, a certain value can be assigned to it which will produce, by the addition of the two lenses, any desired chromatic deviation, e.g. sufficient to eliminate one present in other parts of the system. If the lenses I. and II. be cemented and have the same refractive index for 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 $D = v_1 f_1 + v_2 f_2$; 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

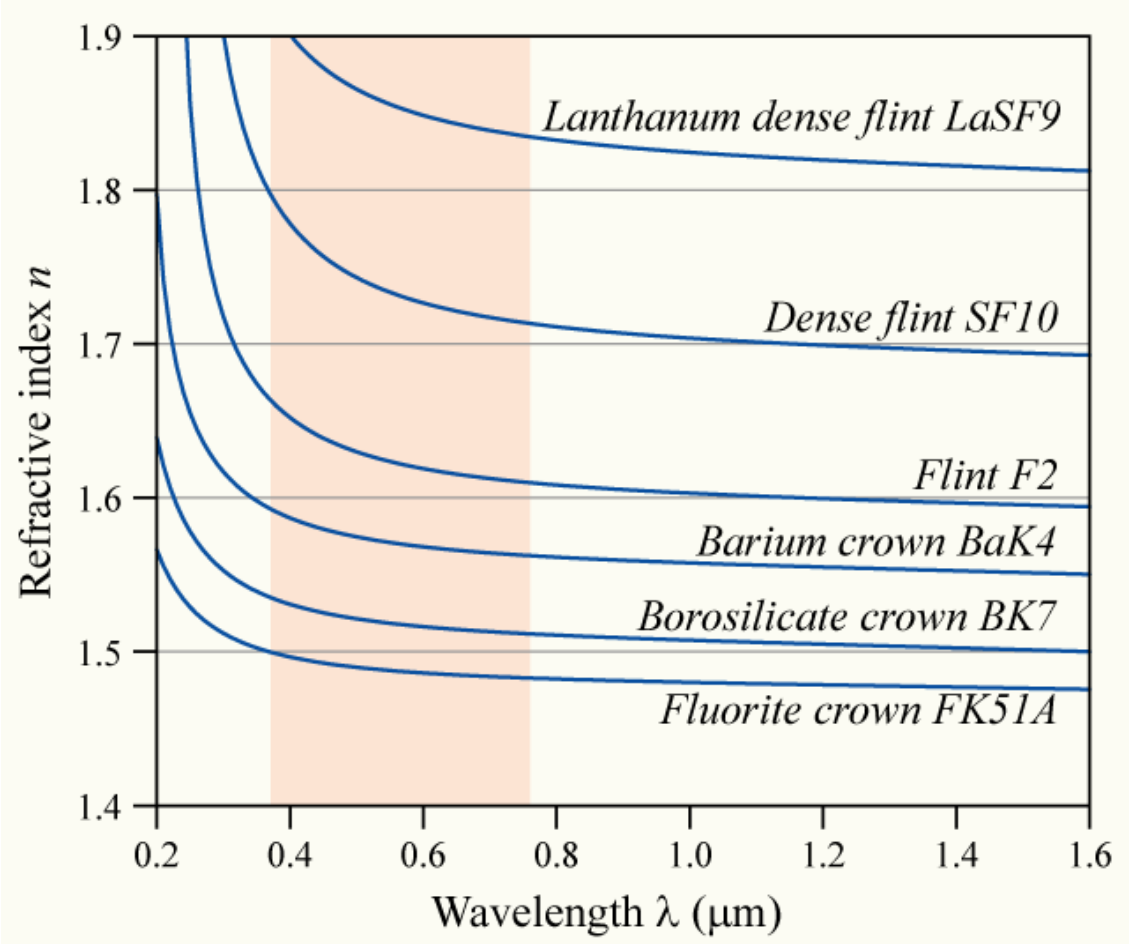
Dispersion (Optics)

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.

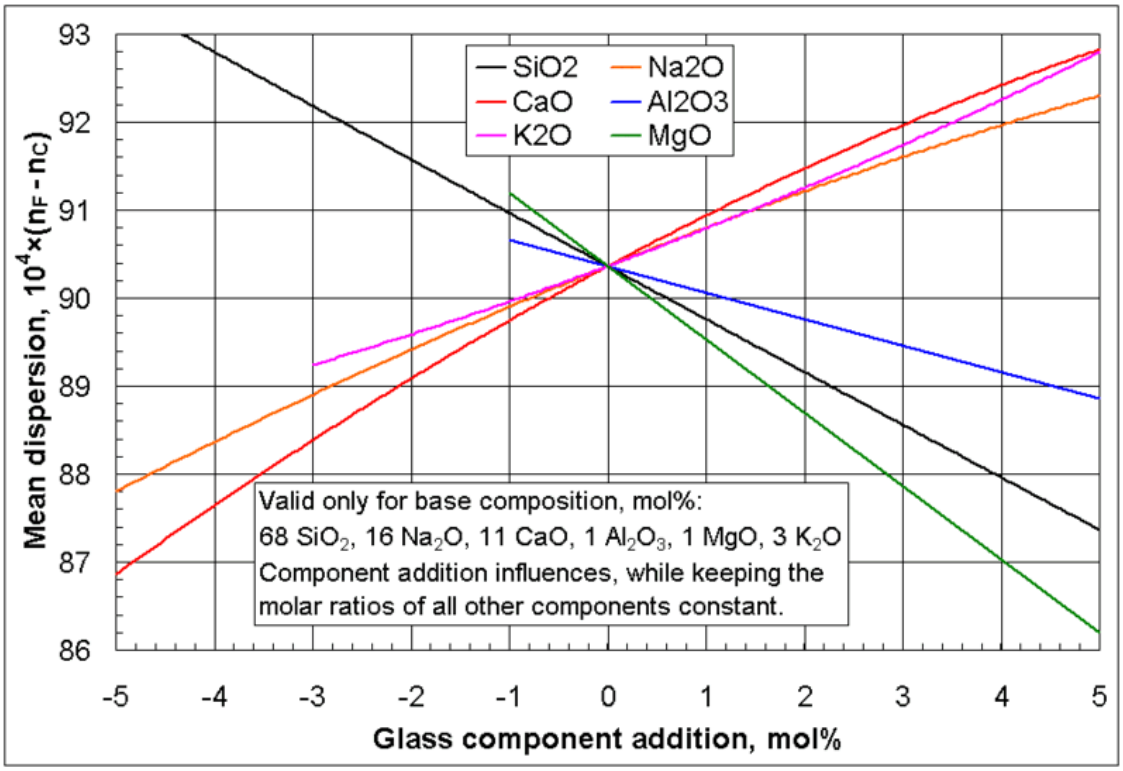
The most familiar example of dispersion is probably a rainbow, in which dispersion causes the spatial separation of a white light into components of different wavelengths (different colors). However, dispersion also has an effect in many other circumstances: for example, GVD causes pulses to spread in 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 $\lambda = 486$ nm (blue), n_C valid for $\lambda = 656$ 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 $\omega(\beta)$ 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 $\beta(\omega)$ 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 ($1\mathrm{pc/cm}^3 = 30.857 \times 10^{21}\ \mathrm{m}^{-2}$).

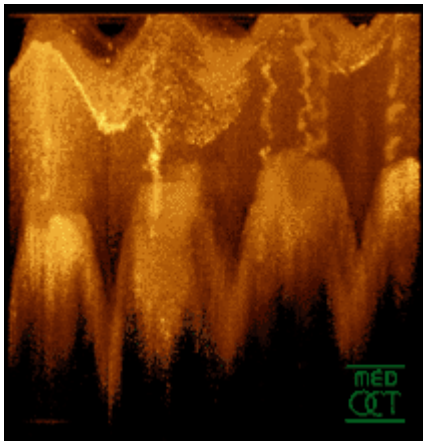
Typically for astronometric observations, this delay cannot be measured directly, since the emission time is unknown. What *can* be measured is the difference in arrival times at two different frequencies. The delay ΔT between a high frequency ν_{hi} and a low frequency ν_{lo} component of a pulse will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

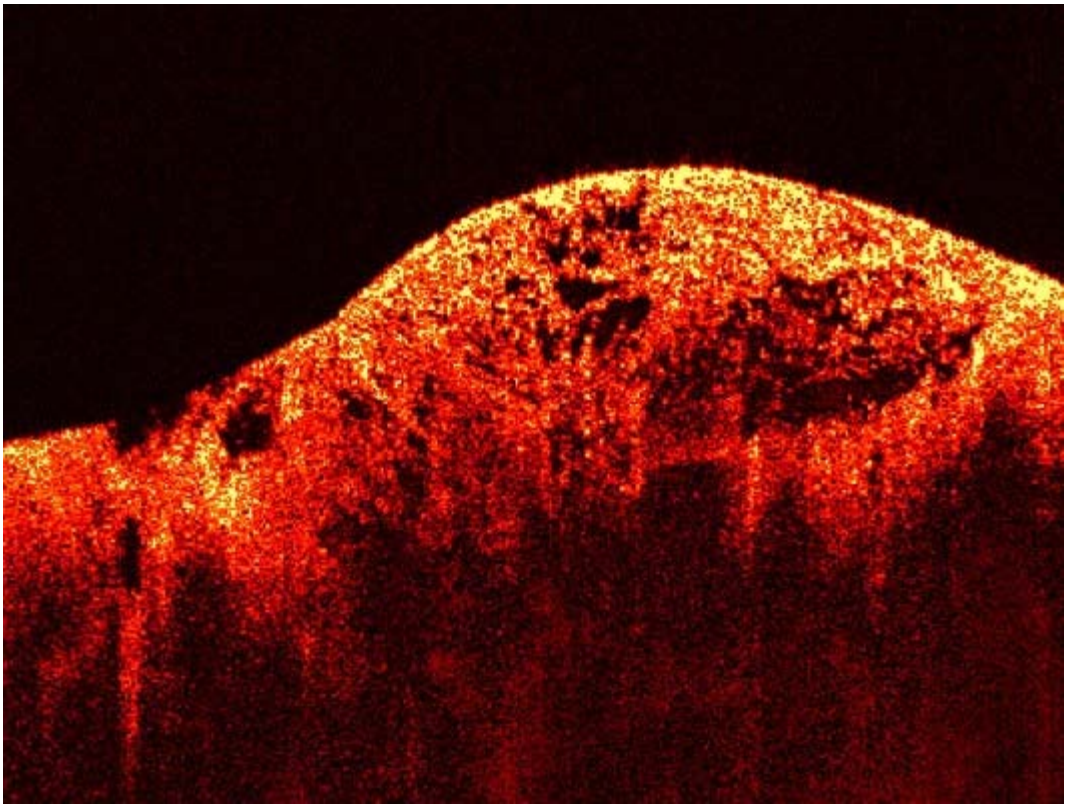
Optical coherence tomography (OCT) is an optical signal acquisition and processing method. It captures micrometer-resolution, three-dimensional images from within optical scattering media (e.g., biological tissue). Optical coherence tomography is an interferometric technique, typically employing near-infrared 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.

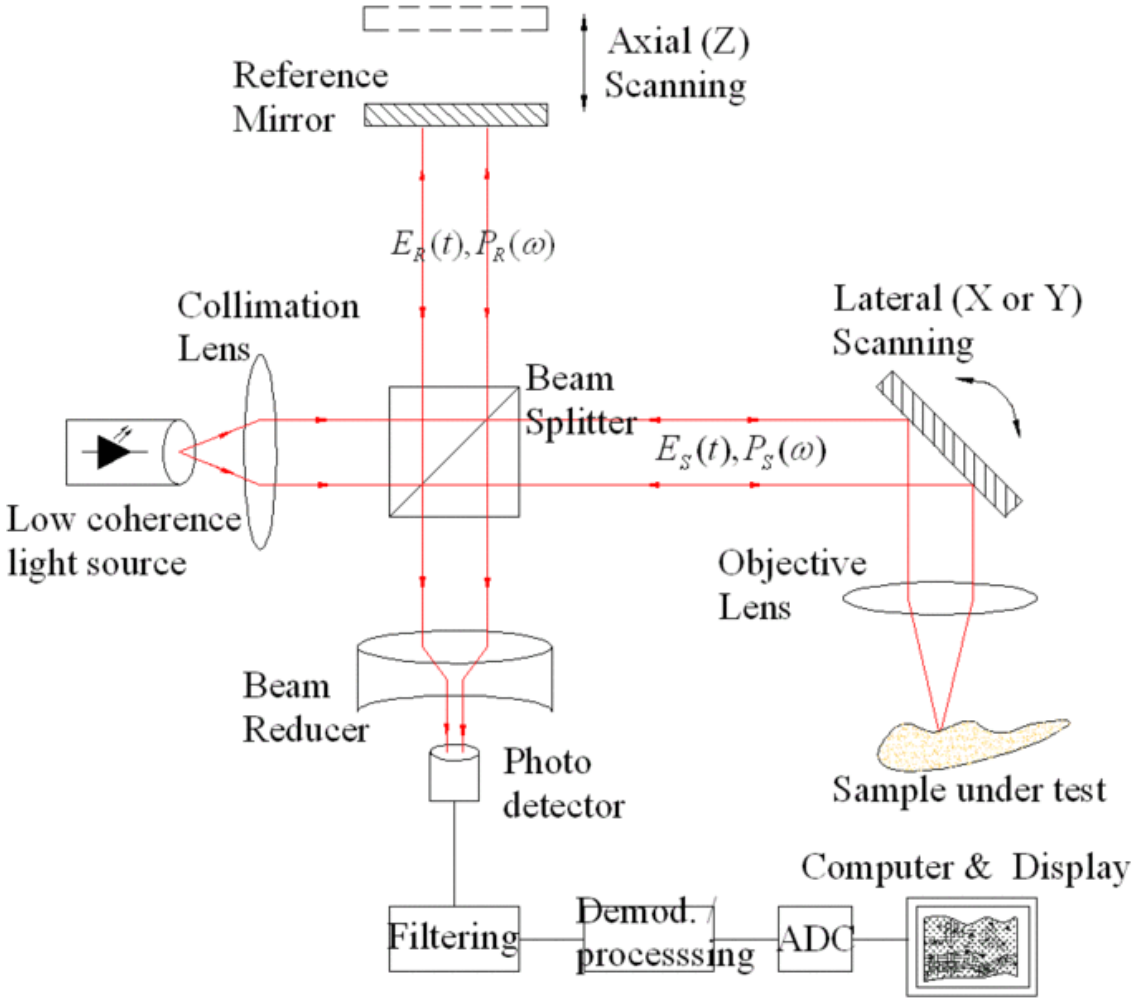


Fig. 2 Typical optical setup of single point OCT. Scanning the light beam on the sample enables non-invasive cross-sectional imaging up to 3 mm in depth with micrometer resolution.

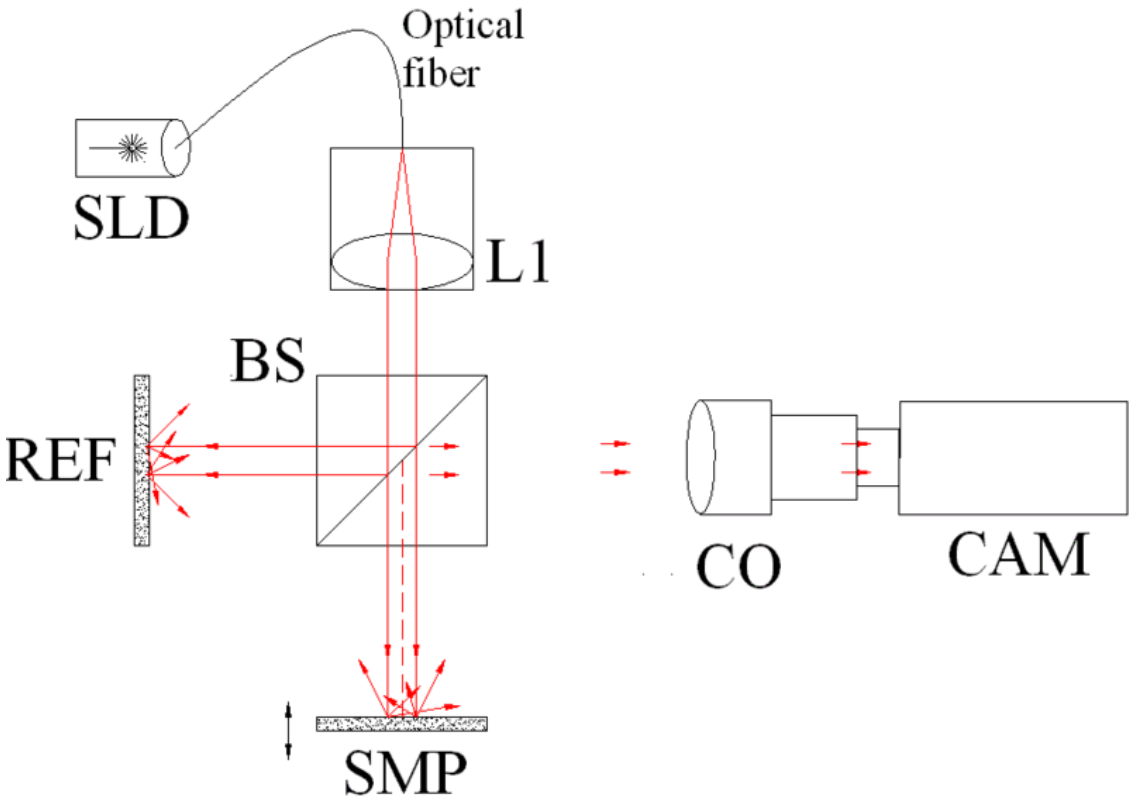


Fig. 1 Full-field OCT optical setup. Components include: super-luminescent diode (SLD), convex lens (L1), 50/50 beamsplitter (BS), camera objective (CO), CMOS-DSP camera (CAM), reference (REF) and sample (SMP). The camera functions as a two-dimensional detector array, and with the OCT technique facilitating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

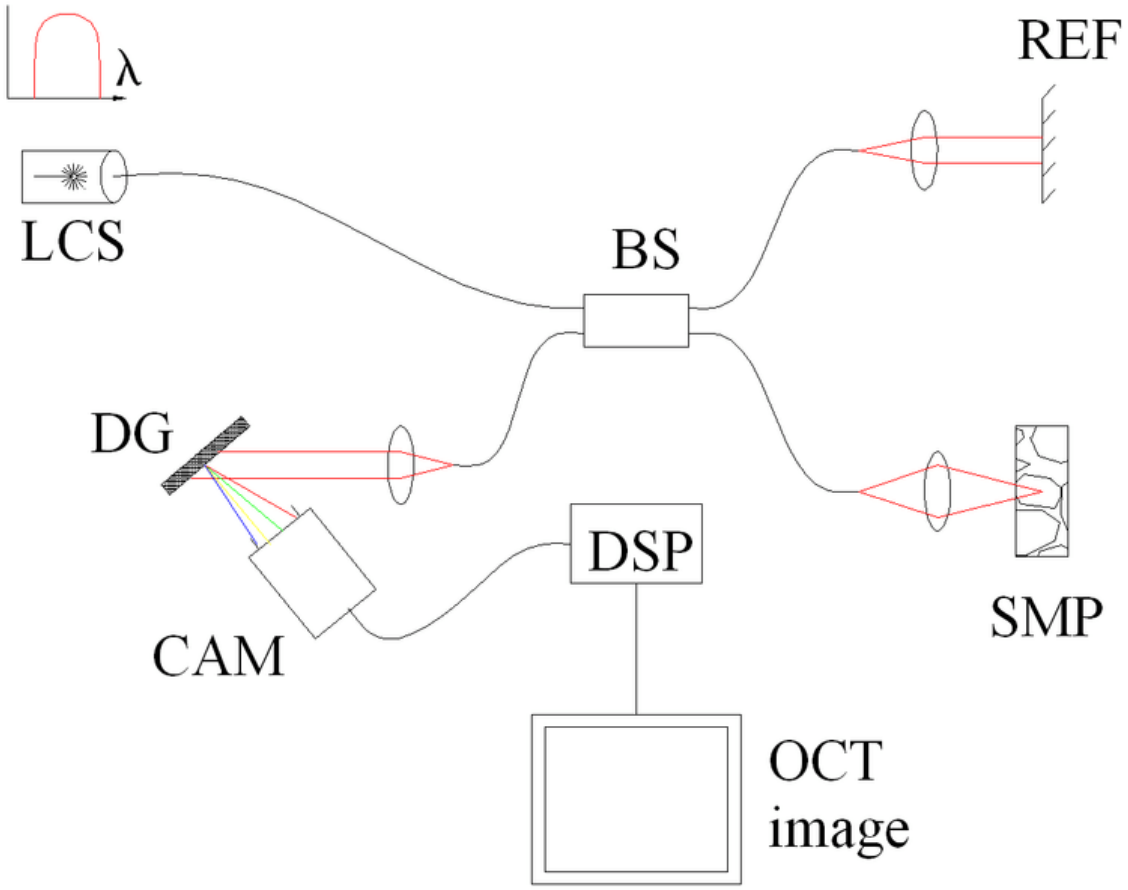


Fig. 4 Spectral discrimination by fourier-domain OCT. Components include: low coherence source (LCS), beamsplitter (BS), reference mirror (REF), sample (SMP), diffraction grating (DG) and full-field detector (CAM) act as a spectrometer, and digital signal processing (DSP)

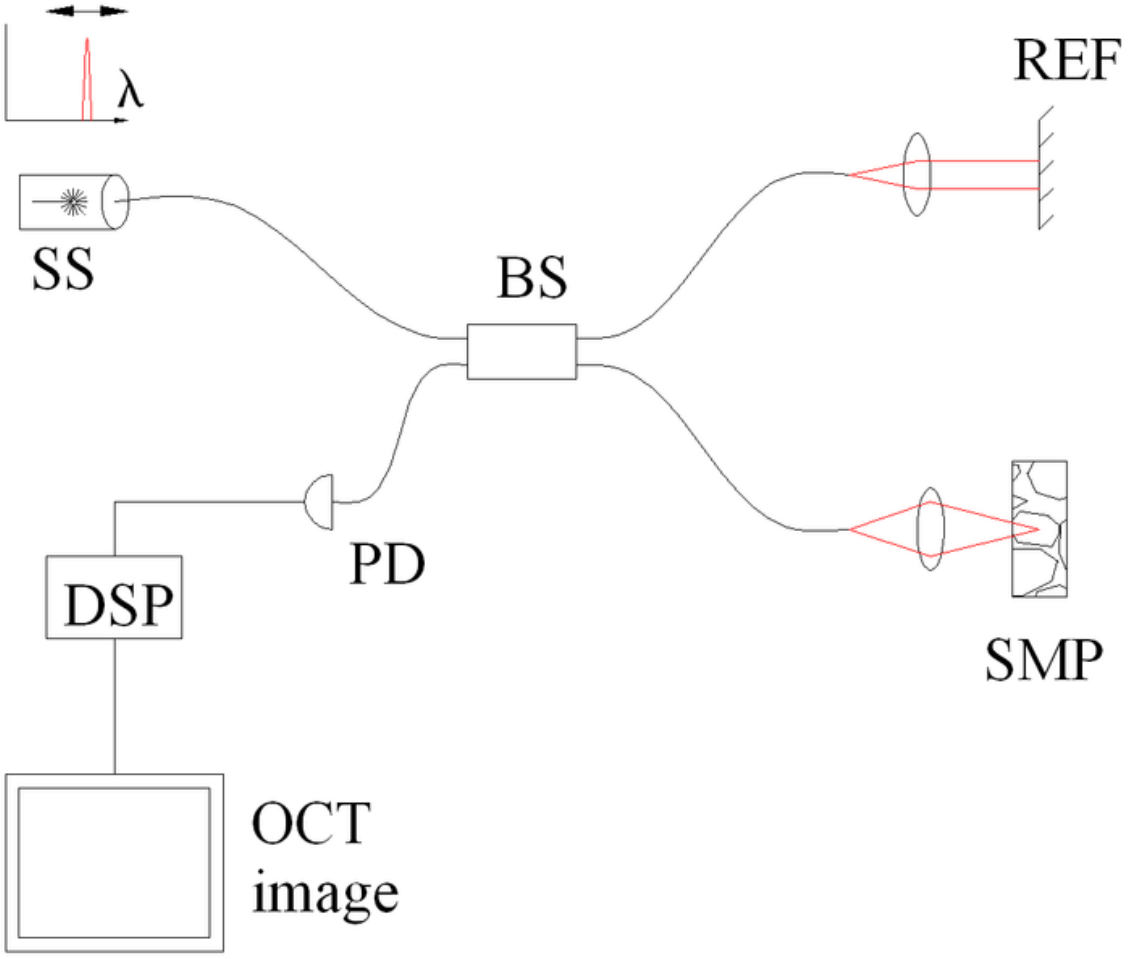


Fig. 3 Spectral discrimination by swept-source OCT. Components include: swept source or tunable laser (SS), beamsplitter (BS), reference mirror (REF), sample (SMP), photodetector (PD), digital signal processing (DSP)

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

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

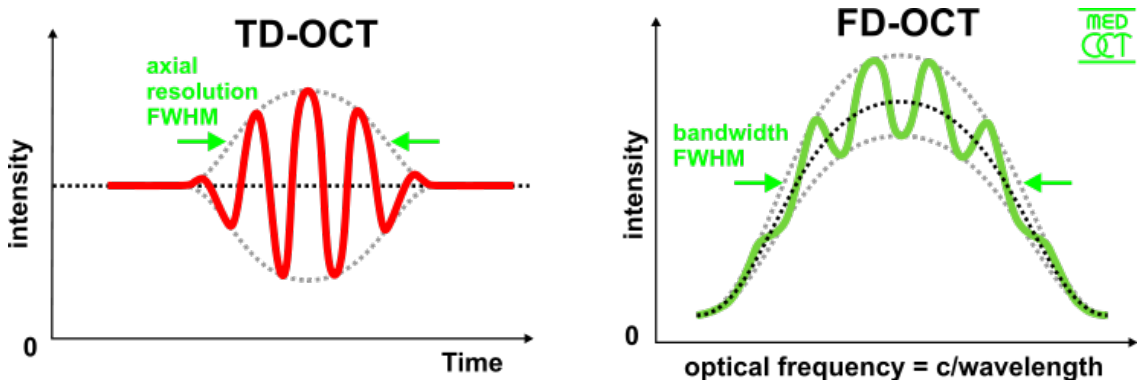
where $k_1 + k_2 < 1$ represents the interferometer beam splitting ratio, and $\gamma(\tau)$ is called the complex degree of coherence, i.e. the interference envelope and carrier dependent on reference arm scan or time delay τ, and whose recovery of interest in OCT. Due to the coherence gating effect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

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

where $\Delta\nu$ 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

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

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.

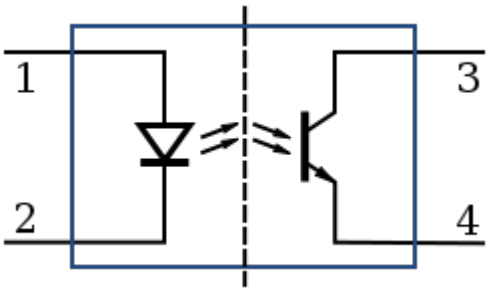
Selected applications

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

Opto-Isolator



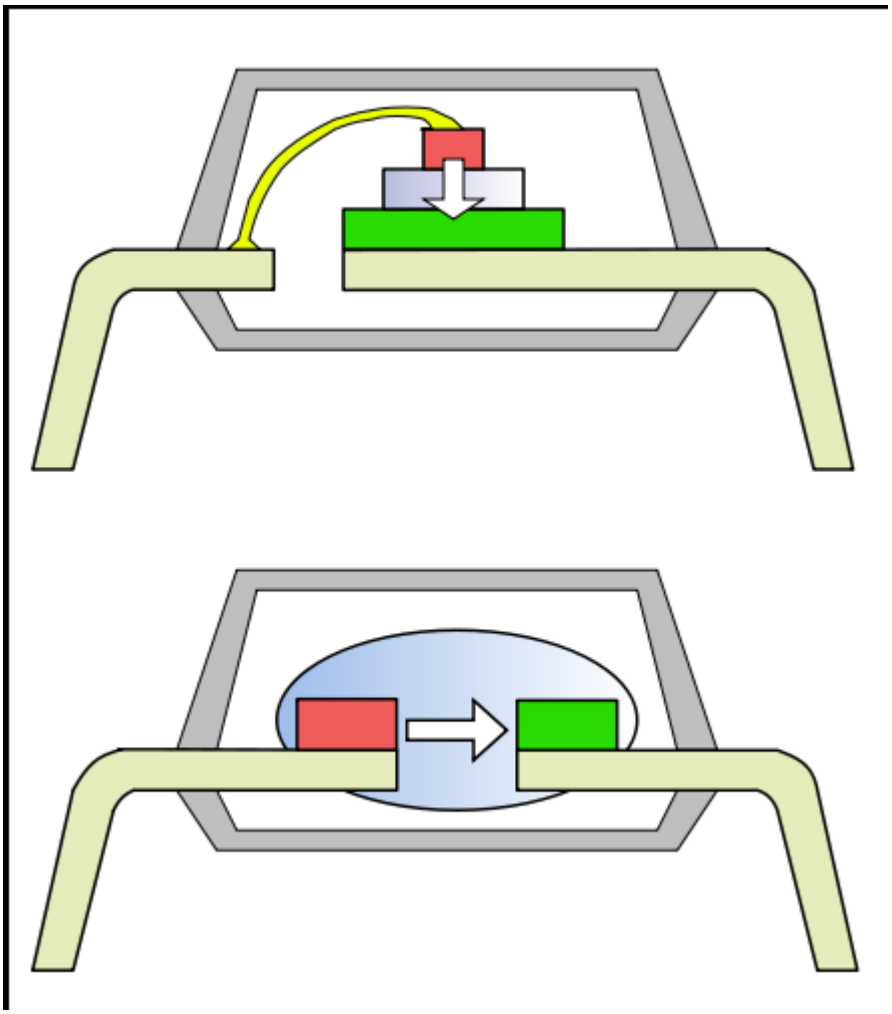
Schematic diagram of an opto-isolator showing source of light (LED) on the left, dielectric barrier in the center, and sensor (phototransistor) on the right.

In electronics, an **opto-isolator**, also called an **optocoupler**, **photocoupler**, or **optical isolator**, is "an electronic device designed to transfer electrical signals by utilizing light waves to provide coupling with electrical isolation between its input and output". The main purpose of an opto-isolator is "to prevent high voltages or rapidly changing voltages on one side of the circuit from damaging components or distorting transmissions on the other side." Commercially available opto-isolators withstand input-to-output voltages up to 10 kV and voltage transients with speeds up to 10 kV/μs.

An opto-isolator contains a source (emitter) of light, almost always a near infrared light-emitting diode (LED), that converts electrical input signal into light, a closed optical channel (also called dielectrical channel), and a photosensor, which detects incoming light and either generates electric energy directly, or modulates electric current flowing from an external power supply. The sensor can be a photoresistor, a photodiode, a phototransistor, a silicon-controlled rectifier (SCR) or a triac. Because LEDs can sense light in addition to emitting it, construction of symmetrical, bidirectional opto-isolators is possible. An optocoupled solid state relay contains a photodiode opto-isolator which drives a power switch, usually a complementary pair of MOSFET transistors. A slotted optical switch contains a source of light and a sensor, but its optical channel is open, allowing modulation of light by external objects obstructing the path of light or reflecting light into the sensor.

Photoresistor-based opto-isolators were introduced in the 1960s. They are the slowest, but also the most linear isolators and still retain a niche market in audio and music industry. Commercialization of LED technology in 1968–1970 caused a boom in optoelectronics, and by the end of the 1970s the industry developed all principal types of opto-isolators. The majority of opto-isolators on the market use bipolar silicon phototransistor sensors. They attain medium data transfer speed, sufficient for applications like electroencephalography. The fastest opto-isolators use PIN diodes in photoconductive mode and contain electronic circuitry for amplification, shaping and interfacing of the signal detected by the sensor, and can attain data transfer rates of 50 MBd. Their role in computing and communications is being challenged by new integrated isolation devices based on microminiature transformers, capacitive coupling or spin valves.

Electric isolation



Planar (top) and silicone dome (bottom) layouts - cross-section through a standard dual in-line package.

Electronic equipment and signal and power transmission lines can be subjected to voltage surges induced by lightning, electrostatic discharge, radio frequency transmissions, switching pulses (spikes) and perturbations in power supply. Remote lightning strikes can induce surges up to 10 kV, one thousand times more than the voltage limits of many electronic components. A circuit can also incorporate high voltages by design, in which case it needs safe, reliable means of interfacing its high-voltage components with low-voltage ones.

The main function of an opto-isolator is to block such high voltages and voltage transients, so that a surge in one part of the system will not disrupt or destroy the other parts. Or, according to the authors of *The Art of Electronics*, "in a nutshell, opto-couplers let you send digital (and sometimes analog) signals between circuits with separate grounds." Historically, this function was delegated to isolation transformers, which use inductive coupling between galvanically isolated input and output sides. Transformers and opto-isolators are the only two classes of electronic devices that offer *reinforced protection* — they protect both the equipment *and* the human user operating this equipment. They contain a single physical isolation barrier, but provide protection equivalent to double isolation. Safety, testing and approval of opto-couplers are regulated by national and international standards: IEC 60747-5-2, EN (CENELEC) 60747-5-2, UL 1577, CSA Component Acceptance Notice #5, etc. Opto-isolator specifications published by manufacturers always follow at least one of these regulatory frameworks.

An opto-isolator connects input and output sides with a beam of light modulated by input current. It transforms useful input signal into light, sends it across the dielectric channel, captures light on the output side and transforms it back into electric signal. Unlike transformers, which pass energy in both directions with very low losses, opto-isolators are unidirectional and they cannot transmit *power*. Typical opto-isolators can only modulate the flow of energy already present on the output side. Unlike transformers, opto-isolators can pass DC or slow-moving signals and do not require matching impedances between input and output sides. Both transformers and opto-isolators are effective in breaking ground loops, common in industrial and stage equipment, caused by high or noisy return currents in ground wires.

The physical layout of an opto-isolator depends primarily on the desired isolation voltage. Devices rated for less than a few kV have planar (or sandwich) construction. The sensor die is mounted directly on the lead frame of its package (usually, a six-pin or a four-pin dual in-line package). The sensor is covered with a sheet of glass or clear plastic, which is topped with the LED die. The LED beam fires downward. To minimize losses of light, the useful absorption spectrum of the sensor must match the output spectrum of the LED, which almost invariably lies in the near infrared. The optical channel is made as thin as possible for a desired breakdown voltage. For example, to be rated for short-term voltages of 3.75 kV and transients of 1 kV/μs, the clear polyimide sheet in the Avago ASSR-300 series is only 0.08 mm thick. Breakdown voltages of planar assemblies depend on the thickness of the transparent sheet and the configuration of bonding wires that connect the dies with external pins. Real in-circuit isolation voltage is further reduced by creepage over the PCB and the surface of the package. Safe design rules

require a minimal clearance of 25 mm/kV for bare metal conductors or 8.3 mm/kV for coated conductors.

Opto-isolators rated for 2.5 to 6 kV employ a different layout called *silicone (sic) dome*. Here, the LED and sensor dies are placed on the opposite sides of the package; the LED fires into the sensor horizontally. The LED, the sensor and the gap between them are encapsulated in a blob, or dome, of transparent silicone. The dome acts as a reflector, retaining all stray light and reflecting it onto the surface of the sensor, minimizing losses in a relatively long optical channel. In *double mold* designs the space between the silicone blob ("inner mold") and the outer shell ("outer mold") is filled with dark dielectric compound with a matched coefficient of thermal expansion.

Types of opto-isolators

Device type	Source of light	Sensor type	Speed	Current transfer ratio
Resistive opto-isolator (Vactrol)	Incandescent light bulb	CdS or CdSe photoresistor (LDR)	Very low	<100%
	Neon lamp		Low	
	GaAs infrared LED		Low	
Diode opto-isolator	GaAs infrared LED	Silicon photodiode	Highest	0.1% - 0.2%
Transistor opto-isolator	GaAs infrared LED	Bipolar silicon phototransistor	Medium	2% - 120%
		Darlington phototransistor	Medium	100% - 600%
Opto-isolated SCR	GaAs infrared LED	Silicon-controlled rectifier	Low to medium	>100%
Opto-isolated triac	GaAs infrared LED	TRIAC	Low to medium	Very high
Opto-isolated maus	DoNs infrared LED	TRIAC	Low to high	Extremely high
Solid-state relay	Stack of GaAs infrared LEDs	Stack of photodiodes driving a pair of MOSFETs or an IGBT	Low to high	Practically unlimited

Resistive opto-isolators

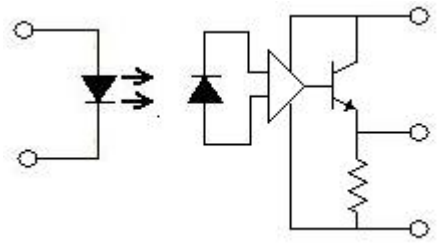
The earliest opto-isolators, originally marketed as *light cells*, emerged in the 1960s. They employed miniature incandescent light bulbs as sources of light, and cadmium sulfide (CdS) or cadmium selenide (CdSe) photoresistors (also called light-dependent resistors, LDRs) as receivers. In applications where control linearity was not important, or where available current was too low for driving an incandescent bulb (as was the case in vacuum tube amplifiers), it was replaced with a neon lamp. These devices (or just their LDR component) were commonly named *Vactrols*, after a trademark of Vactec, Inc. The trademark has since been genericized, but the original Vactrols are still being manufactured by PerkinElmer.

The turn-on and turn-off lag of an incandescent bulb lies in hundreds of milliseconds range, which makes the bulb an effective low-pass filter and rectifier but limits the practical modulation frequency range to a few Hertz. With the introduction of light-emitting diodes (LEDs) in 1968–1970, the manufacturers replaced incandescent and neon lamps with LEDs and achieved response times of 5 milliseconds and modulation frequencies up to 250 Hz. The name *Vactrol* was carried over on LED-based devices which are, as of 2010, still produced in small quantities.

Photoresistors used in opto-isolators rely on bulk effects in a uniform film of semiconductor; there are no p-n junctions. Uniquely among photosensors, photoresistors are non-polar devices suited for either AC or DC circuits. Their resistance drops in reverse proportion to the intensity of incoming light, from virtually infinity to a residual floor that may be as low as less than a hundred Ohms. These properties made the original Vactrol a convenient and cheap automatic gain control and compressor for telephone networks. The photoresistors easily withstood voltages up to 400 Volts, which made them ideal for driving vacuum fluorescent displays. Other industrial applications included photocopiers, industrial automation, professional light measurement instruments and auto-exposure meters. Most of these applications are now obsolete, but resistive opto-isolators retained a niche in audio, in particular guitar amplifier, markets.

American guitar and organ manufacturers of the 1960s embraced the resistive opto-isolator as a convenient and cheap tremolo modulator. Fender's early tremolo effects used two vacuum tubes; after 1964 one of these tubes was replaced by an optocouple made of a LDR and a neon lamp. To date, Vactrols activated by pressing the stompbox pedal are ubiquitous in the music industry. Shortages of genuine PerkinElmer Vactrols forced the DIY guitar community to "roll their own" resistive opto-isolators. Guitarists to date prefer opto-isolated effects because their superior separation of audio and control grounds results in "inherently high quality of the sound". However, the distortion introduced by a photoresistor at line level signal is higher than that of a professional electrically-coupled voltage-controlled amplifier. Performance is further compromised by slow fluctuations of resistance owing to light history, a memory effect inherent in cadmium compounds. Such fluctuations take hours to settle and can be only partially offset with feedback in the control circuit.

Photodiode opto-isolators



A fast photodiode opto-isolator with an output-side amplifier circuit.

Diode opto-isolators employ LEDs as sources of light and silicon photodiodes as sensors. When the photodiode is reverse-biased with an external voltage source, incoming light increases the reverse current flowing through the diode. The diode itself does not generate energy; it modulates the flow of energy from an external source. This mode of operation is called photoconductive mode. Alternatively, in the absence of external bias the diode converts the energy of light into electric energy by charging its terminals to a voltage of up to 0.7 V. The rate of charge is proportional to the intensity of incoming light. The energy is harvested by draining the charge through an external high-impedance path; the ratio of current transfer can reach 0.2%. This mode of operation is called photovoltaic mode.

The fastest opto-isolators employ PIN diodes in photoconductive mode. The response times of PIN diodes lie in the subnanosecond range; overall system speed is limited by delays in LED output and in biasing circuitry. To minimize these delays, fast digital opto-isolators contain their own LED drivers and output amplifiers optimized for speed. These devices are called *full logic opto-isolators*: their LEDs and sensors are fully encapsulated within a digital logic circuit. The Hewlett-Packard 6N137/HPCL2601 family of devices equipped with internal output amplifiers was introduced in the late 1970s and attained 10 MBd data transfer speeds. It remained an industry standard until the introduction of the 50 MBd Agilent Technologies 7723/0723 family in 2002. The 7723/0723 series opto-isolators contain CMOS LED drivers and a CMOS buffered amplifiers, which require two independent external power supplies of 5 V each.

Photodiode opto-isolators can be used for interfacing analog signals, although their non-linearity invariably distorts the signal. A special class of analog opto-isolators introduced by Burr-Brown uses *two* photodiodes and an input-side operational amplifier to compensate for diode non-linearity. One of two identical diodes is wired into the feedback loop of the amplifier, which maintains overall current transfer ratio at a constant level regardless of the non-linearity in the second (output) diode.

Solid-state relays built around MOSFET switches usually employ a photodiode opto-isolator to drive the switch. The gate of a MOSFET requires relatively small total charge to turn on and its leakage current in steady state is very low. A photodiode in photovoltaic mode can generate turn-on *charge* in a reasonably short time but it's output

voltage is many times less than the MOSFET's threshold voltage. To reach the required threshold, solid-state relays contain stacks of up to thirty photodiodes wired in series.

Phototransistor opto-isolators

Phototransistors are inherently slower than photodiodes. The earliest and the slowest but still common 4N35 opto-isolator, for example, has rise and fall times of 5 μs into a 100 Ohm load and its bandwidth is limited at around 10 kiloHertz - sufficient for applications like electroencephalography or pulse-width motor control. Devices like PC-900 or 6N138 recommended in the original 1983 Musical Instrument Digital Interface specification allow digital data transfer speeds of tens of kiloBauds. Phototransistors must be properly biased and loaded to achieve their maximum speeds, for example, the 4N28 operates at up to 50 kHz with optimum bias and less than 4 kHz without it.

Design with transistor opto-isolators requires generous allowances for wide fluctuations of parameters found in commercially available devices. Such fluctuations may be destructive, for example, when an opto-isolator in the feedback loop of a DC-to-DC converter changes its transfer function and causes spurious oscillations, or when unexpected delays in opto-isolators cause a short circuit through one side of an H-bridge. Manufacturers' datasheets typically list only worst-case values for critical parameters; actual devices surpass these worst-case estimates in an unpredictable fashion. Bob Pease observed that current transfer ratio in a batch of 4N28's can vary from 15% to more than 100%; the datasheet specified only a minimum of 10%. Transistor beta in the same batch can vary from 300 to 3000, resulting in 10:1 variance in bandwidth.

Opto-isolators using field-effect transistors (FETs) as sensors are rare and, like vactrols, can be used as remote-controlled analog potentiometers provided that the voltage across the FET's output terminal does not exceed a few hundred mV. Opto-FETs turn on without injecting switching charge in the output circuit, which is particularly useful in sample and hold circuits.

Bidirectional opto-isolators

All opto-isolators described so far are uni-directional. Optical channel always works one way, from the source (LED) to the sensor. The sensors, be it photoresistors, photodiodes or phototransistors, cannot emit light. But LEDs, like all semiconductor diodes, are capable of detecting incoming light, which makes possible construction of a two-way opto-isolator from a pair of LEDs. The simplest bidirectional opto-isolator is merely a pair of LEDs placed face to face and held together with heat-shrink tubing. If necessary, the gap between two LEDs can be extended with a glass fiber insert.

Visible spectrum LEDs have relatively poor transfer efficiency, thus near infrared spectrum GaAs, GaAs:Si and AlGaAs:Si LEDs are the preferred choice for bidirectional devices. Bidirectional opto-isolators built around pairs of GaAs:Si LEDs have current transfer ratio of around 0.06% in either photovoltaic or photoconductive mode — less than photodiode-based isolators, but sufficiently practical for real-world applications.

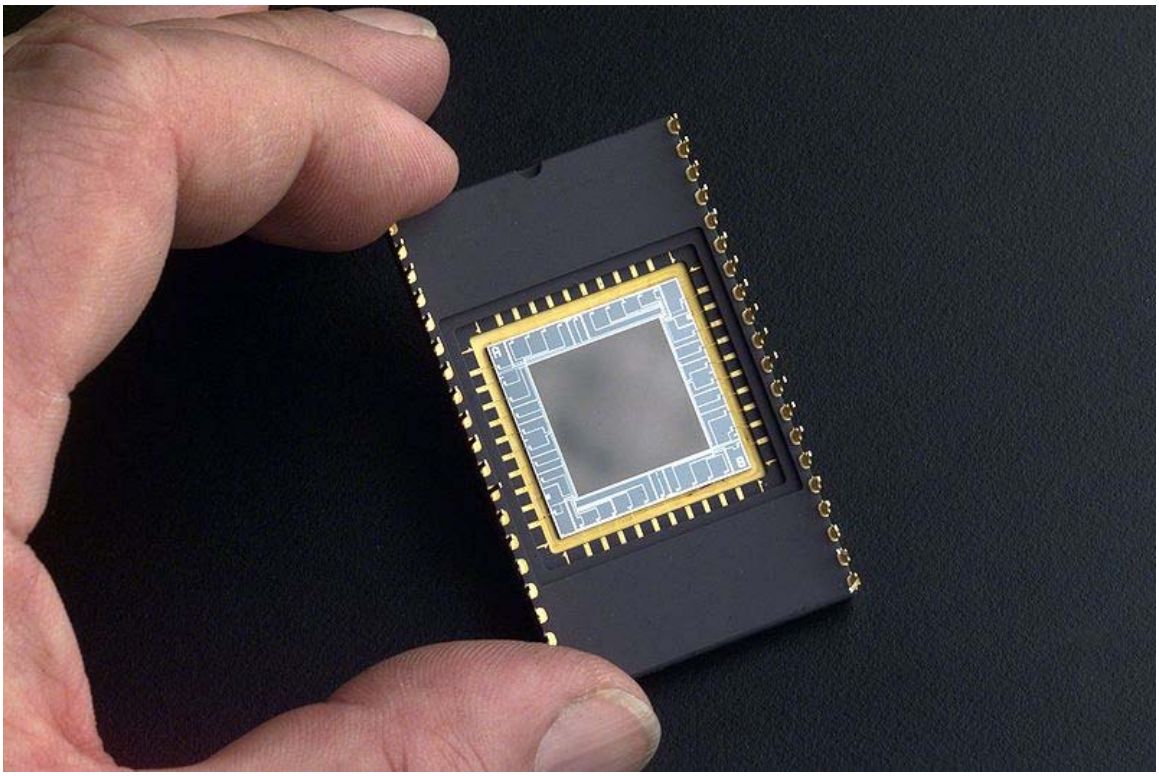
Alternatives

Opto-isolators can be too slow and bulky for modern digital applications. Since the 1990s, researchers have examined and perfected alternative, faster and more compact isolation technologies. Two of these technologies, magnetic isolators and capacitor-coupled isolators, reached the mass market in the 2000s. The third alternative, based on giant magnetoresistance, has been present on the market since 2002 in limited quantities. As of 2010, production models of all three types allow data transfer speeds of 150 MBit/s and resist voltage transients of up to 25 kV/μs, compared to 10 kV/μs for opto-isolators. Unlike opto-isolators, which are stacks of discrete LEDs and sensors, the new devices are monolithic integrated circuits, and are easily scalable into multi-bit data bus isolators.

- In 2000 Analog Devices introduced integrated magnetic isolators — electrically-decoupled 100 MBit/s, 2.5 kV isolation circuits employing air core transformers micromachined on the surface of silicon integrated circuits. They featured lesser power consumption, lesser cost and were four times faster than the fastest contemporary opto-isolators. In 2010, Analog increased the speed of their magnetic isolators to 150 MBit/s and offered isolation up to 5 kV. Microtransformer-based isolators can work as dc-dc converters, passing both signal *and* power. Commercially available ICs can carry up to four isolated digital channels and a 2 W isolated power channel in miniature 20-pin packages. According to Analog Devices, by October 2010 the company has more "than 450 million [magnetic isolator] channels deployed". In the same year NEC and Renesas announced transformer-based CMOS devices with transfer rates of 250 MBit/s.
- High-speed capacitive-coupled isolators were introduced in 2000 by Silicon Laboratories and commercialized by Texas Instruments. These devices convert an incoming data stream into an amplitude-modulated UHF signal, pass it through a silicon dioxide isolation layer, and demodulate the received signal. The spectra of spurious voltage transients, which can pass through the capacitive barrier and disrupt operation, lie far below the modulation frequency and can be effectively blocked. As of 2010, capacitive-coupled isolators offer data transfer speeds of 150 MBit/s and voltage isolation of 560 V continuous and 4 kV peak across the barrier.
- NVE Corporation, the pioneer of magnetoresistive random access memory, markets an alternative type of isolator based on giant magnetoresistance (GMR) effect (*Spintronic* and *IsoLoop* trademarks). Each isolation cell of these devices is formed by a flat square coil which is micromachined above four spin valve sensors buried in the silicon wafer. These sensors, wired into a Wheatstone bridge circuit, generate binary on/off output signals. At the time of their introduction in 2002, NVE advertised speeds 5 to 10 times higher than the fastest opto-isolators; and in March 2008 commercial devices marketed by NVE were rated for speeds up to 150 MBit/s.

Chapter 8

Charge-Coupled Device



A specially developed CCD used for ultraviolet imaging in a wire bonded package.

A **charge-coupled device** (**CCD**) is a device for the movement of electrical charge, usually from within the device to an area where the charge can be manipulated, for example conversion into a digital value. This is achieved by "shifting" the signals between stages within the device one at a time. CCDs move charge between capacitive *bins* in the device, with the shift allowing for the transfer of charge between bins.

Often the device is integrated with an image sensor, such as a photoelectric device to produce the charge that is being read, thus making the CCD a major technology for

digital imaging. Although CCDs are not the only technology to allow for light detection, CCDs are widely used in professional, medical, and scientific applications where high-quality image data are required.

History

The charge-coupled device was invented in 1969 at AT&T Bell Labs by Willard Boyle and George E. Smith. The lab was working on semiconductor bubble memory when Boyle and Smith conceived of the design of what they termed, in their notebook, "Charge 'Bubble' Devices". A description of how the device could be used as a shift register and as a linear and area imaging devices was described in this first entry. The essence of the design was the ability to transfer charge along the surface of a semiconductor from one storage capacitor to the next. The concept was similar in principle to the bucket-brigade device (BBD), which was developed at Philips Research Labs during the late 1960's.

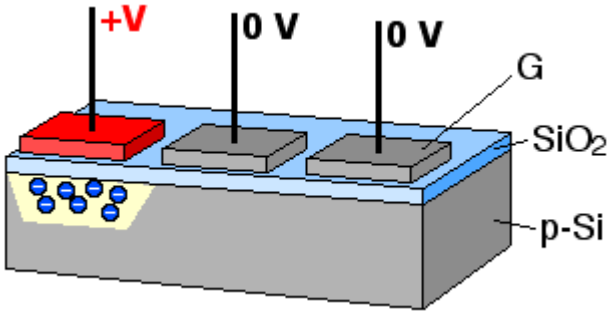
The initial paper describing the concept listed possible uses as a memory, a delay line, and an imaging device. The first experimental device demonstrating the principle was a row of closely spaced metal squares on an oxidized silicon surface electrically accessed by wire bonds.

The first working CCD made with integrated circuit technology was a simple 8-bit shift register. This device had input and output circuits and was used to demonstrate its use as a shift register and as a crude eight pixel linear imaging device. Development of the device progressed at a rapid rate. By 1971, Bell researchers Michael F. Tompsett et al. were able to capture images with simple linear devices.

Several companies, including Fairchild Semiconductor, RCA and Texas Instruments, picked up on the invention and began development programs. Fairchild's effort, led by ex-Bell researcher Gil Amelio, was the first with commercial devices, and by 1974 had a linear 500-element device and a 2-D 100 x 100 pixel device. Under the leadership of Kazuo Iwama, Sony also started a big development effort on CCDs involving a significant investment. Eventually, Sony managed to mass produce CCDs for their camcorders. Before this happened, Iwama died in August 1982. Subsequently, a CCD chip was placed on his tombstone to acknowledge his contribution.

In January 2006, Boyle and Smith were awarded the National Academy of Engineering Charles Stark Draper Prize, and in 2009 they were awarded the Nobel Prize for Physics, for their work on the CCD.

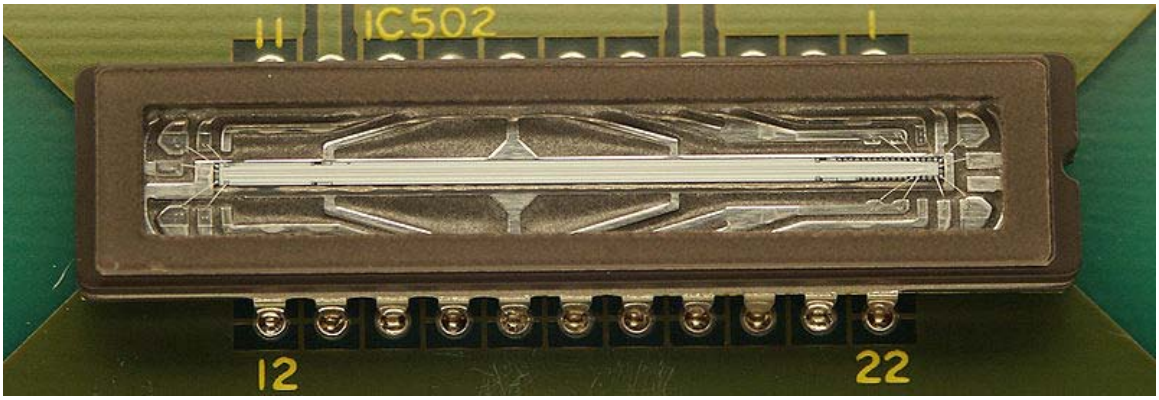
Basics of operation



The charge packets (electrons, blue) are collected in potential wells (yellow) created by applying positive voltage at the gate electrodes (G). Applying positive voltage to the gate electrode in the correct sequence transfers the charge packets.

In a CCD for capturing images, there is a photoactive region (an epitaxial layer of silicon), and a transmission region made out of a shift register (the CCD, properly speaking).

An image is projected through a lens onto the capacitor array (the photoactive region), causing each capacitor to accumulate an electric charge proportional to the light intensity at that location. A one-dimensional array, used in line-scan cameras, captures a single slice of the image, while a two-dimensional array, used in video and still cameras, captures a two-dimensional picture corresponding to the scene projected onto the focal plane of the sensor. Once the array has been exposed to the image, a control circuit causes each capacitor to transfer its contents to its neighbor (operating as a shift register). The last capacitor in the array dumps its charge into a charge amplifier, which converts the charge into a voltage. By repeating this process, the controlling circuit converts the entire contents of the array in the semiconductor to a sequence of voltages. In a digital device, these voltages are then sampled, digitized, and usually stored in memory; in an analog device (such as an analog video camera), they are processed into a continuous analog signal (e.g. by feeding the output of the charge amplifier into a low-pass filter) which is then processed and fed out to other circuits for transmission, recording, or other processing.



"One-dimensional" CCD image sensor from a fax machine.

Detailed physics of operation

The photoactive region of the CCD is, generally, an epitaxial layer of silicon. It has a doping of p+ (Boron) and is grown upon a substrate material, often p++. In buried channel devices, the type of design utilized in most modern CCDs, certain areas of the surface of the silicon are ion implanted with phosphorus, giving them an n-doped designation. This region defines the channel in which the photogenerated charge packets will travel. The gate oxide, i.e. the capacitor dielectric, is grown on top of the epitaxial layer and substrate.

Later on in the process polysilicon gates are deposited by chemical vapor deposition, patterned with photolithography, and etched in such a way that the separately phased gates lie perpendicular to the channels. The channels are further defined by utilization of the LOCOS process to produce the channel stop region.

Channel stops are thermally grown oxides that serve to isolate the charge packets in one column from those in another. These channel stops are produced before the polysilicon gates are, as the LOCOS process utilizes a high temperature step that would destroy the gate material. The channels stops are parallel to, and exclusive of, the channel, or "charge carrying", regions.

Channel stops often have a p+ doped region underlying them, providing a further barrier to the electrons in the charge packets (this discussion of the physics of CCD devices assumes an electron transfer device, though hole transfer, is possible).

The clocking of the gates, alternately high and low, will forward and reverse bias to the diode that is provided by the buried channel (n-doped) and the epitaxial layer (p-doped). This will cause the CCD to deplete, near the p-n junction and will collect and move the charge packets beneath the gates—and within the channels—of the device.

CCD manufacturing and operation can be optimized for different uses. The above process describes a frame transfer CCD. While CCDs may be manufactured on a heavily doped

p++ wafer it is also possible to manufacture a device inside p-wells that have been placed on an n-wafer. This second method, reportedly, reduces smear, dark current, and infrared and red response. This method of manufacture is used in the construction of interline transfer devices.

Another version of CCD is called a peristaltic CCD. In a peristaltic charge-coupled device, the charge packet transfer operation is analogous to the peristaltic contraction and dilation of the digestive system. The peristaltic CCD has an additional implant that keeps the charge away from the silicon/silicon dioxide interface and generates a large lateral electric field from one gate to the next. This provides an additional driving force to aid in transfer of the charge packets.

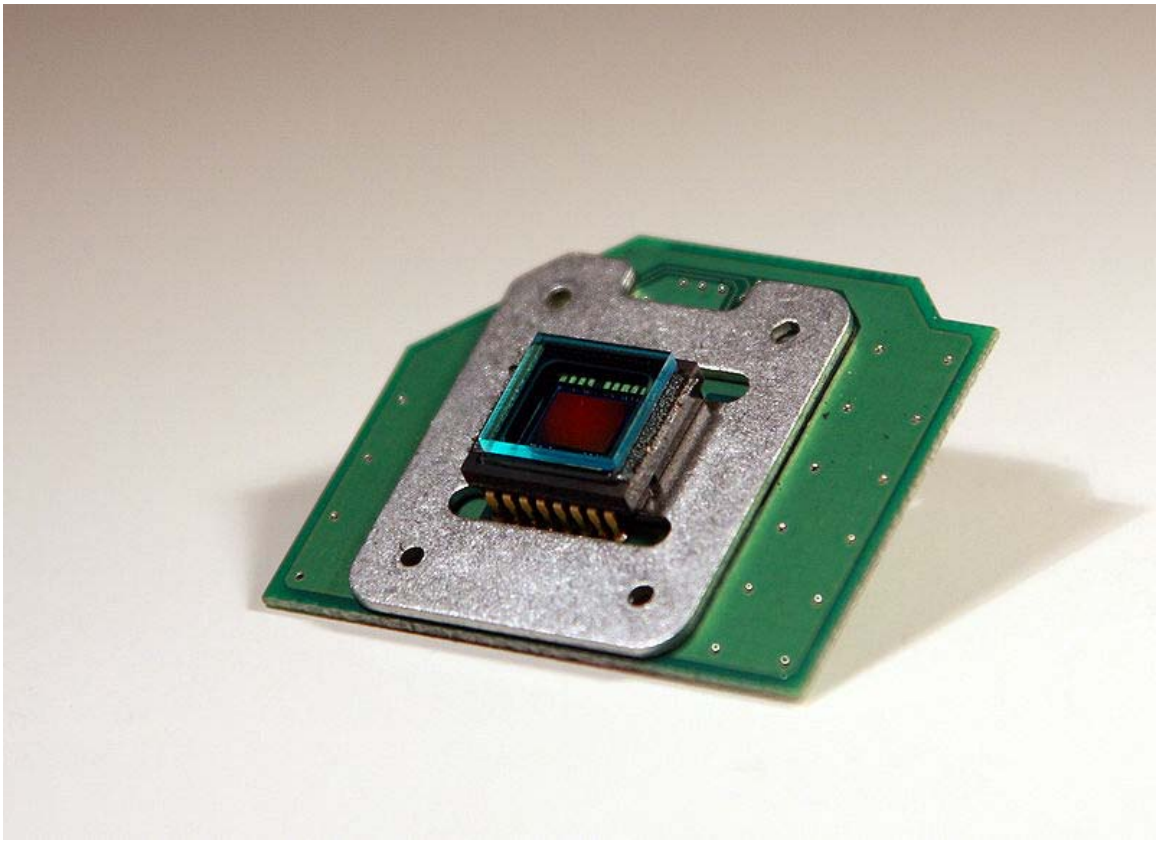
Architecture

The CCD image sensors can be implemented in several different architectures. The most common are full-frame, frame-transfer, and interline. The distinguishing characteristic of each of these architectures is their approach to the problem of shuttering.

In a full-frame device, all of the image area is active, and there is no electronic shutter. A mechanical shutter must be added to this type of sensor or the image smears as the device is clocked or read out.

With a frame-transfer CCD, half of the silicon area is covered by an opaque mask (typically aluminum). The image can be quickly transferred from the image area to the opaque area or storage region with acceptable smear of a few percent. That image can then be read out slowly from the storage region while a new image is integrating or exposing in the active area. Frame-transfer devices typically do not require a mechanical shutter and were a common architecture for early solid-state broadcast cameras. The downside to the frame-transfer architecture is that it requires twice the silicon real estate of an equivalent full-frame device; hence, it costs roughly twice as much.

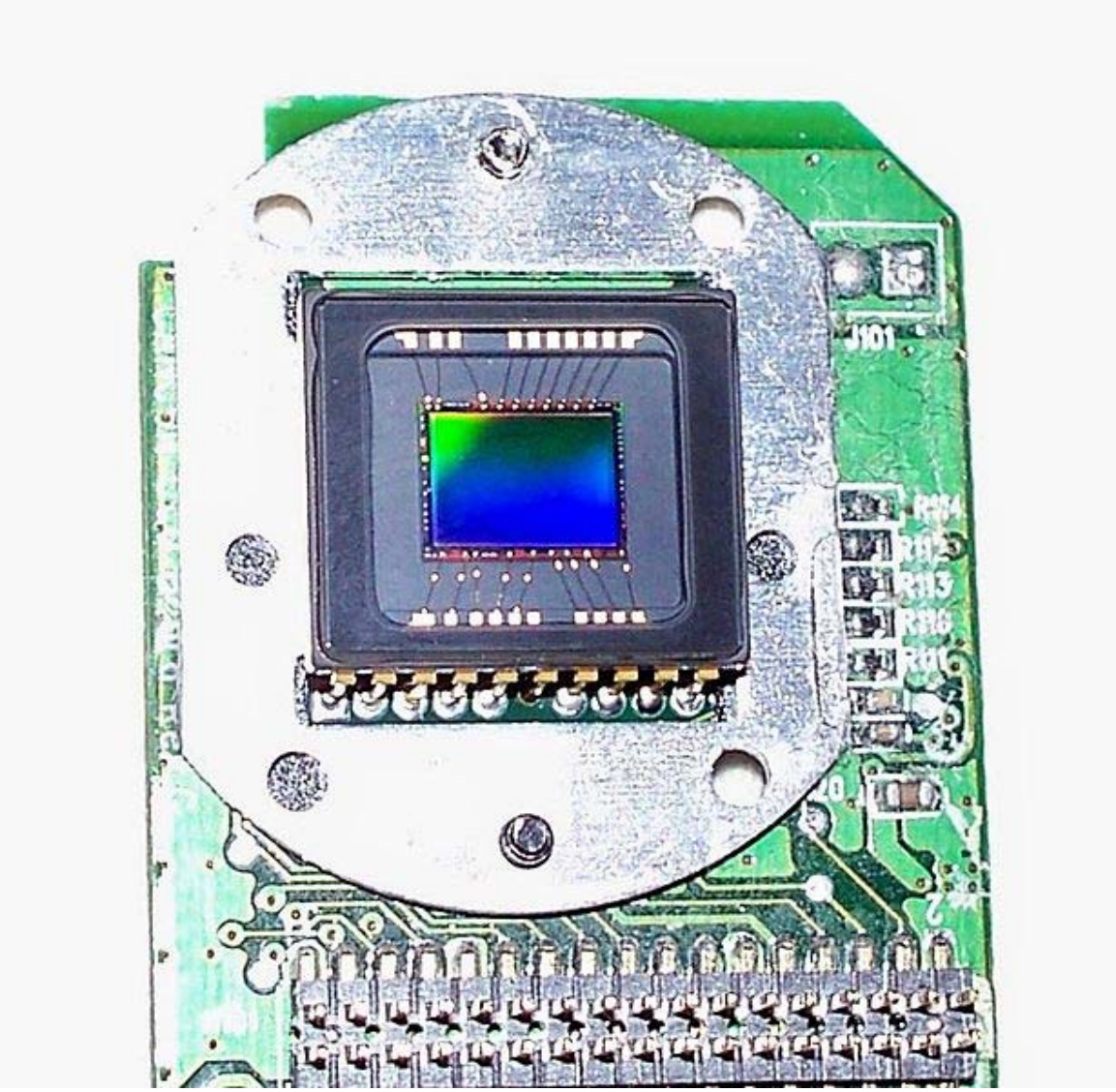
The interline architecture extends this concept one step further and masks every other column of the image sensor for storage. In this device, only one pixel shift has to occur to transfer from image area to storage area; thus, shutter times can be less than a microsecond and smear is essentially eliminated. The advantage is not free, however, as the imaging area is now covered by opaque strips dropping the fill factor to approximately 50 percent and the effective quantum efficiency by an equivalent amount. Modern designs have addressed this deleterious characteristic by adding microlenses on the surface of the device to direct light away from the opaque regions and on the active area. Microlenses can bring the fill factor back up to 90 percent or more depending on pixel size and the overall system's optical design.



CCD from a 2.1 megapixel Argus digital camera.

The choice of architecture comes down to one of utility. If the application cannot tolerate an expensive, failure-prone, power-intensive mechanical shutter, an interline device is the right choice. Consumer snap-shot cameras have used interline devices. On the other hand, for those applications that require the best possible light collection and issues of money, power and time are less important, the full-frame device is the right choice. Astronomers tend to prefer full-frame devices. The frame-transfer falls in between and was a common choice before the fill-factor issue of interline devices was addressed. Today, frame-transfer is usually chosen when an interline architecture is not available, such as in a back-illuminated device.

CCDs containing grids of pixels are used in digital cameras, optical scanners, and video cameras as light-sensing devices. They commonly respond to 70 percent of the incident light (meaning a quantum efficiency of about 70 percent) making them far more efficient than photographic film, which captures only about 2 percent of the incident light.



CCD from a 2.1 megapixel Hewlett-Packard digital camera.

Most common types of CCDs are sensitive to near-infrared light, which allows infrared photography, night-vision devices, and zero lux (or near zero lux) video-recording/photography. For normal silicon-based detectors, the sensitivity is limited to 1.1 μm. One other consequence of their sensitivity to infrared is that infrared from remote controls often appears on CCD-based digital cameras or camcorders if they do not have infrared blockers.

Cooling reduces the array's dark current, improving the sensitivity of the CCD to low light intensities, even for ultraviolet and visible wavelengths. Professional observatories often cool their detectors with liquid nitrogen to reduce the dark current, and therefore the thermal noise, to negligible levels.

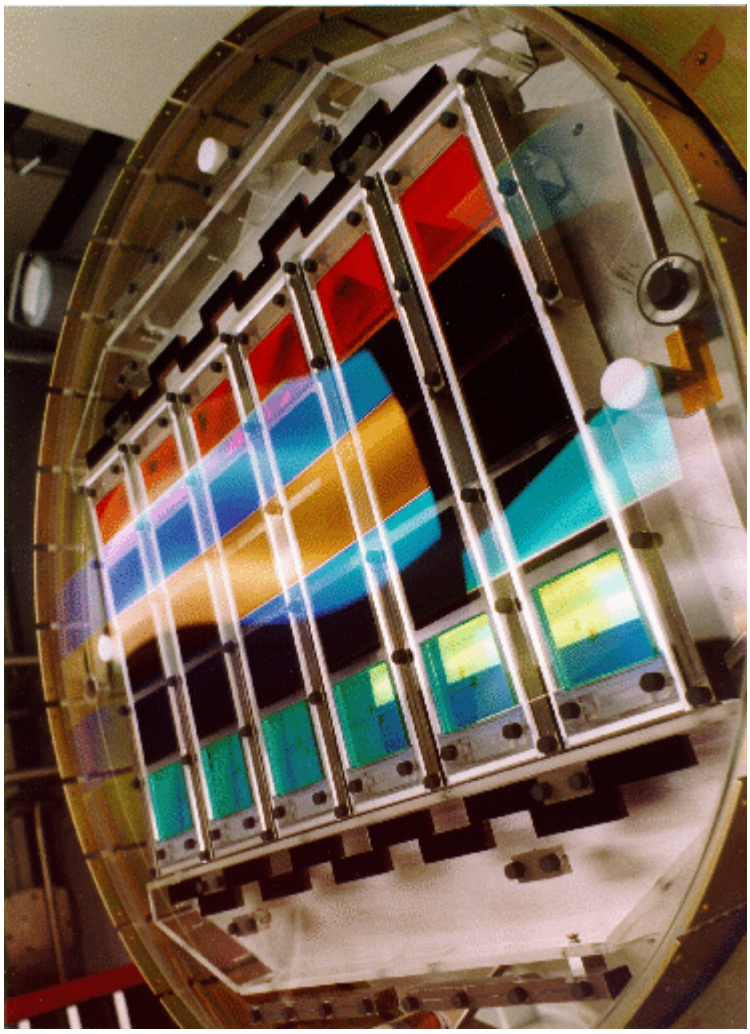
Use in astronomy

Due to the high quantum efficiencies of CCDs, linearity of their outputs (one count for one photon of light), ease of use compared to photographic plates, and a variety of other reasons, CCDs were very rapidly adopted by astronomers for nearly all UV-to-infrared applications.

Thermal noise and cosmic rays may alter the pixels in the CCD array. To counter such effects, astronomers take several exposures with the CCD shutter closed and opened. The average of images taken with the shutter closed is necessary to lower the random noise. Once developed, the *dark frame* average image is then subtracted from the open-shutter image to remove the dark current and other systematic defects (dead pixels, hot pixels, etc.) in the CCD.

The Hubble Space Telescope, in particular, has a highly developed series of steps ("data reduction pipeline") to convert the raw CCD data to useful images.

CCD cameras used in astrophotography often require sturdy mounts to cope with vibrations from wind and other sources, along with the tremendous weight of most imaging platforms. To take long exposures of galaxies and nebulae, many astronomers use a technique known as auto-guiding. Most autoguiders use a second CCD chip to monitor deviations during imaging. This chip can rapidly detect errors in tracking and command the mount motors to correct for them.

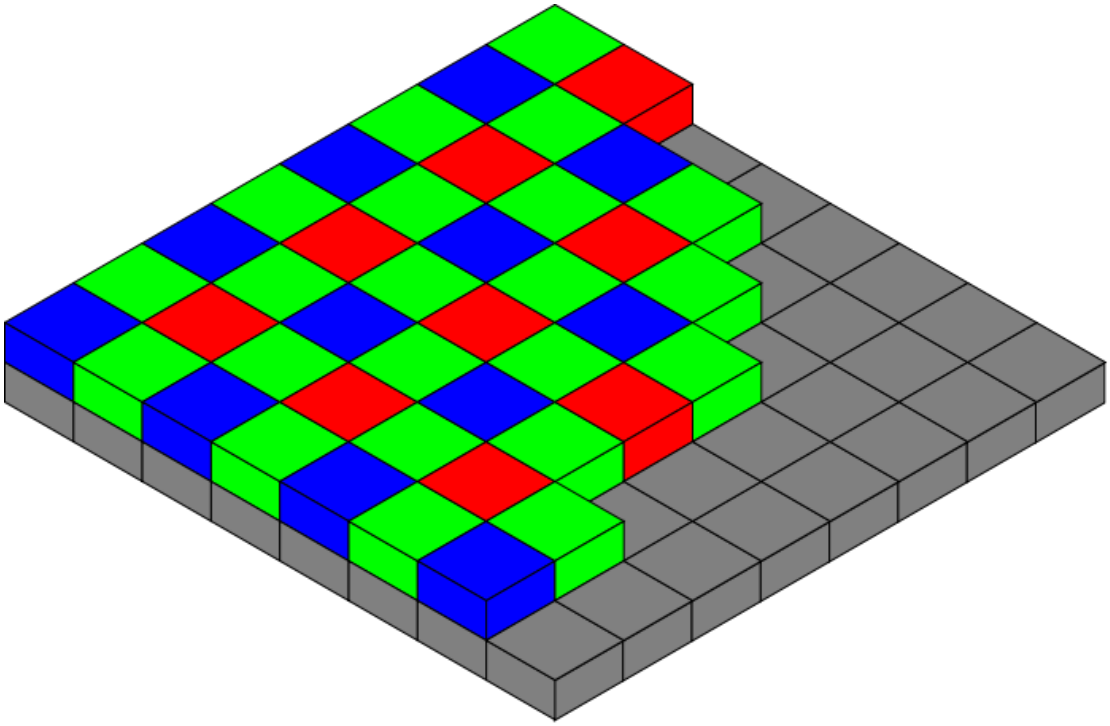


Array of 30 CCDs used on Sloan Digital Sky Survey telescope imaging camera, an example of "drift-scanning."

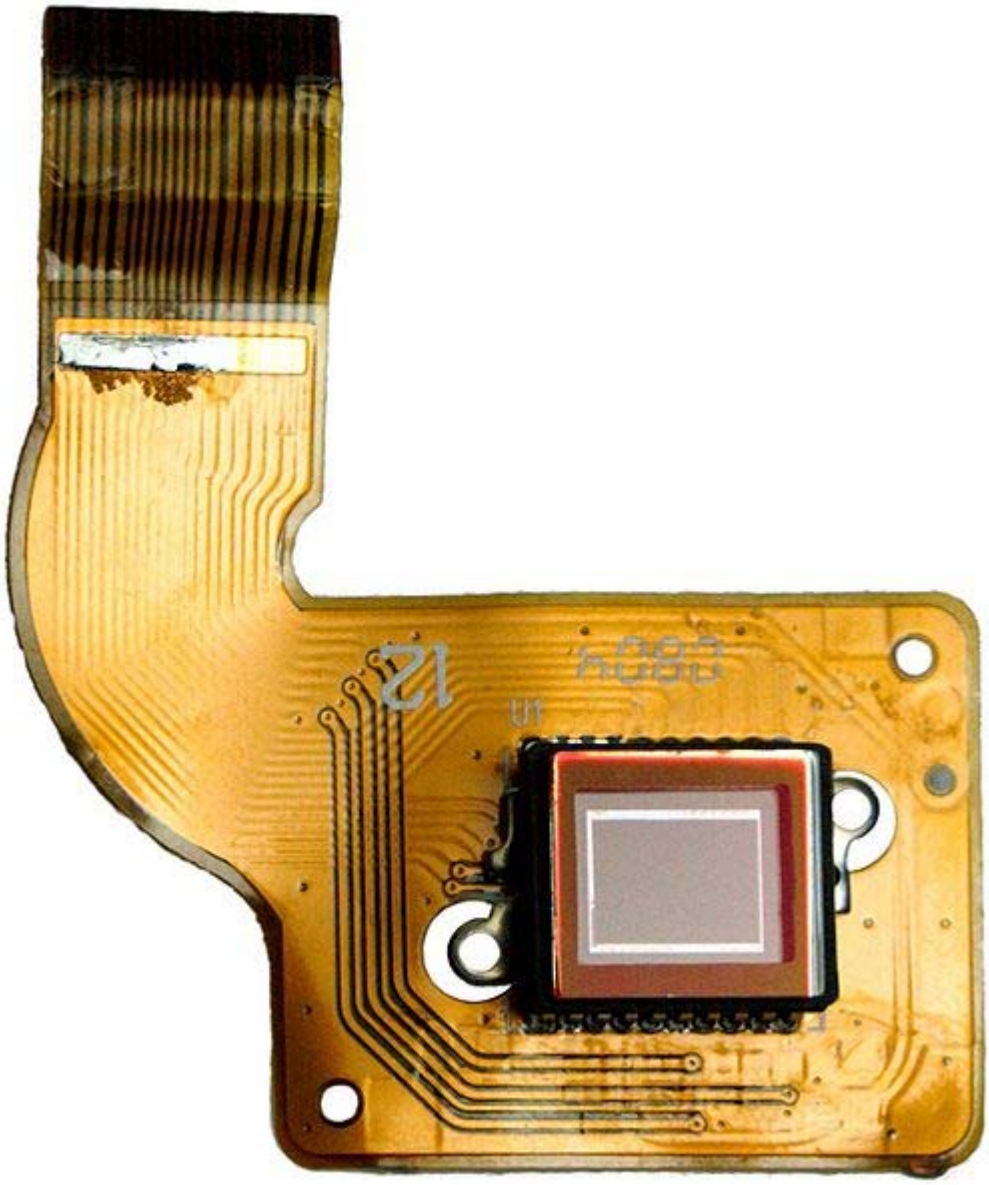
An interesting unusual astronomical application of CCDs, called *drift-scanning*, uses a CCD to make a fixed telescope behave like a tracking telescope and follow the motion of the sky. The charges in the CCD are transferred and read in a direction parallel to the motion of the sky, and at the same speed. In this way, the telescope can image a larger region of the sky than its normal field of view. The Sloan Digital Sky Survey is the most famous example of this, using the technique to produce the largest uniform survey of the sky yet accomplished.

In addition to astronomy, CCDs are also used in laboratory analytical instrumentation such as monochromators, spectrometers, and N-slit laser interferometers.

Color cameras



A Bayer filter on a CCD



CCD-Colorsensor

Digital color cameras generally use a Bayer mask over the CCD. Each square of four pixels has one filtered red, one blue, and two green (the human eye is more sensitive to green than either red or blue). The result of this is that luminance information is collected at every pixel, but the color resolution is lower than the luminance resolution.

Better color separation can be reached by three-CCD devices (3CCD) and a dichroic beam splitter prism, that splits the image into red, green and blue components. Each of the three CCDs is arranged to respond to a particular color. Most professional video camcorders, and some semi-professional camcorders, use this technique. Another advantage of 3CCD over a Bayer mask device is higher quantum efficiency (and therefore higher light sensitivity for a given aperture size). This is because in a 3CCD

device most of the light entering the aperture is captured by a sensor, while a Bayer mask absorbs a high proportion (about 2/3) of the light falling on each CCD pixel.

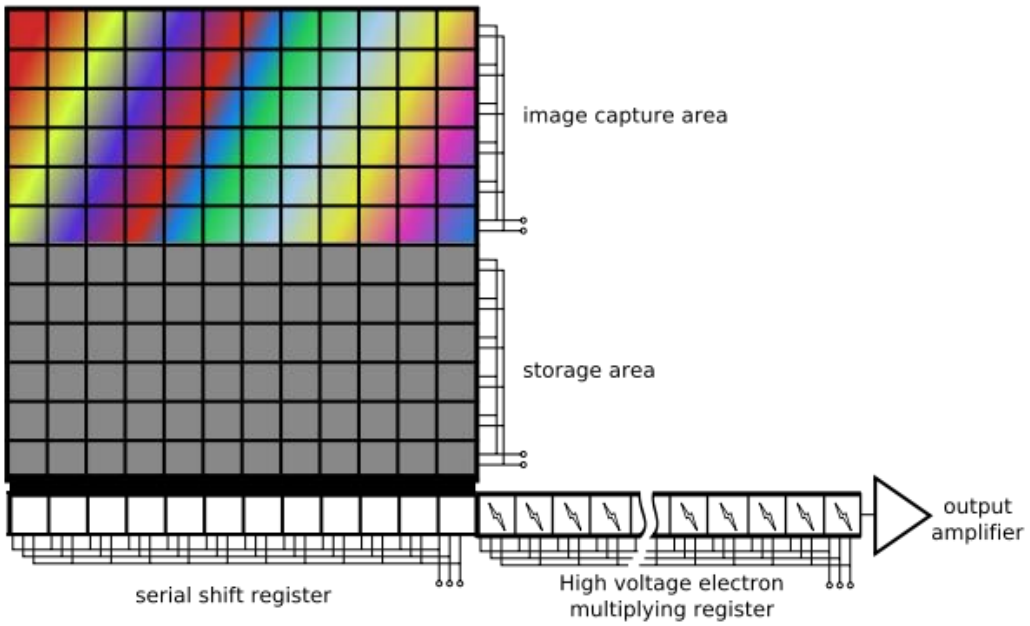
For still scenes, for instance in microscopy, the resolution of a Bayer mask device can be enhanced by Microscanning technology. During the process of color co-site sampling, several frames of the scene are produced. Between acquisitions, the sensor is moved in pixel dimensions, so that each point in the visual field is acquired consecutively by elements of the mask that are sensitive to the red, green and blue components of its color. Eventually every pixel in the image has been scanned at least once in each color and the resolution of the three channels become equivalent (the resolutions of red and blue channels are quadrupled while the green channel is doubled).

Sensor sizes

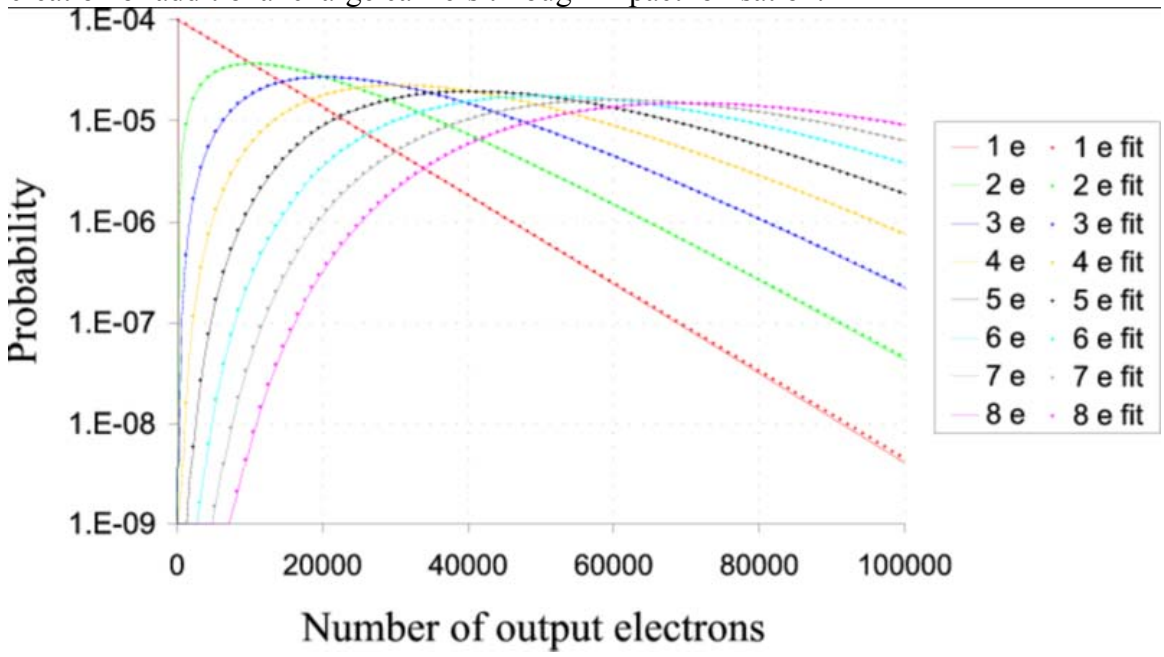
Sensors (CCD / CMOS) are often referred to with an inch fraction designation such as 1/1.8" or 2/3" called the optical format. This measurement actually originates back in the 1950s and the time of Vidicon tubes. Compact digital cameras and Digicams typically have much smaller sensors than a digital SLR and are thus less sensitive to light and inherently more prone to noise. Some examples of the CCDs found in modern cameras can be found in this table in a Digital Photography Review article

Type	Aspect Ratio	Width mm	Height mm	Diagonal mm	Area mm^2	Relative Area
1/6"	4:3	2.300	1.730	2.878	3.979	1.000
1/4"	4:3	3.200	2.400	4.000	7.680	1.930
1/3.6"	4:3	4.000	3.000	5.000	12.000	3.016
1/3.2"	4:3	4.536	3.416	5.678	15.495	3.894
1/3"	4:3	4.800	3.600	6.000	17.280	4.343
1/2.7"	4:3	5.270	3.960	6.592	20.869	5.245
1/2"	4:3	6.400	4.800	8.000	30.720	7.721
1/1.8"	4:3	7.176	5.319	8.932	38.169	9.593
2/3"	4:3	8.800	6.600	11.000	58.080	14.597
1"	4:3	12.800	9.600	16.000	122.880	30.882
4/3"	4:3	18.000	13.500	22.500	243.000	61.070
Other image sizes as a comparison						
APS-C	3:2	25.100	16.700	30.148	419.170	105.346
35mm	3:2	36.000	24.000	43.267	864.000	217.140
645	4:3	56.000	41.500	69.701	2324.000	584.066

Electron-multiplying CCD



Electrons are transferred serially through the gain stages making up the multiplication register of an EMCCD. The high voltages used in these serial transfers induce the creation of additional charge carriers through impact ionisation.



There is a dispersion (variation) in the number of electrons output by the multiplication register for a given (fixed) number of input electrons (shown in the legend on the right). The probability distribution for the number of output electrons is plotted logarithmically

on the vertical axis for a simulation of a multiplication register. Also shown are results from the empirical fit equation shown on this page.

An **electron-multiplying CCD** (EMCCD, also known as an L3Vision CCD, L3CCD or Impactron CCD) is a charge-coupled device in which a gain register is placed between the shift register and the output amplifier. The gain register is split up into a large number of stages. In each stage the electrons are multiplied by impact ionization in a similar way to an avalanche diode. The gain probability at every stage of the register is small ($P < 2\%$) but as the number of elements is large ($N > 500$), the overall gain can be very high ($g = (1 + P)^N$), with single input electrons giving many thousands of output electrons. Reading a signal from a CCD gives a noise background, typically a few electrons. In an EMCCD this noise is superimposed on many thousands of electrons rather than a single electron; the devices thus have negligible readout noise.

EMCCDs show a similar sensitivity to Intensified CCDs (ICCDs). However, as with ICCDs, the gain that is applied in the gain register is stochastic and the *exact* gain that has been applied to a pixel's charge is impossible to know. At high gains (> 30), this uncertainty has the same effect on the signal-to-noise ratio (SNR) as halving the quantum efficiency with respect to operation with a gain of unity. However, at very low light levels (where the quantum efficiency is most important) it can be assumed that a pixel either contains an electron - or not. This removes the noise associated with the stochastic multiplication at the cost of counting multiple electrons in the same pixel as a single electron. The dispersion in the gain is shown in the graph on the right. For multiplication registers with many elements and large gains it is well modelled by the equation:

$$P(n) = \frac{(n-m+1)^{m-1}}{(m-1)!\left(g-1+\frac{1}{m}\right)^m} \exp\left(-\frac{n-m+1}{g-1+\frac{1}{m}}\right) \quad \text{if } n \geq m$$

where P is the probability of getting n output electrons given m input electrons and a total mean multiplication register gain of g.

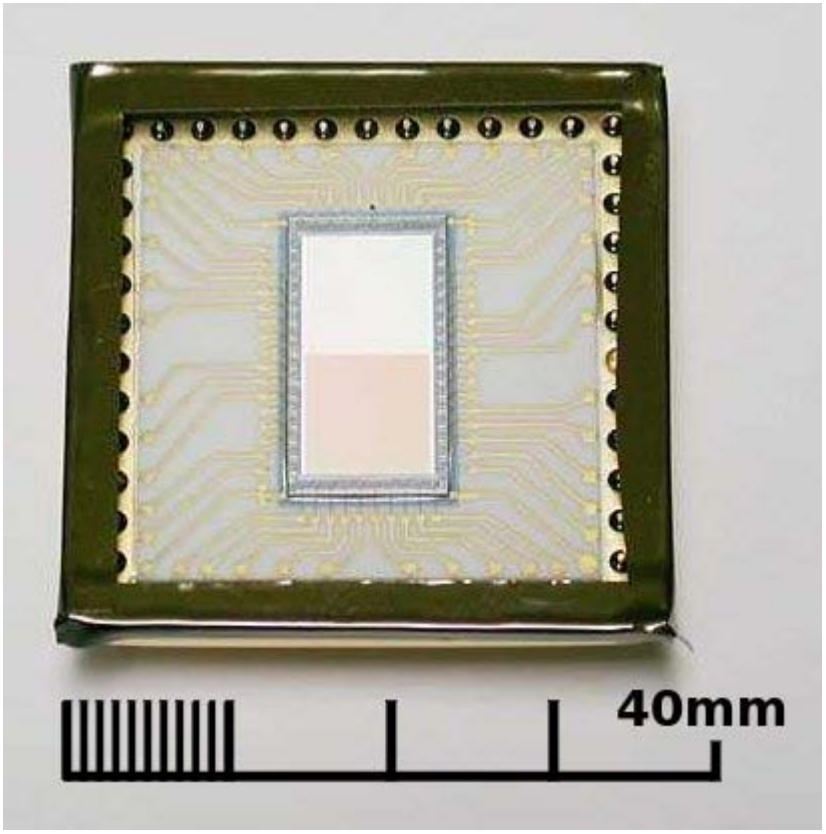
Because of the lower costs and the somewhat better resolution EMCCDs are capable of replacing ICCDs in many applications. ICCDs still have the advantage that they can be gated very fast and thus are useful in applications like range-gated imaging. EMCCD cameras indispensably need a cooling system to cool the chip down to temperatures around 170 K. This cooling system unfortunately adds additional costs to the EMCCD imaging system and often yields heavy condensation problems in the application.

The low-light capabilities of L3CCDs are starting to find use in astronomy. In particular their low noise at high readout speeds makes them very useful for lucky imaging of faint stars, and high speed photon counting photometry.

Commercial EMCCD cameras typically have clock-induced charge and darkcurrent (dependent on the extent of cooling) that leads to an effective readout noise ranging from 0.01 to 1 electrons per pixel read. Custom-built deep-cooled non-inverting mode

EMCCD cameras have provided effective readout noise lower than 0.1 electrons per pixel read for lucky imaging observations.

Frame transfer CCD



A frame transfer CCD for startracker applications.

Vertical smear.

A **frame transfer CCD** is a specialized CCD, often used in astronomy and some professional video cameras, designed for high exposure efficiency and correctness.

The normal functioning of a CCD, astronomical or otherwise, can be divided into two phases: exposure and readout. During the first phase, the CCD passively collects incoming photons, storing electrons in its cells. After the exposure time is passed, the cells are read out one line at a time.

During the readout phase, cells are shifted down the entire area of the CCD. While they are shifted, they continue to collect light. Thus, if the shifting is not fast enough, errors can result from light that falls on a cell holding charge during the transfer. These errors are referred to as "vertical smear" and cause a strong light source to create a vertical line above and below its exact location. In addition, the CCD cannot be used to collect light while it is being read out. Unfortunately, a faster shifting requires a faster readout, and a faster readout can introduce errors in the cell charge measurement, leading to a higher noise level.

A frame transfer CCD solves both problems: it has a hidden, not normally used, area containing as many cells as the area exposed to light. Typically, this area is covered by a reflective material such as aluminium. When the exposure time is up, the cells are transferred very rapidly to the hidden area. Here, safe from any incoming light, cells can be read out at any speed one deems necessary to correctly measure the cells' charge. At the same time, the exposed part of the CCD is collecting light again, so no delay occurs between successive exposures.

The disadvantage of such a CCD is the higher cost: the cell area is basically doubled, and more complex control electronics are needed.

Intensified charge-coupled device

An intensified charge-coupled device (ICCD) is a CCD that is optically connected to an image intensifier that is mounted in front of the CCD.

An image intensifier includes three functional elements: a photocathode, a micro-channel plate (MCP) and a phosphor screen. These three elements are mounted one close behind the other in the mentioned sequence. The photons which are coming from the light source fall onto the photocathode, thereby generating photoelectrons. The photoelectrons are accelerated towards the MCP by an electrical control voltage, applied between photocathode and MCP. The electrons are multiplied inside of the MCP and thereafter accelerated towards the phosphor screen. The phosphor screen finally converts the multiplied electrons back to photons which are guided to the CCD by a fiber optic or a lens.

An image intensifier inherently includes a shutter functionality: If the control voltage between the photocathode and the MCP is reversed, the emitted photoelectrons are not accelerated towards the MCP but return to the photocathode. Thus, no electrons are multiplied and emitted by the MCP, no electrons are going to the phosphor screen and no light is emitted from the image intensifier. In this case no light falls onto the CCD, which means that the shutter is closed. The process of reversing the control voltage at the photocathode is called gating and therefore ICCDs are also called gateable CCD cameras.

Beside of the extremely high sensitivity of ICCD cameras, which enable single photon detection, the gateability is one of the major advantages of the ICCD over the EMCCD cameras. The highest performing ICCD cameras enable shutter times as short as 200 picoseconds.

ICCD cameras are in general somewhat higher in price than EMCCD cameras because they need the expensive image intensifier. On the other hand EMCCD cameras need a cooling system to cool the EMCCD chip down to temperatures around 170 K. This cooling system adds additional costs to the EMCCD camera and often yields heavy condensation problems in the application.

ICCDs are used in night vision devices and in a large variety of scientific applications.

Chapter 9

Blinky (Novelty)

Various Blinkies.

Blinkies are small electronic devices that make very bright light (usually flashing) using LEDs and small batteries. They are often sold by vendors at night-time events that have fireworks displays such as Independence Day, Canada Day, or Guy Fawkes Night. They are also popular at raves, New Year's Eve parties and night time sporting events.

Other names

There is no industry standard or official name for Blinkies, but most common names use some combination of the terms flash, magnet, strobe, body, blink, light, and/or jewellery. Common examples are, Blinkys, Blinkies, Blinkees, Body Lights, Blinky Body Lights, Magnetic Flashers, or Flashing Jewelry.

Blinky has also recently become the registered trademark of Blinky Ltd. A Cheshire (UK) company that specializes in solar powered flashing LCD promotional merchandise, despite the previous industry wide use of the term for the devices described here.

Construction

Body

A typical blinky is a small metal cylinder that has threads on one end and a very small etched circuit board on the other. The threaded end is open to accept small button cell batteries, and another cylinder that screws on to hold them in place. The circuit board can

be round and inside the cylinder, or larger, shaped, and glued to the outside of the cylinder end. Common designs have a rubber gasket inside the front (between the batteries and circuit board). Tightening the base causes the gasket to flatten and allows the batteries to complete the circuit with the back of the circuit board.

Back

The most common designs use a set of strong magnets, one at the back of the Body Light, and another that can be removed. This allows the Body Light to be easily attached to clothes, or stuck on any magnetic metal such as buttons or belt buckles. Clips are often used to make earrings, a loop can make a pendant, or a ring can be welded to the back to make a finger ring. Double sided adhesive pads are sometimes used to stick the blinky directly to the body, most often in the navel.

Circuit board

The circuit board typically has anywhere from 2 to as many as 25 micro-LEDs. Current LED technology allows for every colour of the rainbow, even infra-red (for military/police), and ultra-violet (black light). Blue, white, violet, and ultra-violet LEDs often need 2 or more batteries because of their high voltage requirements. Because it is an etched circuit board the front can be constructed to flash in a variety of ways, especially where there are multiple LEDs in multiple colours. A clear plastic material such as silicone, acrylic or epoxy protects the fragile LEDs on the front (outside) of the board. Shaped boards have literally hundreds of variations combined with imprinting. Common shapes (besides the classic small round) are stars, hearts, flowers, flags, animals, holiday symbols (like Halloween jack-o-lanterns), and sports team logos.

Uses

Most often Blinkies are used for amusement at raves, parties and night time events. But they can have other uses as well such as:

- Blinkies imprinted with company logos at conventions.
- Safety lights for children during Halloween, or night time events.
- Fun and safety during camping trips.
- Emergency flashers for disabled automobiles or lost hikers (most blinkies have over a 1 mile visibility range at night).
- The term Blinky is often used for bicycle lights which flash. In some countries blinkies can be used as a primary light on a bicycle.
- Blinkies also can be attached to mobiles (cell phones). When the mobile turns on, make a call, receive a call and during calls the blinky will keep flashing.
- "Winky blinkies" can refer to stage and film props which display lighting effects, or "gags," during a dramatic production.

Blinky batteries

While a few blinkies are made to be used once (like glowsticks), most can be reused with a fresh set of batteries. Typical blinkies use two to three AG3/AG4, SG3/SG4, or two CR927 batteries. Where the same sized battery is available in alkaline or silver oxide chemistry types, silver oxide is preferred as they are longer lasting. Although such batteries cost about $3-$5 at watch and drug stores, they can be had for $15 per hundred from online stores. (However, watch stores usually install the battery in your watch, which can be very difficult, while blinkies are usually just 'twist, remove, replace, and twist'. With some blinkies, take care to retain the insulating plastic or tape around the outer ring of the set of two or three batteries.)

Similar devices

Although the term 'Blinky', 'Body Light' or 'Flashing Body Light' usually means the round or shaped devices listed above the term has also come to sometimes broadly define a large group of similar items. These include:

- Body Lights that don’t flash, but rather stay on brightly or slowly change colours.
- Any small novelty that makes light, like a light up whistle, or a light up keychain.
- Light up batons, mouth pieces, jewellery, or fibre optic wands.
- Electroluminescent wire and badges.
- Clothing with flashing LEDs, like belt buckles or shoes. In the case of the shoes, which are usually running shoes or attractive women's sandals, the flashing light is in the heel. Blinkies are also popular on small children's shoes, the eye-catching flashes making them highly attractive to young eyes and therefore making the product more saleable. The light starts flashing when the person wearing the shoes walks, runs, etc.

Chapter 10

Digital Light Processing

Digital Light Processing (DLP) is a trademark owned by Texas Instruments, representing a technology used in some TVs and video projectors. It was originally developed in 1987 by Dr. Larry Hornbeck of Texas Instruments.

DLP is used in DLP front projectors (small standalone projection units) and DLP rear projection television.

DLP, along with LCD and LCoS, are the current display technologies behind rear-projection television, having supplanted CRT rear projectors. These rear-projection technologies compete against LCD and plasma flat panel displays in the HDTV market.

The single-chip version of DLP and 3LCD are the two main technologies used in modern color digital projectors, with the two technologies being used in over 95% of the projectors sold in 2008.

DLP is also one of the leading technologies used in digital cinema projection.

In March 2008, TI announced the initial production of the DPP1500 chipset, which are micro projectors to be used in mobile devices. Availability for final products would show up in the market early 2009.

Digital micromirror device

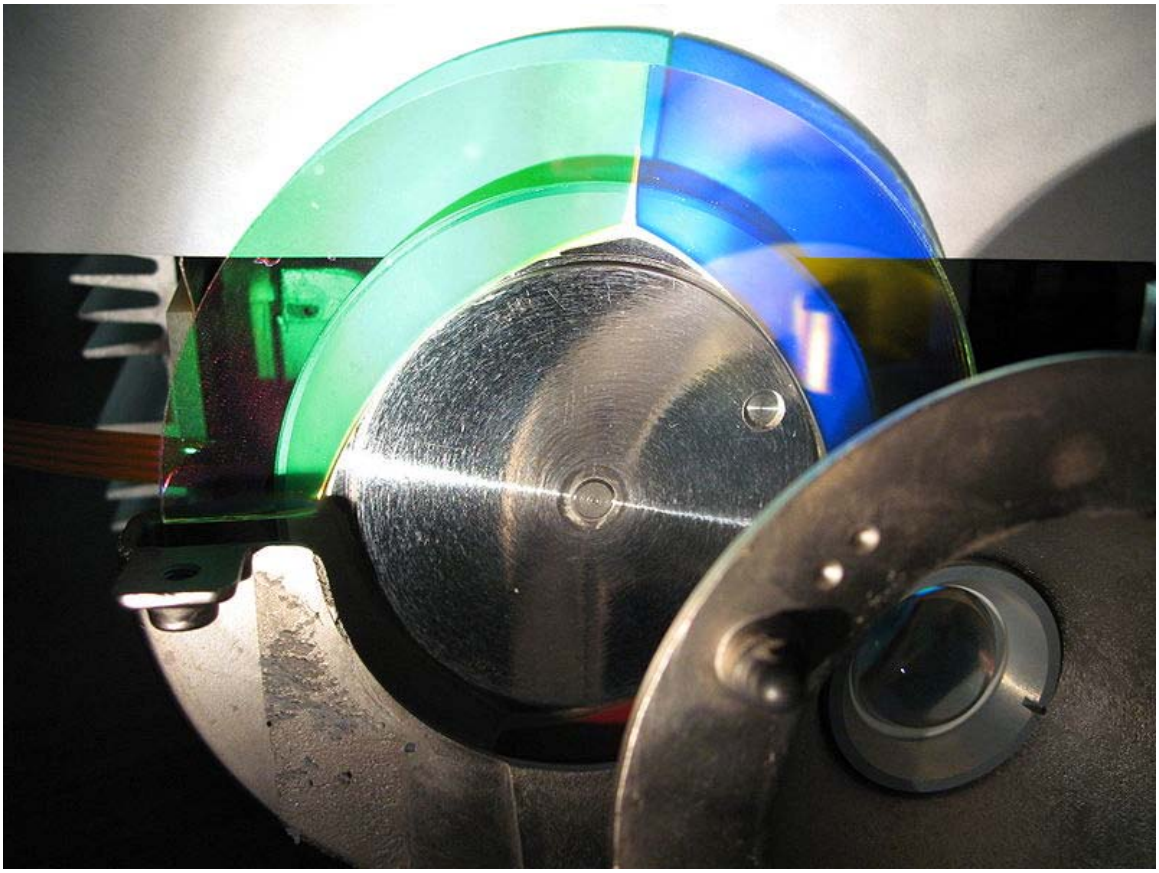
In DLP projectors, the image is created by microscopically small mirrors laid out in a matrix on a semiconductor chip, known as a Digital Micromirror Device (DMD). Each mirror represents one or more pixels in the projected image. The number of mirrors corresponds to the resolution of the projected image (often half as many mirrors as the advertised resolution due to wobulation). 800x600, 1024x768, 1280x720, and 1920x1080 (HDTV) matrices are some common DMD sizes. These mirrors can be repositioned rapidly to reflect light either through the lens or on to a heat sink (called a *light dump* in Barco terminology).

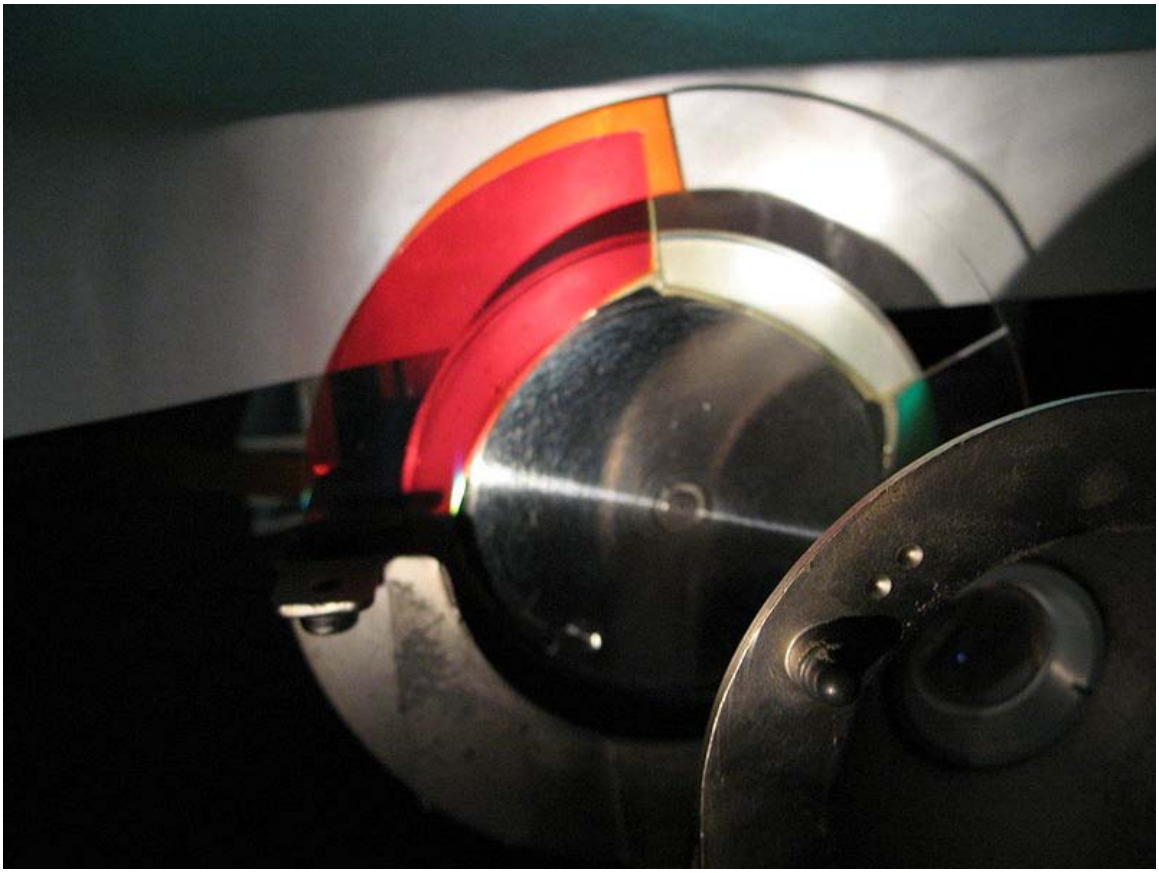
Rapidly toggling the mirror between these two orientations (essentially on and off) produces grayscales, controlled by the ratio of on-time to off-time.

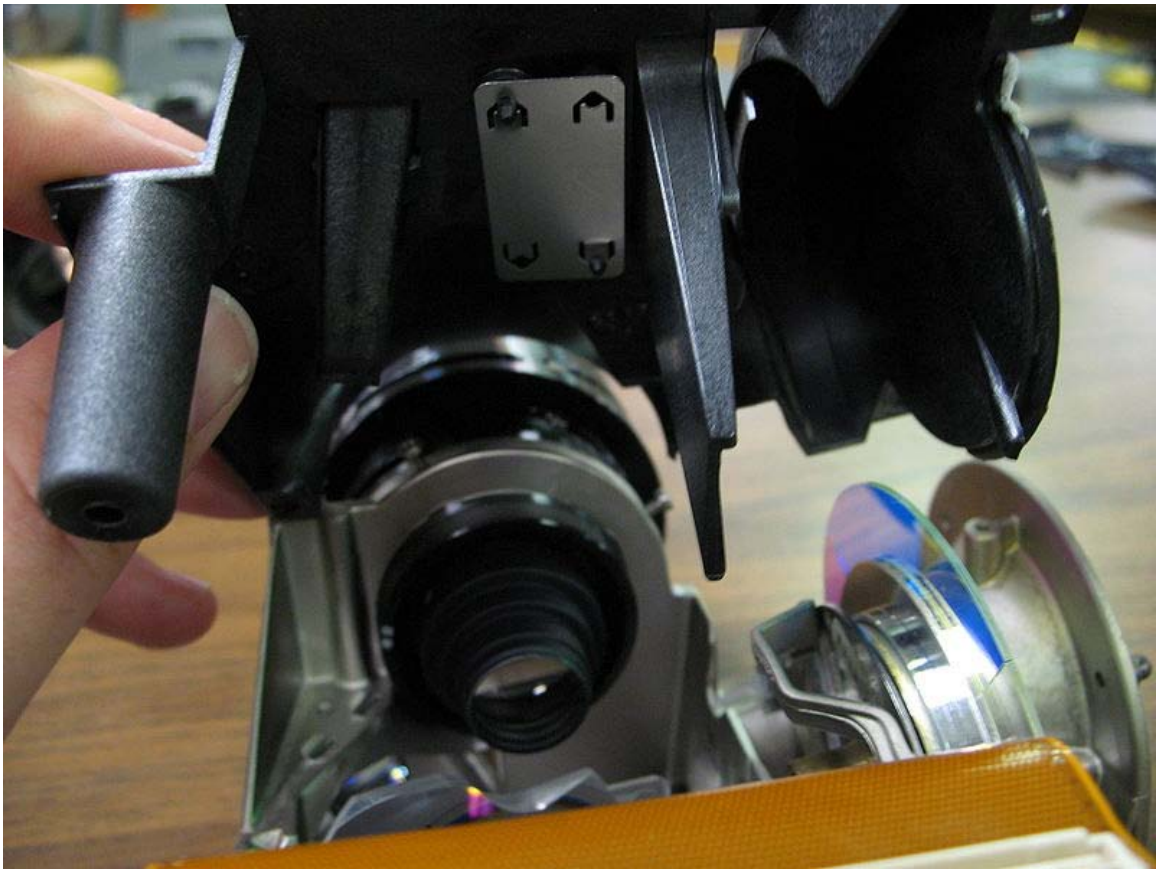
Color in DLP projection

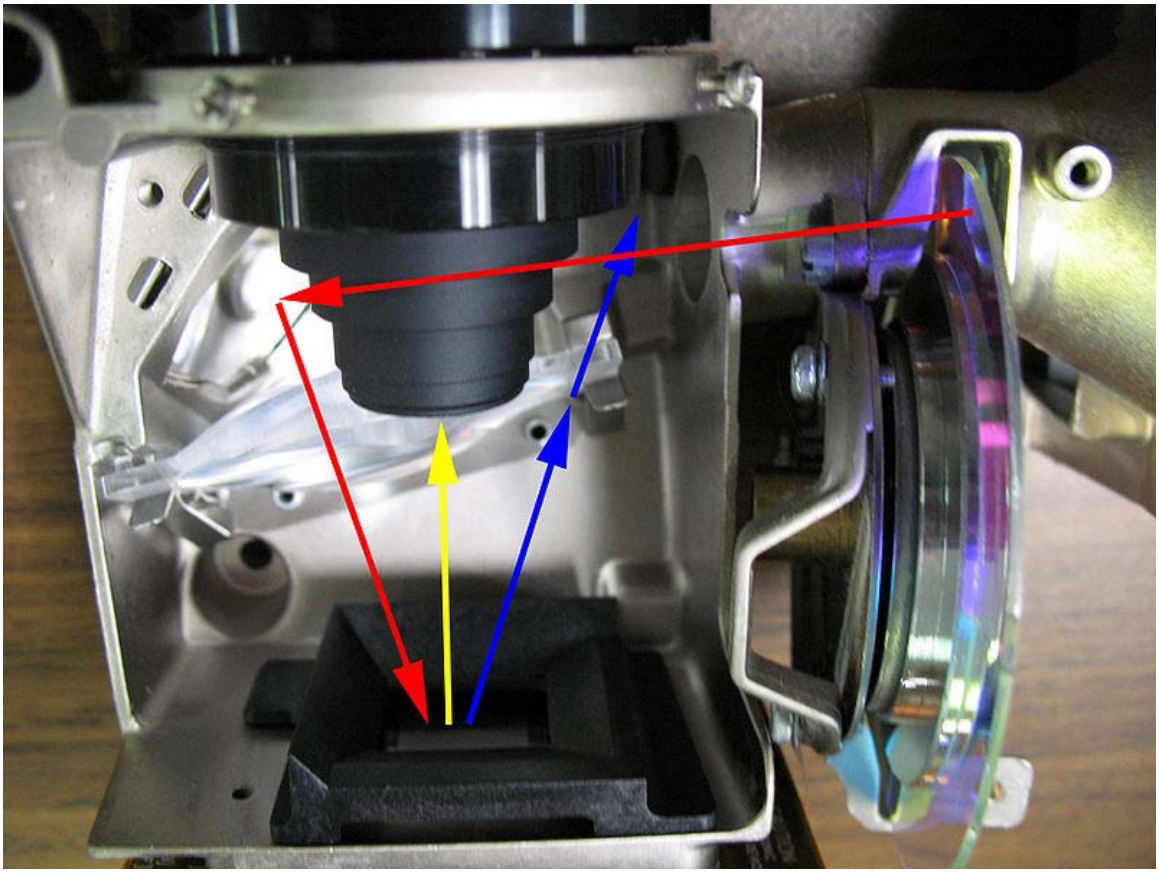
There are two primary methods by which DLP projection systems create a color image: those utilized by single-chip DLP projectors, and those used by three-chip projectors. A third method, sequential illumination by three colored light emitting diodes, is being developed, and is currently used in televisions manufactured by Samsung. Yet another method, color LASERs, is currently in use by Mitsubishi in their LASERVUE products.

Single-chip projectors









Interior view of a single-chip DLP projector, showing the light path. Light from the lamp enters a reverse-fisheye, passes through the spinning color wheel, crosses underneath the main lens, reflects off a front-surfaced mirror, and is spread onto the DMD (red arrows). From there, light either enters the lens (yellow) or is reflected off the top cover down into a light-sink (blue arrows) to absorb unneeded light. Top row shows overall components, closeups of 4-segment RGBW color wheel, and light-sink diffuser/reflection plate on top cover.

In a projector with a single DLP chip, colors are produced either by placing a color wheel between a white lamp and the DLP chip or by using individual light sources to produce the primary colors, LEDs or LASERs for example. The color wheel is divided into multiple sectors: the primary colors: red, green, and blue, and in many cases secondary colors including cyan, magenta, yellow and white. The use of the secondary colors is part of the new color performance system called BrilliantColor which processes the primary colors along with the secondary colors to create a broader spectrum of possible color combinations on the screen.

The DLP chip is synchronized with the rotating motion of the color wheel so that the green component is displayed on the DMD when the green section of the color wheel is in front of the lamp. The same is true for the red, blue and other sections. The colors are thus displayed sequentially at a sufficiently high rate that the observer sees a composite

"full color" image. In early models, this was one rotation per frame. Now, most systems operate at up to 10x the frame rate.

The color wheel "rainbow effect"

A single-chip projector alternates between colors and produces separate red, green, and blue images when displaying a moving image, or in this case, illuminating a moving hand.

DLP projectors utilizing a mechanical spinning color wheel may exhibit an anomaly known as the “rainbow effect.” This is best described as brief flashes of perceived red, blue, and green "shadows" observed most often when the projected content features high contrast areas of moving bright/white objects on a mostly dark/black background. The scrolling end credits of many movies are a common example, and also in animations where moving objects are surrounded by a thick black outline. Brief visible separation of the colours can also be apparent when the viewer moves their eyes quickly across the projected image. Some people perceive these rainbow artifacts frequently, while others may never see them at all.

This effect is caused by the way the eye follows a moving object on the projection. When an object on the screen moves, the eye will follow the object with a constant motion, but the projector will display each alternating color of the frame at the same location, for the

duration of the whole frame. So, while the eye is moving, it will see a frame of a specific color (red for example). Then, when the next color is displayed (green for example), although it gets displayed at the same location overlapping the previous color, the eye will have moved toward the object's next frame target. Thus, the eye will see that specific frame color slightly shifted. Then, the third color gets displayed (blue for example), and the eye will see that frame's color slightly shifted again. This effect is not perceived only for the moving object, but the whole picture.

The effect varies with the rotational speed of the color wheel and the frame refresh rate of the video signal. There is a maximum rotational speed limit for the wheel, typically 10,000 to 15,000 RPM. Video framerate is usually measured in frames per second and must be multiplied by 60 to find the wheel speed, whereas 60 frames/sec equals 3,600 frames/minute. If the color wheel spins 4 times per frame, it is rotating at a speed of 14,400 RPM. (Projector specifications often list the wheel speed at specific framerates as 2x, 3x, 4x, etc.) Increasing the video refresh rate to 85 frames per second does not necessarily further reduce the rainbow effect since this rate would increase the wheel speed to 20,400 RPM, potentially exceeding the safe limits of wheel rotation and requiring the projector to drop back to 3x speed, at 15,300 RPM.

Multi-color LED-based and LASER-based single-chip projectors are able to eliminate the spinning wheel and minimize the rainbow effect since the pulse rate of LEDs and LASERs are not limited by physical motion.

Three-chip projectors

A three-chip DLP projector uses a prism to split light from the lamp, and each primary color of light is then routed to its own DLP chip, then recombined and routed out through the lens. Three chip systems are found in higher-end home theater projectors, large venue projectors and DLP Cinema projection systems found in digital movie theaters.

According to DLP.com, the three-chip projectors used in movie theaters can produce 35 trillion colors, which many suggest is more than the human eye can detect. The human eye is suggested to be able to detect around 16 million colors, which is theoretically possible with the single chip solution. However, this high color precision does not mean that three-chip DLP projectors are capable of displaying the entire gamut of colors we can distinguish (this is fundamentally impossible with any system composing colors by adding three constant base colors). In contrast, it is the one-chip DLP projectors that have the advantage of allowing any number of primary colors in a sufficiently fast color filter wheel, and so the possibility of improved color gamuts is available.

Light source

The main light source used on DLP-based rear screen projection TVs is based on a replaceable high-pressure mercury-vapor metal halide arc lamp unit (containing a quartz arc tube, reflector, electrical connections, and sometimes a quartz/glass shield), while in

some newer DLP projectors high-power LEDs or lasers are used as a source of illumination.

Metal-halide lamps

For metal-halide lamps, during start-up, the lamp is ignited by a 5000 volt pulse from a current-regulating ballast to initiate an arc between two electrodes in the quartz tube. After warmup, the ballast's output voltage drops to approximately 60 volts while keeping the relative current high. As the lamp ages, the arc tube's electrodes wear out and light output declines somewhat while waste heating of the lamp increases. The mercury lamp's end of life is typically indicated via an LED on the unit or an onscreen text warning, necessitating replacement of the lamp unit.

Older projectors would simply give a warning that the lamp life had expired but would continue to operate. Newer projectors will not power up until the lamp is replaced and the lamp hours are reset. Most devices include a lamp hours reset function for when a new lamp is installed, but it is possible to reset a projector to continue to use an old lamp past its rated lifespan.

When a metal-halide lamp is operated past its rated lifespan, the efficiency declines significantly, the lightcast may become uneven, and the lamp starts to operate extremely hot, to the point that the power wires can melt off the lamp terminals. Eventually, the required startup voltage will also rise to the point where ignition can no longer occur. Secondary protections such as a temperature monitor may shut down the projector, but a thermally overstressed quartz arc tube can also crack and/or explode, releasing a cloud of hot mercury vapor inside and around the projector. However, practically all lamp housings contain heat-resistant barriers (in addition to those on the lamp unit itself) to prevent the red-hot quartz fragments from leaving the area.

LED-based DLPs

The first commercially-available LED-based DLP HDTV was the Samsung HL-S5679W in 2006, which also eliminated the use of color wheel. Besides long lifetime eliminating the need for lamp replacement and elimination of the color wheel, other advantages of LED illumination include instant-on operation and improved color, with increased color saturation and improved color gamut to over 140% of the NTSC color gamut. Samsung expanded the LED model line-up in 2007 with products available in 50", 56" and 61" screen sizes. For spring 2008, the third generation of Samsung LED DLP products are available in 61" (HL61A750) and 67" (HL67A750) screen sizes.

Ordinary LED technology does not produce the intensity and high lumen output characteristics required to replace arc lamps. The special patented LEDs used in all of the Samsung DLP TVs are PhlatLight LEDs, designed and manufactured by US based Luminus Devices. A single RGB PhlatLight LED chipset illuminates these projection TVs. The PhlatLight LEDs are also used in a new class of ultra-compact DLP front projector commonly referred to as a "pocket projector" and have been introduced in new

models from LG Electronics (HS101), Samsung electronics (SP-P400) and Casio (XJ-A series). Home Theater projectors will be the next category of DLP projectors that will use PhlatLight LED technology. At InfoComm, June 2008 Luminus and TI announced their collaboration on using their technology on home theater and business projectors and demonstrated a prototype PhlatLight LED based DLP home theater front projector. They also announced products will be available in the marketplace later in 2008 from Optoma and other companies to be named later in the year.

Luminus Devices PhlatLight LEDs have also been used by Christie Digital in their DLP-based MicroTiles display system. It is a modular system built from small (20 inch diagonal) rear projection cubes, which can be stacked and tiled together to form large display canvasses with very small seams. The scale and shape of the display can be any size, only constrained by practical limits.

LASER-based DLPs

The first commercially-available LASER-based DLP HDTV was the Mitsubishi L65-A90 LASERVUE in 2008, which also eliminated the use of a color wheel. Three separate color LASERs illuminate the digital micromirror device (DMD) in these projection TVs, producing a richer, more vibrant color palette than other methods.

Digital cinema

On February 2, 2000, Philippe Binant, technical manager of Digital Cinema Project at Gaumont in France, realized the first digital cinema projection in Europe with the DLP CINEMA technology developed by Texas Instruments.

DLP is the current market-share leader in professional digital movie projection, largely because of its high contrast ratio and available resolution as compared to other digital front-projection technologies. As of December 2008, there are over 6,000 DLP-based Digital Cinema Systems installed worldwide.

DLP projectors are also used in RealD Cinema and newer Imax theatres for 3-D films.

Manufacturers and market place

Texas Instruments remains the primary manufacturer of DLP technology, which is used by many licensees who market products based on T.I.'s chipsets. The Fraunhofer Institute of Dresden, Germany, also manufactures Digital Light Processors, termed Spatial Light Modulators, for use in specialized applications. For example, Micronic Laser Systems of Sweden utilizes Fraunhofer's SLMs to generate deep-ultraviolet imaging in its Sigma line of silicon mask lithography writers.

DLP technology has quickly gained market share in the front projection market and now holds roughly 50% of the worldwide share in front projection. Over 30 manufacturers use the DLP chipset to power their projectors.

Pros

- Smooth (at 1080p resolution), jitter-free images.
- Perfect geometry and excellent grayscale linearity achievable.
- Usually great ANSI contrast.
- No possibility of screen burn-in.
- Less "screen-door effect" than with LCD projectors.
- DLP rear projection TVs generally have a smaller form factor than comparable CRT projectors.
- DLP rear projection TVs are considerably cheaper than LCD or plasma flat-panel displays and can still offer 1080p resolution.
- The use of a replaceable light source means a potentially longer life than CRTs and plasma displays (this may also be a con as listed below).
- The light source is more-easily replaceable than the backlights used with LCDs, and on DLPs is often user-replaceable.
- New LED and LASER DLP TVs and projectors eliminate the need for lamp replacement.
- Using two projectors, one can project full color stereoscopic images using polarized process (because beams can be polarized).
- Lighter weight than LCD and plasma televisions.
- Unlike their LCD and plasma counterparts, DLP screens do not rely on fluids as their projection medium and are therefore not limited in size by their inherent mirror mechanisms, making them ideal for increasingly larger high-definition theater and venue screens.
- DLP Projectors can process up to 7 separate colors, giving them strong color performance.
- DLP projectors do not suffer from “Color Decay” often seen with LCD projectors in which the image on the screen turns yellow after extended periods of usage.

Cons

- Some viewers are bothered by the "rainbow effect," explained above.
- Not as thin as LCD or plasma flat-panel displays (although approximately comparable in weight), although some models as of 2008 are becoming wall-mountable (while still being 10" to 14" thick)
- Replacement of the lamp / light bulb. The average life span of a TV light source averages 2000-5000 hours and the replacement cost for these range from $99 – $350, depending on the brand and model. After replacing the bulb a few times the cost can easily exceed the original purchase price of the television itself. Newer generations units use LEDs or LASERs which effectively eliminates this issue, although replacement LED chips could potentially be required over the extended lifespan of the television.

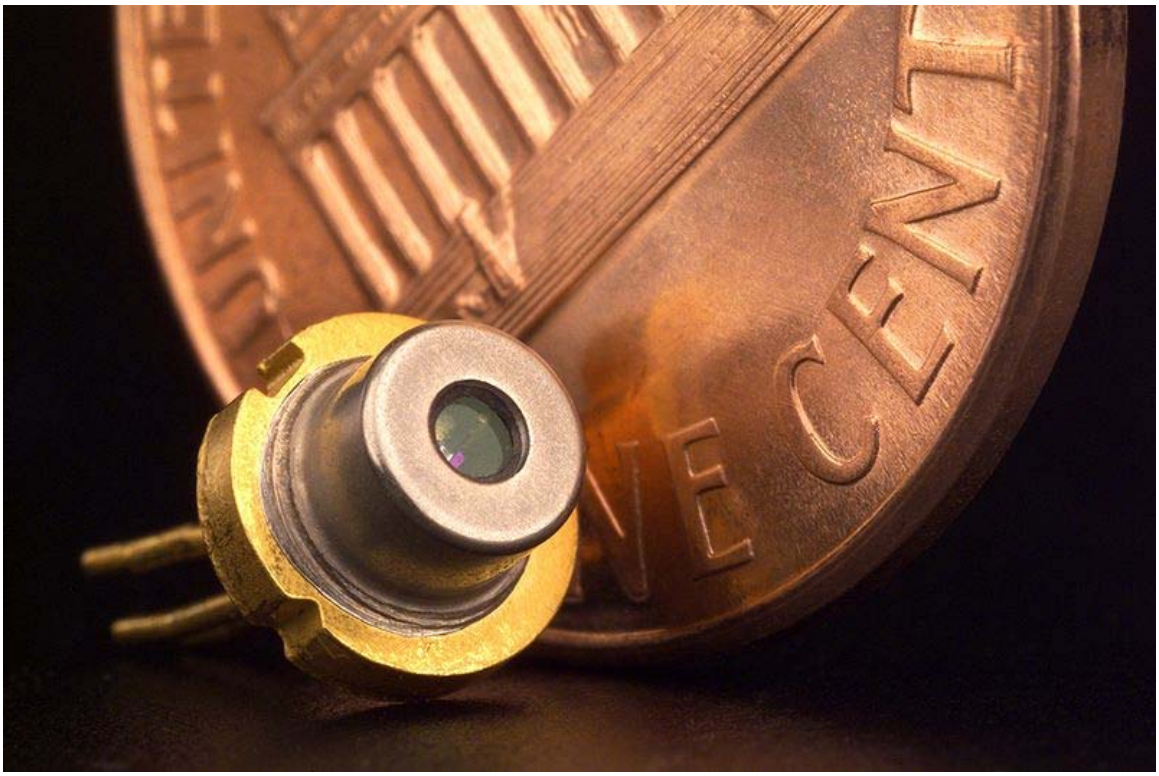
- Some devices may have fan noise.
- Dithering noise may be noticeable, especially in dark image areas. Newer (post ~2004) chip generations have less noise than older ones.
- Error-diffusion artifacts caused by averaging a shade over different pixels, since one pixel cannot render the shade exactly.
- Response time in video games may be affected by upscaling lag. While all HDTVs have some lag when upscaling lower resolution input to their native resolution, DLPs are commonly reported to have longer delays. Newer consoles such as the Xbox 360 do not have this problem as long as they are connected with HD-capable cables.
- Reduced viewing angle as compared to direct-view technologies such as CRT, plasma, and LCD.

DLP, LCD, and LCoS rear projection TV

The most similar competing system to DLP is known as LCoS (liquid crystal on silicon), which creates images using a stationary mirror mounted on the surface of a chip, and uses a liquid crystal matrix (similar to a liquid crystal display) to control how much light is reflected. DLP-based television systems are also arguably considered to be smaller in depth than traditional projection television.

Chapter 11

Laser Diode



A packaged laser diode with penny for scale.

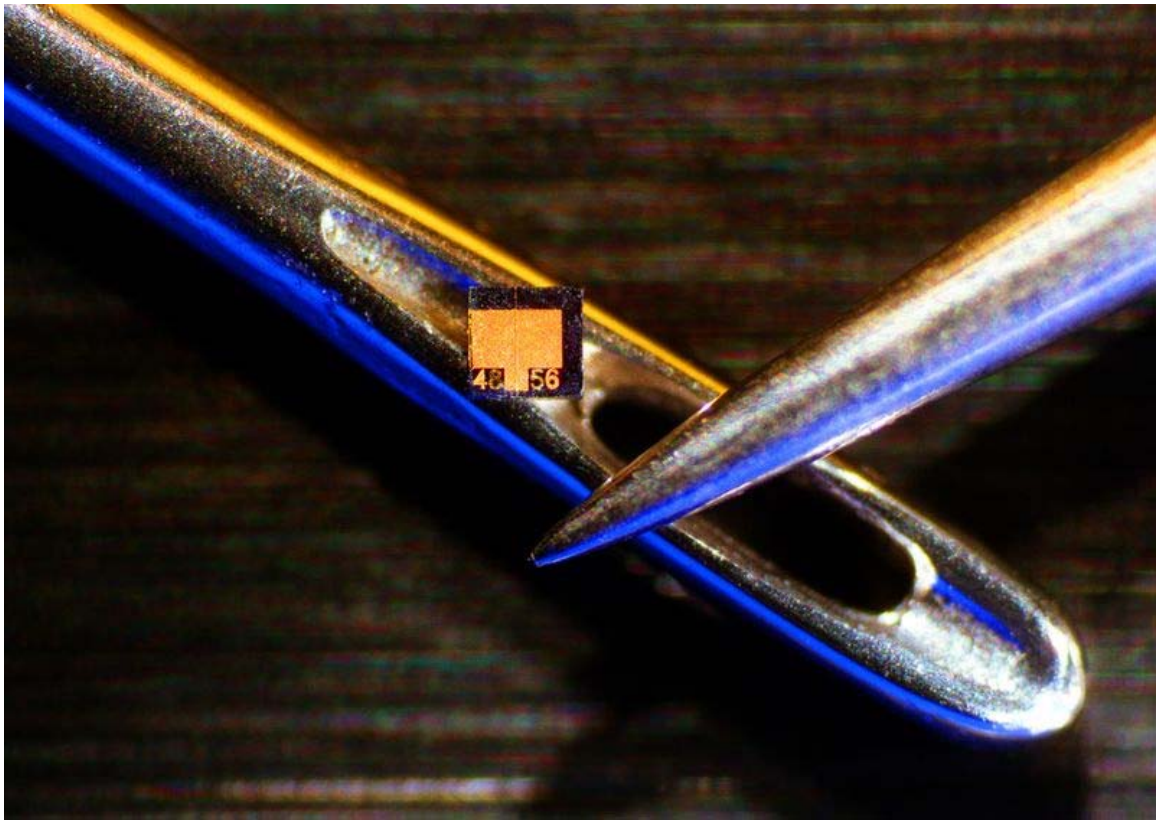
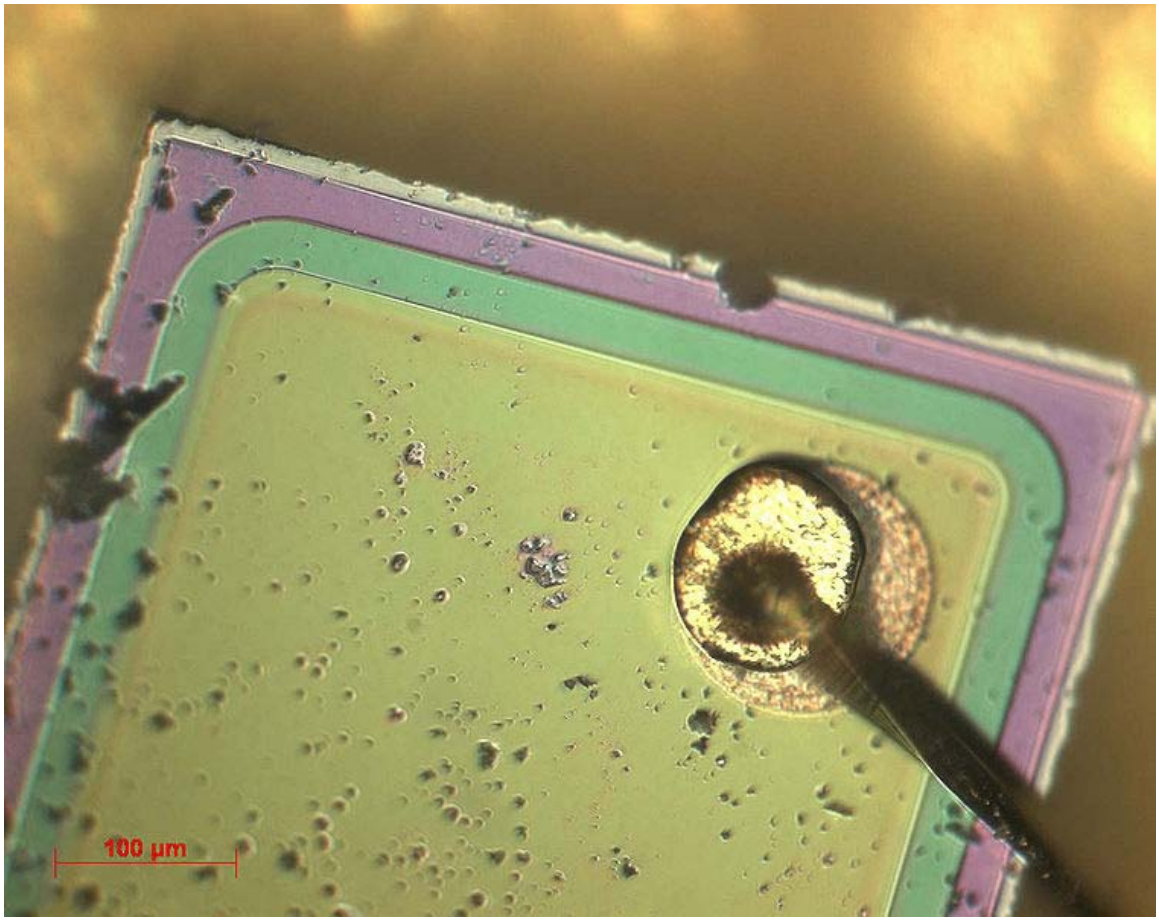


Image of the actual laser diode chip (shown on the eye of a needle for scale) contained within the package shown in the above image.



This is a visible light micrograph of a laser diode taken from a CD-ROM drive. Visible are the P and N layers distinguished by different colours. Also visible are scattered glass fragments from a broken collimating lens.

A **laser diode** is a laser where the active medium is a semiconductor similar to that found in a light-emitting diode. The most common type of laser diode is formed from a p-n junction and powered by injected electric current. The former devices are sometimes referred to as *injection laser diodes* to distinguish them from *optically pumped laser diodes*.

Theory of operation

A laser diode is formed by doping a very thin layer on the surface of a crystal wafer. The crystal is doped to produce an n-type region and a p-type region, one above the other, resulting in a *p*-*n* junction, or diode.

Laser diodes form a subset of the larger classification of semiconductor *p*-*n* junction diodes. Forward electrical bias across the laser diode causes the two species of charge carrier – holes and electrons – to be "injected" from opposite sides of the *p*-*n* junction into the depletion region. Holes are injected from the *p*-doped, and electrons from the *n*-

doped, semiconductor. (A depletion region, devoid of any charge carriers, forms as a result of the difference in electrical potential between *n*- and *p*-type semiconductors wherever they are in physical contact.) Due to the use of charge injection in powering most diode lasers, this class of lasers is sometimes termed "injection lasers," or "injection laser diode" (ILD). As diode lasers are semiconductor devices, they may also be classified as semiconductor lasers. Either designation distinguishes diode lasers from solid-state lasers.

Another method of powering some diode lasers is the use of optical pumping. Optically Pumped Semiconductor Lasers (OPSL) use a III-V semiconductor chip as the gain media, and another laser (often another diode laser) as the pump source. OPSL offer several advantages over ILDs, particularly in wavelength selection and lack of interference from internal electrode structures.

When an electron and a hole are present in the same region, they may recombine or "annihilate" with the result being spontaneous emission — i.e., the electron may re-occupy the energy state of the hole, emitting a photon with energy equal to the difference between the electron and hole states involved. (In a conventional semiconductor junction diode, the energy released from the recombination of electrons and holes is carried away as phonons, i.e., lattice vibrations, rather than as photons.) Spontaneous emission gives the laser diode below lasing threshold similar properties to an LED. Spontaneous emission is necessary to initiate laser oscillation, but it is one among several sources of inefficiency once the laser is oscillating.

The difference between the photon-emitting semiconductor laser and conventional phonon-emitting (non-light-emitting) semiconductor junction diodes lies in the use of a different type of semiconductor, one whose physical and atomic structure confers the possibility for photon emission. These photon-emitting semiconductors are the so-called "direct bandgap" semiconductors. The properties of silicon and germanium, which are single-element semiconductors, have bandgaps that do not align in the way needed to allow photon emission and are not considered "direct." Other materials, the so-called compound semiconductors, have virtually identical crystalline structures as silicon or germanium but use alternating arrangements of two different atomic species in a checkerboard-like pattern to break the symmetry. The transition between the materials in the alternating pattern creates the critical "direct bandgap" property. Gallium arsenide, indium phosphide, gallium antimonide, and gallium nitride are all examples of compound semiconductor materials that can be used to create junction diodes that emit light.

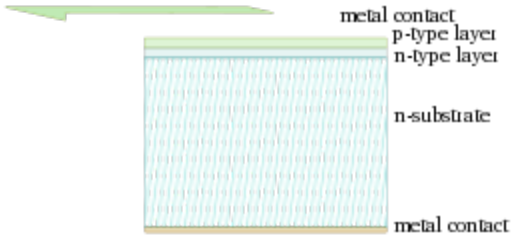


Diagram (not to scale) of a simple laser diode, such as shown above.

In the absence of stimulated emission (e.g., lasing) conditions, electrons and holes may coexist in proximity to one another, without recombining, for a certain time, termed the "upper-state lifetime" or "recombination time" (about a nanosecond for typical diode laser materials), before they recombine. Then a nearby photon with energy equal to the recombination energy can cause recombination by stimulated emission. This generates another photon of the same frequency, travelling in the same direction, with the same polarization and phase as the first photon. This means that stimulated emission causes gain in an optical wave (of the correct wavelength) in the injection region, and the gain increases as the number of electrons and holes injected across the junction increases. The spontaneous and stimulated emission processes are vastly more efficient in direct bandgap semiconductors than in indirect bandgap semiconductors; therefore silicon is not a common material for laser diodes.

As in other lasers, the gain region is surrounded with an optical cavity to form a laser. In the simplest form of laser diode, an optical waveguide is made on that crystal surface, such that the light is confined to a relatively narrow line. The two ends of the crystal are cleaved to form perfectly smooth, parallel edges, forming a Fabry–Pérot resonator. Photons emitted into a mode of the waveguide will travel along the waveguide and be reflected several times from each end face before they are emitted. As a light wave passes through the cavity, it is amplified by stimulated emission, but light is also lost due to absorption and by incomplete reflection from the end facets. Finally, if there is more amplification than loss, the diode begins to "lase".

Some important properties of laser diodes are determined by the geometry of the optical cavity. Generally, in the vertical direction, the light is contained in a very thin layer, and the structure supports only a single optical mode in the direction perpendicular to the layers. In the lateral direction, if the waveguide is wide compared to the wavelength of light, then the waveguide can support multiple lateral optical modes, and the laser is known as "multi-mode". These laterally multi-mode lasers are adequate in cases where one needs a very large amount of power, but not a small diffraction-limited beam; for example in printing, activating chemicals, or pumping other types of lasers.

In applications where a small focused beam is needed, the waveguide must be made narrow, on the order of the optical wavelength. This way, only a single lateral mode is

supported and one ends up with a diffraction-limited beam. Such single spatial mode devices are used for optical storage, laser pointers, and fiber optics. Note that these lasers may still support multiple longitudinal modes, and thus can lase at multiple wavelengths simultaneously.

The wavelength emitted is a function of the band-gap of the semiconductor and the modes of the optical cavity. In general, the maximum gain will occur for photons with energy slightly above the band-gap energy, and the modes nearest the gain peak will lase most strongly. If the diode is driven strongly enough, additional *side modes* may also lase. Some laser diodes, such as most visible lasers, operate at a single wavelength, but that wavelength is unstable and changes due to fluctuations in current or temperature.

Due to diffraction, the beam diverges (expands) rapidly after leaving the chip, typically at 30 degrees vertically by 10 degrees laterally. A lens must be used in order to form a collimated beam like that produced by a laser pointer. If a circular beam is required, cylindrical lenses and other optics are used. For single spatial mode lasers, using symmetrical lenses, the collimated beam ends up being elliptical in shape, due to the difference in the vertical and lateral divergences. This is easily observable with a red laser pointer.

The simple diode described above has been heavily modified in recent years to accommodate modern technology, resulting in a variety of types of laser diodes, as described below.

Types

The simple laser diode structure, described above, is extremely inefficient. Such devices require so much power that they can only achieve pulsed operation without damage. Although historically important and easy to explain, such devices are not practical.

Double heterostructure lasers

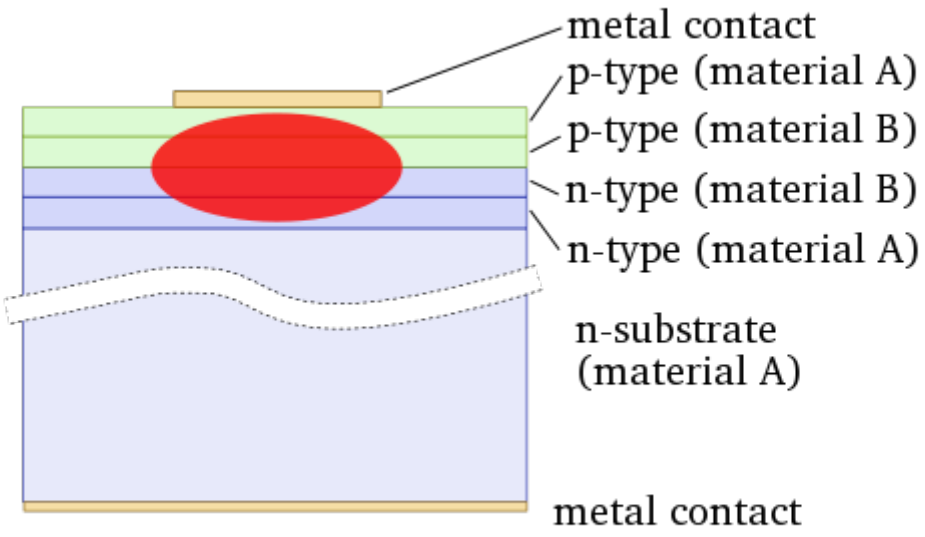


Diagram of front view of a double heterostructure laser diode (not to scale)

In these devices, a layer of low bandgap material is sandwiched between two high bandgap layers. One commonly-used pair of materials is gallium arsenide (GaAs) with aluminium gallium arsenide ($Al_xGa_{(1-x)}As$). Each of the junctions between different bandgap materials is called a *heterostructure*, hence the name "double heterostructure laser" or *DH* laser.

The advantage of a DH laser is that the region where free electrons and holes exist simultaneously—the active region—is confined to the thin middle layer. This means that many more of the electron-hole pairs can contribute to amplification—not so many are left out in the poorly amplifying periphery. In addition, light is reflected from the heterojunction; hence, the light is confined to the region where the amplification takes place.

Quantum well lasers

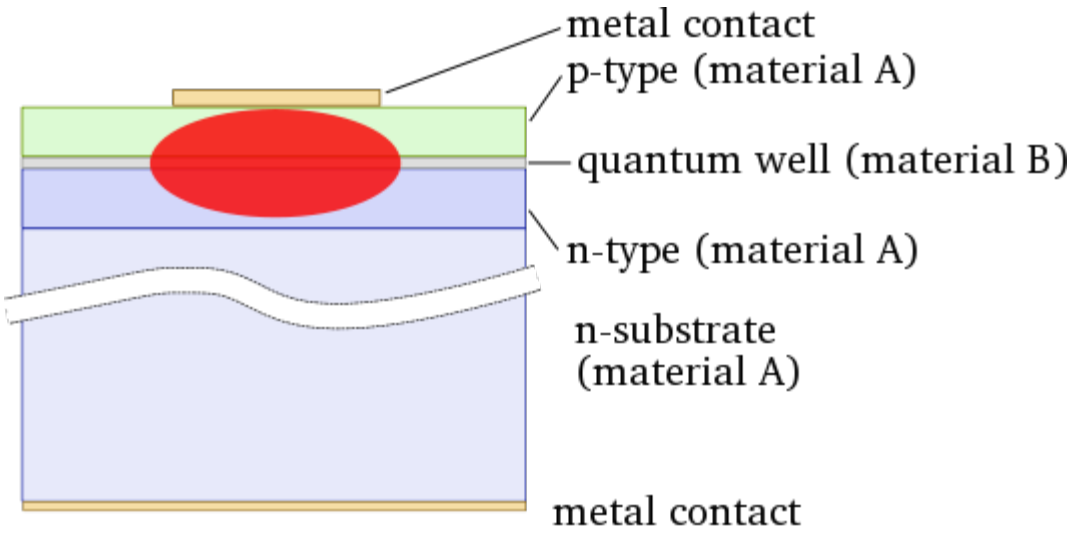


Diagram of front view of a simple quantum well laser diode (not to scale)

If the middle layer is made thin enough, it acts as a quantum well. This means that the vertical variation of the electron's wavefunction, and thus a component of its energy, is quantized. The efficiency of a quantum well laser is greater than that of a bulk laser because the density of states function of electrons in the quantum well system has an abrupt edge that concentrates electrons in energy states that contribute to laser action.

Lasers containing more than one quantum well layer are known as *multiple quantum well* lasers. Multiple quantum wells improve the overlap of the gain region with the optical waveguide mode.

Further improvements in the laser efficiency have also been demonstrated by reducing the quantum well layer to a quantum wire or to a "sea" of quantum dots.

Quantum cascade lasers

In a quantum cascade laser, the difference between quantum well energy levels is used for the laser transition instead of the bandgap. This enables laser action at relatively long wavelengths, which can be tuned simply by altering the thickness of the layer. They are heterojunction lasers.

Separate confinement heterostructure lasers

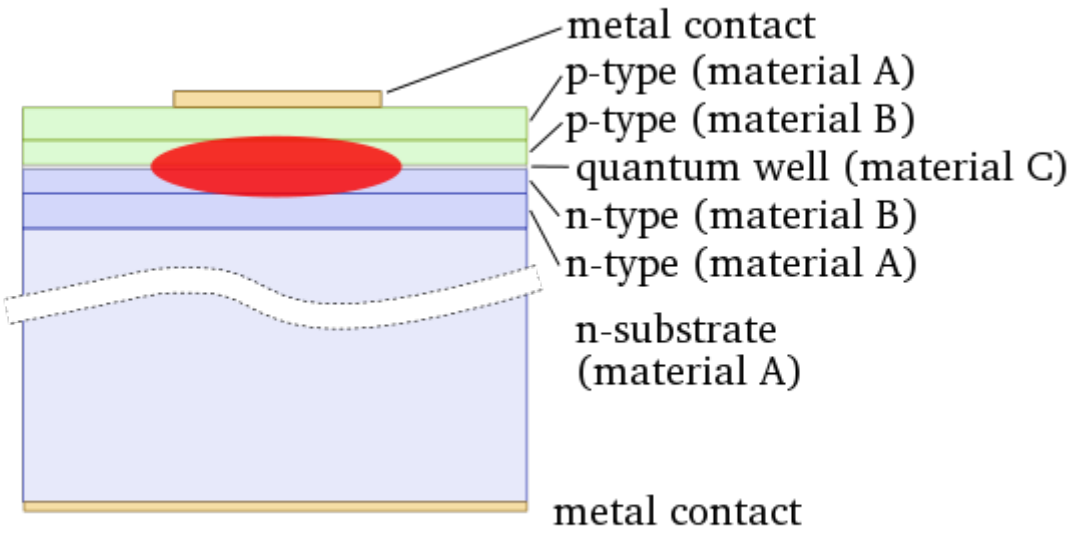


Diagram of front view of a separate confinement heterostructure quantum well laser diode

The problem with the simple quantum well diode described above is that the thin layer is simply too small to effectively confine the light. To compensate, another two layers are added on, outside the first three. These layers have a lower refractive index than the centre layers, and hence confine the light effectively. Such a design is called a separate confinement heterostructure (SCH) laser diode.

Almost all commercial laser diodes since the 1990s have been SCH quantum well diodes.

Distributed feedback lasers

Distributed feedback lasers (DFB) are the most common transmitter type in DWDM-systems. To stabilize the lasing wavelength, a diffraction grating is etched close to the p-n junction of the diode. This grating acts like an optical filter, causing a single wavelength to be fed back to the gain region and lase. Since the grating provides the feedback that is required for lasing, reflection from the facets is not required. Thus, at least one facet of a DFB is anti-reflection coated. The DFB laser has a stable wavelength that is set during manufacturing by the pitch of the grating, and can only be tuned slightly with temperature. DFB lasers are widely used in optical communication applications where a precise and stable wavelength is critical.

VCSELs

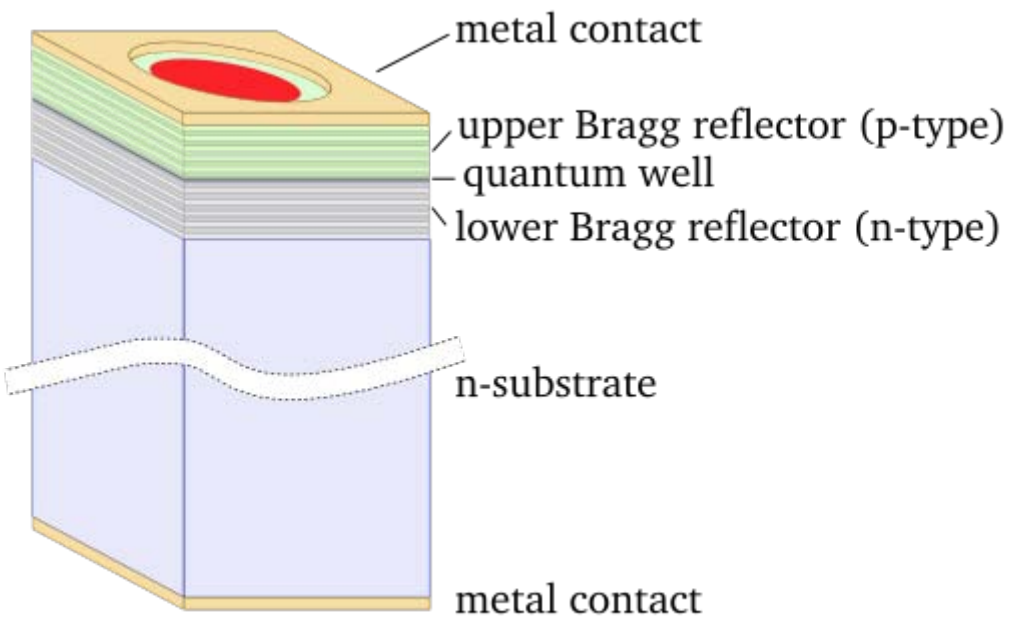


Diagram of a simple VCSEL structure

Vertical-cavity surface-emitting lasers (VCSELs) have the optical cavity axis along the direction of current flow rather than perpendicular to the current flow as in conventional laser diodes. The active region length is very short compared with the lateral dimensions so that the radiation emerges from the surface of the cavity rather than from its edge as shown in the figure. The reflectors at the ends of the cavity are dielectric mirrors made from alternating high and low refractive index quarter-wave thick multilayer.

Such dielectric mirrors provide a high degree of wavelength-selective reflectance at the required free surface wavelength λ if the thicknesses of alternating layers d_1 and d_2 with refractive indices n_1 and n_2 are such that $n_1d_1 + n_2d_2 = \lambda/2$ which then leads to the constructive interference of all partially reflected waves at the interfaces. But there is a disadvantage: because of the high mirror reflectivities, VCSELs have lower output powers when compared to edge-emitting lasers.

There are several advantages to producing VCSELs when compared with the production process of edge-emitting lasers. Edge-emitters cannot be tested until the end of the production process. If the edge-emitter does not work, whether due to bad contacts or poor material growth quality, the production time and the processing materials have been wasted. Additionally, because VCSELs emit the beam perpendicular to the active region of the laser as opposed to parallel as with an edge emitter, tens of thousands of VCSELs can be processed simultaneously on a three inch Gallium Arsenide wafer. Furthermore, even though the VCSEL production process is more labor and material intensive, the

yield can be controlled to a more predictable outcome. However, they normally show a lower power output level.

VECSELs

Vertical external-cavity surface-emitting lasers, or VECSELs, are similar to VCSELs. In VCSELs, the mirrors are typically grown epitaxially as part of the diode structure, or grown separately and bonded directly to the semiconductor containing the active region. VECSELs are distinguished by a construction in which one of the two mirrors is external to the diode structure. As a result, the cavity includes a free-space region. A typical distance from the diode to the external mirror would be 1 cm.

One of the most interesting features of any VECSEL is the small thickness of the semiconductor gain region in the direction of propagation, less than 100 nm. In contrast, a conventional in-plane semiconductor laser entails light propagation over distances of from 250 μm upward to 2 mm or longer. The significance of the short propagation distance is that it causes the effect of "antiguiding" nonlinearities in the diode laser gain region to be minimized. The result is a large-cross-section single-mode optical beam which is not attainable from in-plane ("edge-emitting") diode lasers.

Several workers demonstrated optically pumped VECSELs, and they continue to be developed for many applications including high power sources for use in industrial machining (cutting, punching, etc.) because of their unusually high power and efficiency when pumped by multi-mode diode laser bars.

Electrically pumped VECSELs have also been demonstrated. Applications for electrically pumped VECSELs include projection displays, served by frequency doubling of near-IR VECSEL emitters to produce blue and green light.

External-cavity diode lasers

External-cavity diode lasers are tunable lasers which use mainly double heterostructures diodes of the $Al_xGa_{(1-x)}As$ type. The first external-cavity diode lasers used intracavity etalons and simple tuning Littrow gratings. Other designs include gratings in grazing-incidence configuration and multiple-prism grating configurations.

Failure modes

Laser diodes have the same reliability and failure issues as light emitting diodes. In addition they are subject to *catastrophic optical damage* (COD) when operated at higher power.

Many of the advances in reliability of diode lasers in the last 20 years remain proprietary to their developers. The reliability of a laser diode can make or break a product line. Moreover, "reverse engineering" is not always able to reveal the differences between more-reliable and less-reliable diode laser products.

At the edge of a diode laser, where light is emitted, a mirror is traditionally formed by cleaving the semiconductor wafer to form a specularly reflecting plane. This approach is facilitated by the weakness of the [110] crystallographic plane in III-V semiconductor crystals (such as GaAs, InP, GaSb, etc.) compared to other planes. A scratch made at the edge of the wafer and a slight bending force causes a nearly atomically perfect mirror-like cleavage plane to form and propagate in a straight line across the wafer.

But it so happens that the atomic states at the cleavage plane are altered (compared to their bulk properties within the crystal) by the termination of the perfectly periodic lattice at that plane. Surface states at the cleaved plane, have energy levels within the (otherwise forbidden) bandgap of the semiconductor.

Essentially, as a result when light propagates through the cleavage plane and transits to free space from within the semiconductor crystal, a fraction of the light energy is absorbed by the surface states whence it is converted to heat by phonon-electron interactions. This heats the cleaved mirror. In addition the mirror may heat simply because the edge of the diode laser—which is electrically pumped—is in less-than-perfect contact with the mount that provides a path for heat removal. The heating of the mirror causes the bandgap of the semiconductor to shrink in the warmer areas. The bandgap shrinkage brings more electronic band-to-band transitions into alignment with the photon energy causing yet more absorption. This is thermal runaway, a form of positive feedback, and the result can be melting of the facet, known as *catastrophic optical damage*, or COD.

In the 1970s this problem, which is particularly nettlesome for GaAs-based lasers emitting between 1 μm and 0.630 μm wavelengths (less so for InP based lasers used for long-haul telecommunications which emit between 1.3 μm and 2 μm), was identified. Michael Ettenberg, a researcher and later Vice President at RCA Laboratories' David Sarnoff Research Center in Princeton, New Jersey, devised a solution. A thin layer of aluminum oxide was deposited on the facet. If the aluminum oxide thickness is chosen correctly it functions as an anti-reflective coating, reducing reflection at the surface. This alleviated the heating and COD at the facet.

Since then, various other refinements have been employed. One approach is to create a so-called non-absorbing mirror (NAM) such that the final 10 μm or so before the light emits from the cleaved facet are rendered non-absorbing at the wavelength of interest.

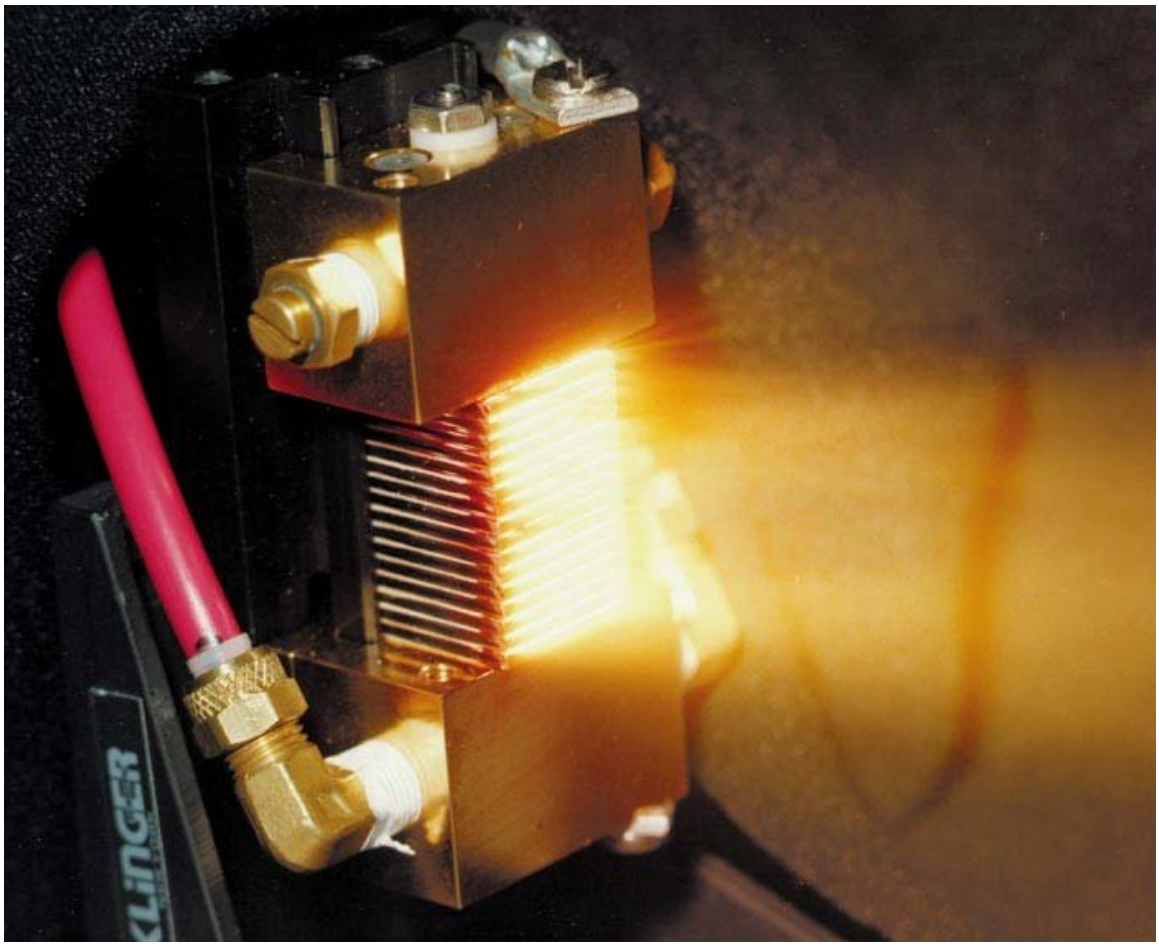
In the very early 1990s, SDL, Inc. began supplying high power diode lasers with good reliability characteristics. CEO Donald Scifres and CTO David Welch presented new reliability performance data at, e.g., SPIE Photonics West conferences of the era. The methods used by SDL to defeat COD were considered to be highly proprietary and have still not been disclosed publicly as of June, 2006.

In the mid-1990s IBM Research (Ruschlikon, Switzerland) announced that it had devised its so-called "E2 process" which conferred extraordinary resistance to COD in GaAs-based lasers. This process, too, has never been disclosed as of June, 2006.

Reliability of high-power diode laser pump bars (employed to pump solid state lasers) remains a difficult problem in a variety of applications, in spite of these proprietary advances. Indeed, the physics of diode laser failure is still being worked out and research on this subject remains active, if proprietary.

Extension of the lifetime of laser diodes is critical to their continued adaptation to a wide variety of applications.

Applications of laser diodes



Laser diodes can be arrayed to produce very high power (continuous wave or pulsed) outputs. Such arrays may be used to efficiently pump solid state lasers for inertial confinement fusion or high average power drilling or burning applications.

Laser diodes are numerically the most common type of laser, with 2004 sales of approximately 733 million diode lasers, as compared to 131,000 of other types of lasers.

Laser diodes find wide use in telecommunication as easily modulated and easily coupled light sources for fiber optics communication. They are used in various measuring

instruments, such as rangefinders. Another common use is in barcode readers. Visible lasers, typically red but later also green, are common as laser pointers. Both low and high-power diodes are used extensively in the printing industry both as light sources for scanning (input) of images and for very high-speed and high-resolution printing plate (output) manufacturing. Infrared and red laser diodes are common in CD players, CD-ROMs and DVD technology. Violet lasers are used in HD DVD and Blu-ray technology. Diode lasers have also found many applications in laser absorption spectrometry (LAS) for high-speed, low-cost assessment or monitoring of the concentration of various species in gas phase. High-power laser diodes are used in industrial applications such as heat treating, cladding, seam welding and for pumping other lasers, such as diode pumped solid state lasers.

Applications of laser diodes can be categorized in various ways. Most applications could be served by larger solid state lasers or optical parametric oscillators, but the low cost of mass-produced diode lasers makes them essential for mass-market applications. Diode lasers can be used in a great many fields; since light has many different properties (power, wavelength and spectral quality, beam quality, polarization, etc.) it is interesting to classify applications by these basic properties.

Many applications of diode lasers primarily make use of the "directed energy" property of an optical beam. In this category one might include the laser printers, bar-code readers, image scanning, illuminators, designators, optical data recording, combustion ignition, laser surgery, industrial sorting, industrial machining, and directed energy weaponry. Some of these applications are emerging while others are well-established.

Laser medicine: medicine and especially dentistry have found many new applications for diode lasers. The shrinking size of the units and their increasing user friendliness makes them very attractive to clinicians for minor soft tissue procedures. The 800 nm – 980 nm units have a high absorption rate for hemoglobin and thus make them ideal for soft tissue applications, where good hemostasis is necessary.

Applications which may make use of the coherence of diode-laser-generated light include interferometric distance measurement, holography, coherent communications, and coherent control of chemical reactions.

Applications which may make use of "narrow spectral" properties of diode lasers include range-finding, telecommunications, infra-red countermeasures, spectroscopic sensing, generation of radio-frequency or terahertz waves, atomic clock state preparation, quantum key cryptography, frequency doubling and conversion, water purification (in the UV), and photodynamic therapy (where a particular wavelength of light would cause a substance such as porphyrin to become chemically active as an anti-cancer agent only where the tissue is illuminated by light).

Applications where the desired quality of laser diodes is their ability to generate ultra-short pulses of light by the technique known as "mode-locking" include clock distribution for high-performance integrated circuits, high-peak-power sources for laser-induced

breakdown spectroscopy sensing, arbitrary waveform generation for radio-frequency waves, photonic sampling for analog-to-digital conversion, and optical code-division-multiple-access systems for secure communication.

Common wavelengths

- **375 nm** – excitation of Hoechst stain, Calcium Blue, and other fluorescent dyes in fluorescence microscopy
- **405 nm** – InGaN blue-violet laser, in Blu-ray Disc and HD DVD drives
- **445 nm** – InGaN Deep blue laser multimode diode recently introduced (2010) for use in mercury free high brightness data projectors
- **473 nm** – Bright blue laser pointers, still very expensive, output of DPSS systems
- **485 nm** – excitation of GFP and other fluorescent dyes
- **510 nm** - Green diodes recently (2010) developed by Nichia for laser projectors.
- **532 nm** – AlGaAs-pumped bright green laser pointers, frequency doubled 1064 nm Nd:YAG laser or (more commonly in laser pointers) Nd:YVO_4 IR lasers (SHG)
- **593 nm** – Yellow-Orange laser pointers, DPSS
- **635 nm** – AlGaInP better red laser pointers, same power subjectively 5 times as bright as 670 nm one
- **640 nm** – High brightness red DPSS laser pointers
- **657 nm** – AlGaInP DVD drives, laser pointers
- **670 nm** – AlGaInP cheap red laser pointers
- **760 nm** – AlGaInP gas sensing: O_2
- **785 nm** – GaAlAs Compact Disc drives
- **808 nm** – GaAlAs pumps in DPSS Nd:YAG lasers (e.g. in green laser pointers or as arrays in higher-powered lasers)
- **848 nm** – laser mice
- **980 nm** – InGaAs pump for optical amplifiers, for Yb:YAG DPSS lasers
- **1064 nm** – AlGaAs fiber-optic communication
- **1310 nm** – InGaAsP fiber-optic communication
- **1480 nm** – InGaAsP pump for optical amplifiers
- **1512 nm** – InGaAsP gas sensing: NH_3
- **1550 nm** – InGaAsP fiber-optic communication
- **1625 nm** – InGaAsP fiber-optic communication, service channel
- **1654 nm** – InGaAsP gas sensing: CH_4
- **1877 nm** – GaSbAs gas sensing: H_2O
- **2004 nm** – GaSbAs gas sensing: CO_2
- **2330 nm** – GaSbAs gas sensing: CO
- **2680 nm** – GaSbAs gas sensing: CO_2

History

The first to demonstrate coherent light emission from a semiconductor diode (the first *laser* diode), is widely acknowledged to have been Robert N. Hall and his team at the

General Electric research center in 1962. The first visible wavelength laser diode was demonstrated by Nick Holonyak, Jr. later in 1962.

Other teams at IBM, MIT Lincoln Laboratory, Texas Instruments, and RCA Laboratories were also involved in and received credit for their historic initial demonstrations of efficient light emission and lasing in semiconductor diodes in 1962 and thereafter. GaAs lasers were also produced in early 1963 in the Soviet Union by the team led by Nikolay Basov.

In the early 1960s liquid phase epitaxy (LPE) was invented by Herbert Nelson of RCA Laboratories. By layering the highest quality crystals of varying compositions, it enabled the demonstration of the highest quality heterojunction semiconductor laser materials for many years. LPE was adopted by all the leading laboratories, worldwide and used for many years. It was finally supplanted in the 1970s by molecular beam epitaxy and organometallic chemical vapor deposition.

Diode lasers of that era operated with threshold current densities of 1000 A/cm^2 at 77 K temperatures. Such performance enabled continuous-lasing to be demonstrated in the earliest days. However, when operated at room temperature, about 300 K, threshold current densities were two orders of magnitude greater, or 100,000 A/cm^2 in the best devices. The dominant challenge for the remainder of the 1960s was to obtain low threshold current density at 300 K and thereby to demonstrate continuous-wave lasing at room temperature from a diode laser.

The first diode lasers were homojunction diodes. That is, the material (and thus the bandgap) of the waveguide core layer and that of the surrounding clad layers, were identical. It was recognized that there was an opportunity, particularly afforded by the use of liquid phase epitaxy using aluminum gallium arsenide, to introduce heterojunctions. Heterostructures consist of layers of semiconductor crystal having varying bandgap and refractive index. Heterojunctions (formed from heterostructures) had been recognized by Herbert Kroemer, while working at RCA Laboratories in the mid-1950s, as having unique advantages for several types of electronic and optoelectronic devices including diode lasers. LPE afforded the technology of making heterojunction diode lasers.

The first heterojunction diode lasers were single-heterojunction lasers. These lasers utilized aluminum gallium arsenide *p*-type injectors situated over *n*-type gallium arsenide layers grown on the substrate by LPE. An admixture of aluminum replaced gallium in the semiconductor crystal and raised the bandgap of the *p*-type injector over that of the *n*-type layers beneath. It worked; the 300 K threshold currents went down by 10× to 10,000 amperes per square centimeter. Unfortunately, this was still not in the needed range and these single-heterostructure diode lasers did not function in continuous wave operation at room temperature.

The innovation that met the room temperature challenge was the double heterostructure laser. The trick was to quickly move the wafer in the LPE apparatus between different "melts" of aluminum gallium arsenide (*p*- and *n*-type) and a third melt of gallium

arsenide. It had to be done rapidly since the gallium arsenide core region needed to be significantly under 1 μm in thickness. This may have been the earliest true example of "nanotechnology." The first laser diode to achieve *continuous wave* operation was a double heterostructure demonstrated in 1970 essentially simultaneously by Zhores Alferov and collaborators (including Dmitri Z. Garbuzov) of the Soviet Union, and Morton Panish and Izuo Hayashi working in the United States. However, it is widely accepted that Zhores I. Alferov and team reached the milestone first.

For their accomplishment and that of their co-workers, Alferov and Kroemer shared the 2000 Nobel Prize in Physics.