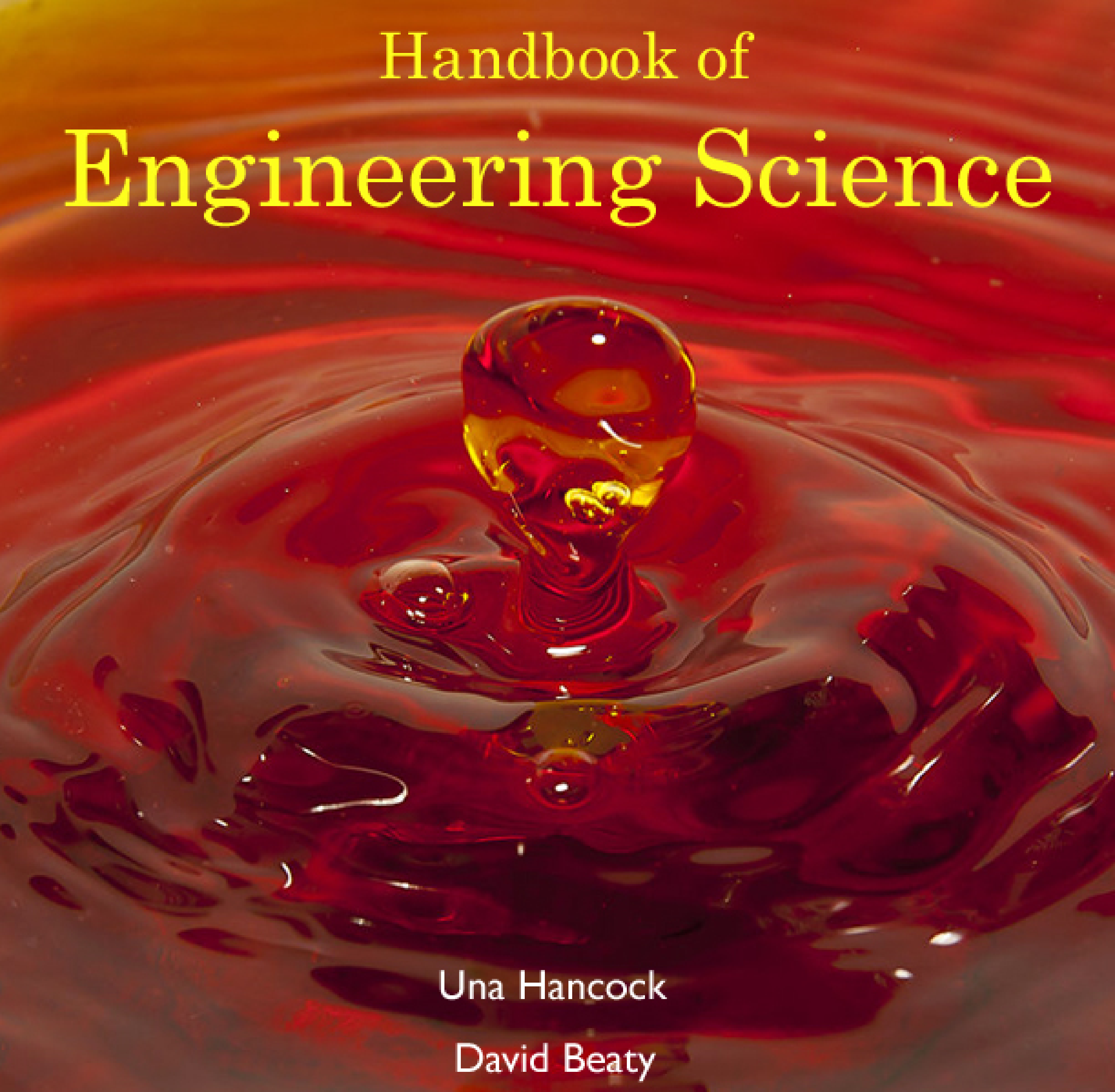


Handbook of

Engineering Science

Una Hancock

David Beaty

First Edition, 2012

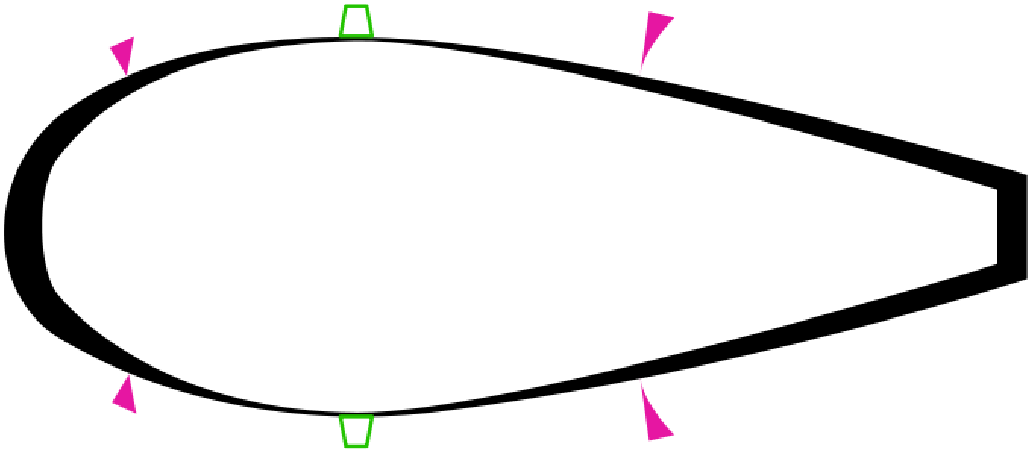
ISBN 978-81-323-0909-3

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

Table of Contents

Chapter 1

Fluid Dynamics



Typical aerodynamic teardrop shape, assuming a viscous medium passing from left to right, the diagram shows the pressure distribution as the thickness of the black line and shows the velocity in the boundary layer as the violet triangles. The green vortex generators prompt the transition to turbulent flow and prevent back-flow also called flow separation from the high pressure region in the back. The surface in front is as smooth as possible or even employs shark like skin, as any turbulence here will reduce the energy of the airflow. The truncation on the right, known as a Kammback, also prevents back flow from the high pressure region in the back across the spoilers to the convergent part.

In physics, **fluid dynamics** is a sub-discipline of fluid mechanics that deals with **fluid flow**—the natural science of fluids (liquids and gases) in motion. It has several subdisciplines itself, including aerodynamics (the study of air and other gases in motion) and **hydrodynamics** (the study of liquids in motion). Fluid dynamics has a wide range of applications, including calculating forces and moments on aircraft, determining the mass flow rate of petroleum through pipelines, predicting weather patterns, understanding nebulae in interstellar space and reportedly modeling fission weapon detonation. Some of its principles are even used in traffic engineering, where traffic is treated as a continuous fluid.

Fluid dynamics offers a systematic structure that underlies these practical disciplines, that embraces empirical and semi-empirical laws derived from flow measurement and used to solve practical problems. The solution to a fluid dynamics problem typically involves calculating various properties of the fluid, such as velocity, pressure, density, and temperature, as functions of space and time.

Historically, *hydrodynamics* meant something different than it does today. Before the twentieth century, hydrodynamics was synonymous with fluid dynamics. This is still reflected in names of some fluid dynamics topics, like magnetohydrodynamics and hydrodynamic stability—both also applicable in, as well as being applied to, gases.

Equations of fluid dynamics

The foundational axioms of fluid dynamics are the conservation laws, specifically, conservation of mass, conservation of linear momentum (also known as Newton's Second Law of Motion), and conservation of energy (also known as First Law of Thermodynamics). These are based on classical mechanics and are modified in quantum mechanics and general relativity. They are expressed using the Reynolds Transport Theorem.

In addition to the above, fluids are assumed to obey the *continuum assumption*. Fluids are composed of molecules that collide with one another and solid objects. However, the continuum assumption considers fluids to be continuous, rather than discrete. Consequently, properties such as density, pressure, temperature, and velocity are taken to be well-defined at infinitesimally small points, and are assumed to vary continuously from one point to another. The fact that the fluid is made up of discrete molecules is ignored.

For fluids which are sufficiently dense to be a continuum, do not contain ionized species, and have velocities small in relation to the speed of light, the momentum equations for Newtonian fluids are the Navier-Stokes equations, which is a non-linear set of differential equations that describes the flow of a fluid whose stress depends linearly on velocity gradients and pressure. The unsimplified equations do not have a general closed-form solution, so they are primarily of use in Computational Fluid Dynamics. The equations can be simplified in a number of ways, all of which make them easier to solve. Some of them allow appropriate fluid dynamics problems to be solved in closed form.

In addition to the mass, momentum, and energy conservation equations, a thermodynamical equation of state giving the pressure as a function of other thermodynamic variables for the fluid is required to completely specify the problem. An example of this would be the perfect gas equation of state:

$$p = \frac{\rho R_u T}{M}$$

where p is pressure, ρ is density, R_u is the gas constant, M is the molar mass and T is temperature.

Compressible vs incompressible flow

All fluids are compressible to some extent, that is changes in pressure or temperature will result in changes in density. However, in many situations the changes in pressure and temperature are sufficiently small that the changes in density are negligible. In this case the flow can be modeled as an incompressible flow. Otherwise the more general compressible flow equations must be used.

Mathematically, incompressibility is expressed by saying that the density ρ of a fluid parcel does not change as it moves in the flow field, i.e.,

$$\frac{\mathrm{D}\rho}{\mathrm{D}t} = 0 \,,$$

where D / Dt is the substantial derivative, which is the sum of local and convective derivatives. This additional constraint simplifies the governing equations, especially in the case when the fluid has a uniform density.

For flow of gases, to determine whether to use compressible or incompressible fluid dynamics, the Mach number of the flow is to be evaluated. As a rough guide, compressible effects can be ignored at Mach numbers below approximately 0.3. For liquids, whether the incompressible assumption is valid depends on the fluid properties (specifically the critical pressure and temperature of the fluid) and the flow conditions (how close to the critical pressure the actual flow pressure becomes). Acoustic problems always require allowing compressibility, since sound waves are compression waves involving changes in pressure and density of the medium through which they propagate.

Viscous vs inviscid flow

Viscous problems are those in which fluid friction has significant effects on the fluid motion.

The Reynolds number, which is a ratio between inertial and viscous forces, can be used to evaluate whether viscous or inviscid equations are appropriate to the problem.

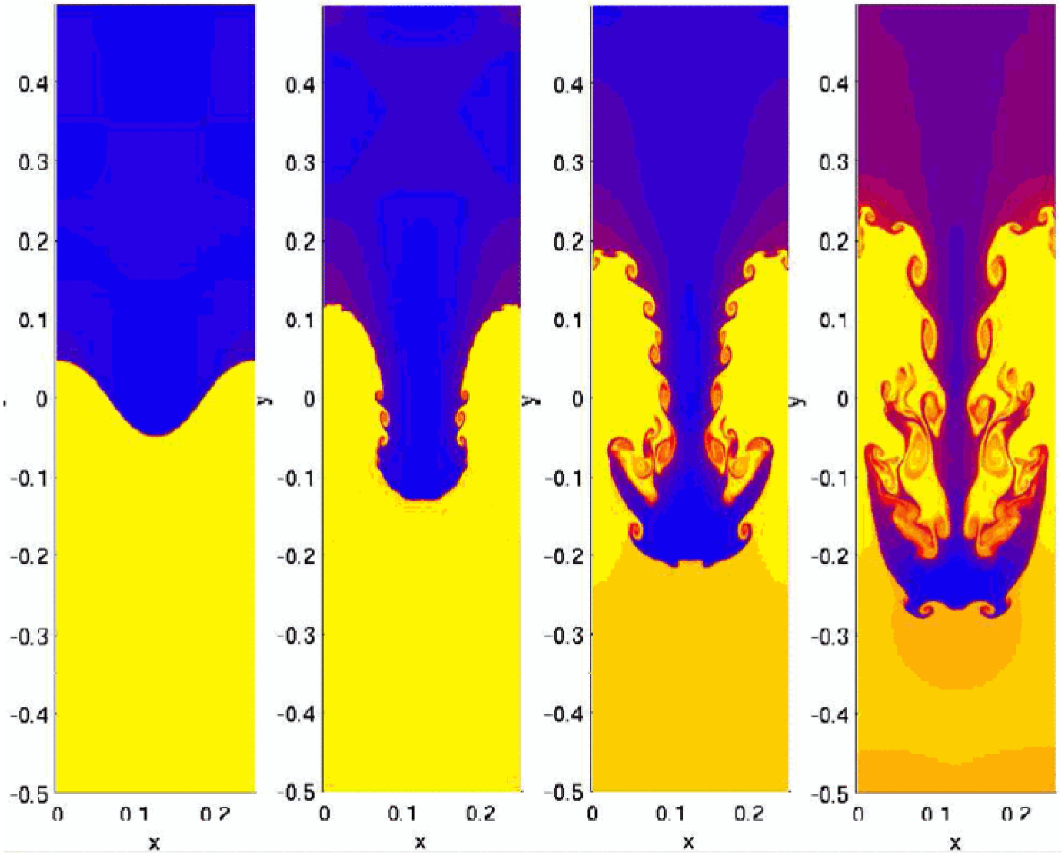
Stokes flow is flow at very low Reynolds numbers, Re<<1, such that inertial forces can be neglected compared to viscous forces.

On the contrary, high Reynolds numbers indicate that the inertial forces are more significant than the viscous (friction) forces. Therefore, we may assume the flow to be an inviscid flow, an approximation in which we neglect viscosity completely, compared to inertial terms.

This idea can work fairly well when the Reynolds number is high. However, certain problems such as those involving solid boundaries, may require that the viscosity be included. Viscosity often cannot be neglected near solid boundaries because the no-slip condition can generate a thin region of large strain rate (known as Boundary layer) which enhances the effect of even a small amount of viscosity, and thus generating vorticity. Therefore, to calculate net forces on bodies (such as wings) we should use viscous flow equations. As illustrated by d'Alembert's paradox, a body in an inviscid fluid will experience no drag force. The standard equations of inviscid flow are the Euler equations. Another often used model, especially in computational fluid dynamics, is to use the Euler equations away from the body and the boundary layer equations, which incorporates viscosity, in a region close to the body.

The Euler equations can be integrated along a streamline to get Bernoulli's equation. When the flow is everywhere irrotational and inviscid, Bernoulli's equation can be used throughout the flow field. Such flows are called potential flows.

Steady vs unsteady flow



Hydrodynamics simulation of the Rayleigh–Taylor instability

When all the time derivatives of a flow field vanish, the flow is considered to be a **steady flow**. Steady-state flow refers to the condition where the fluid properties at a point in the system do not change over time. Otherwise, flow is called unsteady. Whether a particular flow is steady or unsteady, can depend on the chosen frame of reference. For instance, laminar flow over a sphere is steady in the frame of reference that is stationary with respect to the sphere. In a frame of reference that is stationary with respect to a background flow, the flow is unsteady.

Turbulent flows are unsteady by definition. A turbulent flow can, however, be statistically stationary. According to Pope:

The random field $U(x,t)$ is statistically stationary if all statistics are invariant under a shift in time.

This roughly means that all statistical properties are constant in time. Often, the mean field is the object of interest, and this is constant too in a statistically stationary flow.

Steady flows are often more tractable than otherwise similar unsteady flows. The governing equations of a steady problem have one dimension fewer (time) than the governing equations of the same problem without taking advantage of the steadiness of the flow field.

Laminar vs turbulent flow

Turbulence is flow characterized by recirculation, eddies, and apparent randomness. Flow in which turbulence is not exhibited is called laminar. It should be noted, however, that the presence of eddies or recirculation alone does not necessarily indicate turbulent flow—these phenomena may be present in laminar flow as well. Mathematically, turbulent flow is often represented via a Reynolds decomposition, in which the flow is broken down into the sum of an average component and a perturbation component.

It is believed that turbulent flows can be described well through the use of the Navier–Stokes equations. Direct numerical simulation (DNS), based on the Navier–Stokes equations, makes it possible to simulate turbulent flows at moderate Reynolds numbers. Restrictions depend on the power of the computer used and the efficiency of the solution algorithm. The results of DNS have been found to agree well with experimental data for some flows.

Most flows of interest have Reynolds numbers much too high for DNS to be a viable option, given the state of computational power for the next few decades. Any flight vehicle large enough to carry a human (L > 3 m), moving faster than 72 km/h (20 m/s) is well beyond the limit of DNS simulation (Re = 4 million). Transport aircraft wings (such as on an Airbus A300 or Boeing 747) have Reynolds numbers of 40 million (based on the wing chord). In order to solve these real-life flow problems, turbulence models will be a necessity for the foreseeable future. Reynolds-averaged Navier–Stokes equations (RANS) combined with turbulence modeling provides a model of the effects of the turbulent flow. Such a modeling mainly provides the additional momentum transfer by the Reynolds stresses, although the turbulence also enhances the heat and mass transfer. Another promising methodology is large eddy simulation (LES), especially in the guise of detached eddy simulation (DES)—which is a combination of RANS turbulence modeling and large eddy simulation.

Newtonian vs non-Newtonian fluids

Sir Isaac Newton showed how stress and the rate of strain are very close to linearly related for many familiar fluids, such as water and air. These Newtonian fluids are modeled by a coefficient called viscosity, which depends on the specific fluid.

However, some of the other materials, such as emulsions and slurries and some visco-elastic materials (e.g. blood, some polymers), have more complicated *non-Newtonian* stress-strain behaviours. These materials include *sticky liquids* such as latex, honey, and lubricants which are studied in the sub-discipline of rheology.

Subsonic vs transonic, supersonic and hypersonic flows

While many terrestrial flows (e.g. flow of water through a pipe) occur at low mach numbers, many flows of practical interest (e.g. in aerodynamics) occur at high fractions of the Mach Number M=1 or in excess of it (supersonic flows). New phenomena occur at these Mach number regimes (e.g. shock waves for supersonic flow, transonic instability in a regime of flows with M nearly equal to 1, non-equilibrium chemical behavior due to ionization in hypersonic flows) and it is necessary to treat each of these flow regimes separately.

Magnetohydrodynamics

Magnetohydrodynamics is the multi-disciplinary study of the flow of electrically conducting fluids in electromagnetic fields. Examples of such fluids include plasmas, liquid metals, and salt water. The fluid flow equations are solved simultaneously with Maxwell's equations of electromagnetism.

Other approximations

There are a large number of other possible approximations to fluid dynamic problems. Some of the more commonly used are listed below.

- The **Boussinesq approximation** neglects variations in density except to calculate buoyancy forces. It is often used in free convection problems where density changes are small.
- **Lubrication theory** and **Hele-Shaw flow** exploits the large aspect ratio of the domain to show that certain terms in the equations are small and so can be neglected.
- **Slender-body theory** is a methodology used in Stokes flow problems to estimate the force on, or flow field around, a long slender object in a viscous fluid.

- The **shallow-water equations** can be used to describe a layer of relatively inviscid fluid with a free surface, in which surface gradients are small.
- The **Boussinesq equations** are applicable to surface waves on thicker layers of fluid and with steeper surface slopes.
- **Darcy's law** is used for flow in porous media, and works with variables averaged over several pore-widths.
- In rotating systems, the **quasi-geostrophic approximation** assumes an almost perfect balance between pressure gradients and the Coriolis force. It is useful in the study of atmospheric dynamics.

Terminology in fluid dynamics

The concept of pressure is central to the study of both fluid statics and fluid dynamics. A pressure can be identified for every point in a body of fluid, regardless of whether the fluid is in motion or not. Pressure can be measured using an aneroid, Bourdon tube, mercury column, or various other methods.

Some of the terminology that is necessary in the study of fluid dynamics is not found in other similar areas of study. In particular, some of the terminology used in fluid dynamics is not used in fluid statics.

Terminology in incompressible fluid dynamics

The concepts of total pressure and dynamic pressure arise from Bernoulli's equation and are significant in the study of all fluid flows. (These two pressures are not pressures in the usual sense—they cannot be measured using an aneroid, Bourdon tube or mercury column.) To avoid potential ambiguity when referring to pressure in fluid dynamics, many authors use the term static pressure to distinguish it from total pressure and dynamic pressure. Static pressure is identical to pressure and can be identified for every point in a fluid flow field.

In *Aerodynamics*, L.J. Clancy writes: *To distinguish it from the total and dynamic pressures, the actual pressure of the fluid, which is associated not with its motion but with its state, is often referred to as the static pressure, but where the term pressure alone is used it refers to this static pressure.*

A point in a fluid flow where the flow has come to rest (i.e. speed is equal to zero adjacent to some solid body immersed in the fluid flow) is of special significance. It is of such importance that it is given a special name—a stagnation point. The static pressure at the stagnation point is of special significance and is given its own name—stagnation pressure. In incompressible flows, the stagnation pressure at a stagnation point is equal to the total pressure throughout the flow field.

Terminology in compressible fluid dynamics

In a compressible fluid, such as air, the temperature and density are essential when determining the state of the fluid. In addition to the concept of total pressure (also known as stagnation pressure), the concepts of total (or stagnation) temperature and total (or stagnation) density are also essential in any study of compressible fluid flows. To avoid potential ambiguity when referring to temperature and density, many authors use the terms static temperature and static density. Static temperature is identical to temperature; and static density is identical to density; and both can be identified for every point in a fluid flow field.

The temperature and density at a stagnation point are called stagnation temperature and stagnation density.

A similar approach is also taken with the thermodynamic properties of compressible fluids. Many authors use the terms total (or stagnation) enthalpy and total (or stagnation) entropy. The terms static enthalpy and static entropy appear to be less common, but where they are used they mean nothing more than enthalpy and entropy respectively, and the prefix "static" is being used to avoid ambiguity with their 'total' or 'stagnation' counterparts. Because the 'total' flow conditions are defined by isentropically bringing the fluid to rest, the total (or stagnation) entropy is by definition always equal to the "static" entropy.

Chapter 2

Solid Mechanics and Solid State (Electronics)

Solid mechanics

Solid mechanics is the branch of mechanics, physics, and mathematics that concerns the behavior of solid matter under external actions (e.g., external forces, temperature changes, applied displacements, etc.). It is part of a broader study known as continuum mechanics. One of the most common practical applications of solid mechanics is the Euler-Bernoulli beam equation. Solid mechanics extensively uses tensors to describe stresses, strains, and the relationship between them.

Relationship to continuum mechanics

As shown in the following table, solid mechanics inhabits a central place within continuum mechanics. The field of rheology presents an overlap between solid and fluid mechanics.

Continuum mechanics The study of the physics of continuous materials	**Solid mechanics** The study of the physics of continuous materials with a defined rest shape.	Elasticity Describes materials that return to their rest shape after an applied stress.	
		Plasticity Describes materials that permanently deform after a sufficient applied stress.	Rheology The study of materials with both solid and fluid characteristics.
	Fluid mechanics The study of the physics of continuous materials which take the shape of their container.	Non-Newtonian fluids	
		Newtonian fluids	

Response models

A material has a rest shape and its shape departs away from the rest shape due to stress. The amount of departure from rest shape is called deformation, the proportion of deformation to original size is called strain. If the applied stress is sufficiently low (or the imposed strain is small enough), almost all solid materials behave in such a way that the strain is directly proportional to the stress; the coefficient of the proportion is called the modulus of elasticity or Young's modulus. This region of deformation is known as the linearly elastic region.

It is most common for analysts in solid mechanics to use linear material models, due to ease of computation. However, real materials often exhibit non-linear behavior. As new materials are used and old ones are pushed to their limits, non-linear material models are becoming more common.

There are three models that describe how a solid responds to an applied stress:

1. Elastically – When an applied stress is removed, the material returns to its undeformed state. Linearly elastic materials, those that deform proportionally to the applied load, can be described by the linear elasticity equations such as Hooke's law.
2. Viscoelastically – These are materials that behave elastically, but also have damping: when the stress is applied and removed, work has to be done against the damping effects and is converted in heat within the material resulting in a hysteresis loop in the stress–strain curve. This implies that the material response has time-dependence.
3. Plastically – Materials that behave elastically generally do so when the applied stress is less than a yield value. When the stress is greater than the yield stress, the material behaves plastically and does not return to its previous state. That is, deformation that occurs after yield is permanent.

Solid state (electronics)

Solid-state electronics are those circuits or devices built entirely from solid materials and in which the electrons, or other charge carriers, are confined entirely within the solid material. The term is often used to contrast with the earlier technologies of vacuum and gas-discharge tube devices and it is also conventional to exclude electro-mechanical devices (relays, switches, hard drives and other devices with moving parts) from the term solid state. While solid-state can include crystalline, polycrystalline and amorphous solids and refer to electrical conductors, insulators and semiconductors, the building material is most often crystalline semiconductor. Common solid-state devices include transistors, microprocessor chips, and DRAM. DRAM devices are used in computers, flash drives and more recently, solid state drives to replace mechanically rotating magnetic disc hard drives. A considerable amount of electromagnetic and quantum-mechanical action takes place within the device. The expression became prevalent in the 1950s and the 1960s, during the transition from vacuum tube technology to semiconductor diodes and transistors. More recently, the integrated circuit (IC), the light-emitting diode (LED), and the liquid-crystal display (LCD) have evolved as further examples of solid-state devices.

In a solid-state component, the current is confined to solid elements and compounds engineered specifically to switch and amplify it. Current flow can be understood in two forms: as negatively-charged electrons, and as positively-charged electron deficiencies called electron holes or just "holes". In some semiconductors, the current consists mostly of electrons; in other semiconductors, it consists mostly of "holes". Both the electron and the hole are called charge carriers.

For data storage, solid-state devices are much faster and more reliable but are usually more expensive. Although solid-state costs continually drop, disks, tapes, and optical disks also continue to improve their cost/performance ratio.

The first solid-state device was the "cat's whisker" detector, first used in 1930s radio receivers. A whisker-like wire was moved around on a solid crystal (such as a germanium crystal) in order to detect a radio signal. The solid-state device came into its own with the invention of the transistor in 1947.

Chapter 3

Operations Research

Operations research (also referred to as **decision science**, or **management science**) is an interdisciplinary mathematical science that focuses on the effective *use* of technology by organizations. In contrast, many other science & engineering disciplines focus on technology giving secondary considerations to its use.

Employing techniques from other mathematical sciences — such as mathematical modeling, statistical analysis, and mathematical optimization — operations research arrives at optimal or near-optimal solutions to complex decision-making problems. Because of its emphasis on human-technology interaction and because of its focus on practical applications, operations research has overlap with other disciplines, notably industrial engineering and management science, and draws on psychology and organization science. Operations Research is often concerned with determining the maximum (of profit, performance, or yield) or minimum (of loss, risk, or cost) of some real-world objective. Originating in military efforts before World War II, its techniques have grown to concern problems in a variety of industries.

Overview

Operational research encompasses a wide range of problem-solving techniques and methods applied in the pursuit of improved decision-making and efficiency. Some of the tools used by operational researchers are statistics, optimization, probability theory, queuing theory, game theory, graph theory, decision analysis, mathematical modeling and simulation. Because of the computational nature of these fields, OR also has strong ties to computer science and analytics. Operational researchers faced with a new problem must determine which of these techniques are most appropriate given the nature of the system, the goals for improvement, and constraints on time and computing power.

Work in operational research and management science may be characterized as one of three categories:

- Fundamental or foundational work takes place in three mathematical disciplines: probability, optimization, and dynamical systems theory.

- Modeling work is concerned with the construction of models, analyzing them mathematically, implementing them on computers, solving them using software tools, and assessing their effectiveness with data. This level is mainly instrumental, and driven mainly by statistics and econometrics.
- Application work in operational research, like other engineering and economics' disciplines, attempts to use models to make a practical impact on real-world problems.

The major subdisciplines in modern operational research, as identified by the journal *Operations Research*, are:

- Computing and information technologies
- Decision analysis
- Environment, energy, and natural resources
- Financial engineering
- Manufacturing, service sciences, and supply chain management
- Marketing Engineering
- Policy modeling and public sector work
- Revenue management
- Simulation
- Stochastic models
- Transportation

History

As a formal discipline, operational research originated in the efforts of military planners during World War II. In the decades after the war, the techniques began to be applied more widely to problems in business, industry and society. Since that time, operational research has expanded into a field widely used in industries ranging from petrochemicals to airlines, finance, logistics, and government, moving to a focus on the development of mathematical models that can be used to analyze and optimize complex systems, and has become an area of active academic and industrial research.

Historical origins

In the World War II era, operational research was defined as "a scientific method of providing executive departments with a quantitative basis for decisions regarding the operations under their control." Other names for it included operational analysis (UK Ministry of Defence from 1962) and quantitative management.

Prior to the formal start of the field, early work in operational research was carried out by individuals such as Charles Babbage. His research into the cost of transportation and sorting of mail led to England's universal "Penny Post" in 1840, and studies into the dynamical behaviour of railway vehicles in defence of the GWR's broad gauge. Percy Bridgman brought operational research to bear on problems in physics in the 1920s and would later attempt to extend these to the social sciences. The modern field of operational research arose during World War II.

Modern operational research originated at the Bawdsey Research Station in the UK in 1937 and was the result of an initiative of the station's superintendent, A. P. Rowe. Rowe conceived the idea as a means to analyse and improve the working of the UK's early warning radar system, Chain Home (CH). Initially, he analyzed the operating of the radar equipment and its communication networks, expanding later to include the operating personnel's behaviour. This revealed unappreciated limitations of the CH network and allowed remedial action to be taken.

Scientists in the United Kingdom including Patrick Blackett later Lord Blackett OM PRS, Cecil Gordon, C. H. Waddington, Owen Wansbrough-Jones, Frank Yates, Jacob Bronowski and Freeman Dyson, and in the United States with George Dantzig looked for ways to make better decisions in such areas as logistics and training schedules. After the war it began to be applied to similar problems in industry.

Second World War

Patrick Blackett

During the Second World War close to 1,000 men and women in Britain were engaged in operational research. About 200 operational research scientists worked for the British Army.

Patrick Blackett worked for several different organizations during the war. Early in the war while working for the Royal Aircraft Establishment (RAE) he set up a team known as the "Circus" which helped to reduce the number of anti-aircraft artillery rounds needed to shoot down an enemy aircraft from an average of over 20,000 at the start of the Battle of Britain to 4,000 in 1941.

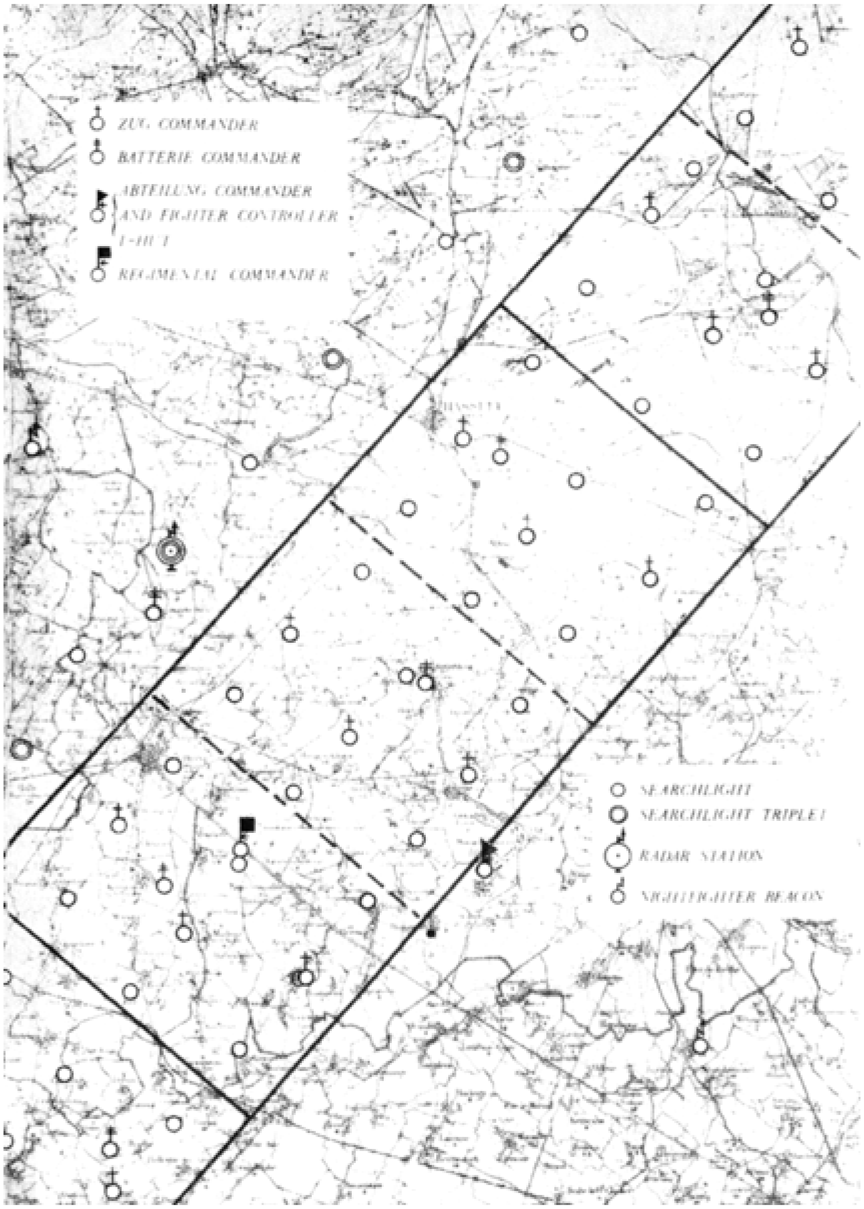
In 1941 Blackett moved from the RAE to the Navy, first to the Royal Navy's Coastal Command, in 1941 and then early in 1942 to the Admiralty. Blackett's team at Coastal Command's Operational Research Section (CC-ORS) included two future Nobel prize winners and many other people who went on to be preeminent in their fields. They undertook a number of crucial analyses that aided the war effort. Britain introduced the convoy system to reduce shipping losses, but while the principle of using warships to accompany merchant ships was generally accepted, it was unclear whether it was better for convoys to be small or large. Convoys travel at the speed of the slowest member, so small convoys can travel faster. It was also argued that small convoys would be harder for German U-boats to detect. On the other hand, large convoys could deploy more warships against an attacker. Blackett's staff showed that the losses suffered by convoys depended largely on the number of escort vessels present, rather than on the overall size of the convoy. Their conclusion, therefore, was that a few large convoys are more defensible than many small ones.

While performing an analysis of the methods used by RAF Coastal Command to hunt and destroy submarines, one of the analysts asked what colour the aircraft were. As most of them were from Bomber Command they were painted black for nighttime operations. At the suggestion of CC-ORS a test was run to see if that was the best colour to camouflage the aircraft for daytime operations in the grey North Atlantic skies. Tests showed that aircraft painted white were on average not spotted until they were 20% closer than those painted black. This change indicated that 30% more submarines would be attacked and sunk for the same number of sightings.

Other work by the CC-ORS indicated that on average if the trigger depth of aerial delivered depth charges (DCs) was changed from 100 feet to 25 feet, the kill ratios would go up. The reason was that if a U-boat saw an aircraft only shortly before it arrived over the target then at 100 feet the charges would do no damage (because the U-boat wouldn't have time to descend as far as 100 feet), and if it saw the aircraft a long way from the target it had time to alter course under water so the chances of it being within the 20 feet kill zone of the charges was small. It was more efficient to attack those submarines close to the surface when these targets' locations were better known than to attempt their destruction at greater depths when their positions could only be guessed. Before the change of settings from 100 feet to 25 feet, 1% of submerged U-boats were sunk and 14% damaged. After the change, 7% were sunk and 11% damaged. (If submarines were caught on the surface, even if attacked shortly after submerging, the numbers rose to 11% sunk and 15% damaged). Blackett observed "there can be few cases where such a great operational gain had been obtained by such a small and simple change of tactics".

Bomber Command's Operational Research Section (BC-ORS), analysed a report of a survey carried out by RAF Bomber Command. For the survey, Bomber Command inspected all bombers returning from bombing raids over Germany over a particular period. All damage inflicted by German air defences was noted and the recommendation was given that armour be added in the most heavily damaged areas. Their suggestion to remove some of the crew so that an aircraft loss would result in fewer personnel loss was rejected by RAF command. Blackett's team instead made the surprising and counter-intuitive recommendation that the armour be placed in the areas which were completely untouched by damage in the bombers which returned. They reasoned that the survey was biased, since it only included aircraft that returned to Britain. The untouched areas of returning aircraft were probably vital areas, which, if hit, would result in the loss of the aircraft.

(Info to help verify some/most of the above paragraph is found in "Dirty Little Secrets of the Twentieth Century" authored by James F. Dunnigan on pages 215-217



Map of *Kammhuber Line*

When Germany organised its air defences into the Kammhuber Line, it was realised that if the RAF bombers were to fly in a bomber stream they could overwhelm the night fighters who flew in individual cells directed to their targets by ground controllers. It was then a matter of calculating the statistical loss from collisions against the statistical loss from night fighters to calculate how close the bombers should fly to minimise RAF losses.

The "exchange rate" ratio of output to input was a characteristic feature of operational research. By comparing the number of flying hours put in by Allied aircraft to the number of U-boat sightings in a given area, it was possible to redistribute aircraft to more productive patrol areas. Comparison of exchange rates established "effectiveness ratios" useful in planning. The ratio of 60 mines laid per ship sunk was common to several campaigns: German mines in British ports, British mines on German routes, and United States mines in Japanese routes.

Operational research doubled the on-target bomb rate of B-29s bombing Japan from the Marianas Islands by increasing the training ratio from 4 to 10 percent of flying hours; revealed that wolf-packs of three United States submarines were the most effective number to enable all members of the pack to engage targets discovered on their individual patrol stations; revealed that glossy enamel paint was more effective camouflage for night fighters than traditional dull camouflage paint finish, and the smooth paint finish increased airspeed by reducing skin friction.

On land, the operational research sections of the Army Operational Research Group (AORG) of the Ministry of Supply (MoS) were landed in Normandy in 1944, and they followed British forces in the advance across Europe. They analysed, among other topics, the effectiveness of artillery, aerial bombing, and anti-tank shooting.

After World War II

With expanded techniques and growing awareness of the field at the close of the war, operational research was no longer limited to only operational, but was extended to encompass equipment procurement, training, logistics and infrastructure.

Academic Denis Bouyssou describes the historical development of operational research from the 1940s to the 1970s as follows. "The historical development of Operational Research (OR) is traditionally seen as the succession of several phases: the 'heroic times' of the Second World War, the 'Golden Age' between the fifties and the sixties during which major theoretical achievements were accompanied by a widespread diffusion of OR techniques in private and public organisations, a 'crisis' followed by a 'decline' starting with the late sixties, a phase during which OR groups in firms progressively disappeared while academia became less and less concerned with the applicability of the techniques developed".

Individuals such as Stafford Beer and George Dantzig pioneered early academic efforts in operational research.

Problems addressed with operational research

- critical path analysis or project planning: identifying those processes in a complex project which affect the overall duration of the project
- floorplanning: designing the layout of equipment in a factory or components on a computer chip to reduce manufacturing time (therefore reducing cost)

- network optimization: for instance, setup of telecommunications networks to maintain quality of service during outages
- allocation problems
- Bayesian search theory : looking for a target
- optimal search
- routing, such as determining the routes of buses so that as few buses are needed as possible
- supply chain management: managing the flow of raw materials and products based on uncertain demand for the finished products
- efficient messaging and customer response tactics
- automation: automating or integrating robotic systems in human-driven operations processes
- globalization: globalizing operations processes in order to take advantage of cheaper materials, labor, land or other productivity inputs
- transportation: managing freight transportation and delivery systems (Examples: LTL Shipping, intermodal freight transport)
- scheduling:
    - personnel staffing
    - manufacturing steps
    - project tasks
    - network data traffic: these are known as queueing models or queueing systems.
    - sports events and their television coverage
- blending of raw materials in oil refineries
- determining optimal prices, in many retail and B2B settings, within the disciplines of pricing science

Operational research is also used extensively in government where evidence-based policy is used.

Management science

In 1967 Stafford Beer characterized the field of management science as "the business use of operations research". However, in modern times the term management science may also be used to refer to the separate fields of organizational studies or corporate strategy. Like operational research itself, management science (MS), is an interdisciplinary branch of applied mathematics devoted to optimal decision planning, with strong links with economics, business, engineering, and other sciences. It uses various scientific research-based principles, strategies, and analytical methods including mathematical modeling, statistics and numerical algorithms to improve an organization's ability to enact rational and meaningful management decisions by arriving at optimal or near optimal solutions to complex decision problems. In short, management sciences help businesses to achieve their goals using the scientific methods of operational research.

The management scientist's mandate is to use rational, systematic, science-based techniques to inform and improve decisions of all kinds. Of course, the techniques of management science are not restricted to business applications but may be applied to military, medical, public administration, charitable groups, political groups or community groups.

Management science is concerned with developing and applying models and concepts that may prove useful in helping to illuminate management issues and solve managerial problems, as well as designing and developing new and better models of organizational excellence.

The application of these models within the corporate sector became known as Management science.

Techniques

Some of the fields that have considerable overlap with Management Science include:

- Data mining
- Decision analysis
- Engineering
- Forecasting
- Game theory
- Industrial engineering
- Logistics
- Mathematical modeling
- Optimization
- Probability and statistics
- Project management
- Simulation
- Social network/Transportation forecasting models
- Supply chain management
- Financial engineering

Applications of management science

Applications of management science are abundant in industry as airlines, manufacturing companies, service organizations, military branches, and in government. The range of problems and issues to which management science has contributed insights and solutions is vast. It includes:

- scheduling airlines, including both planes and crew,
- deciding the appropriate place to site new facilities such as a warehouse, factory or fire station,
- managing the flow of water from reservoirs,
- identifying possible future development paths for parts of the telecommunications industry,
- establishing the information needs and appropriate systems to supply them within the health service, and
- identifying and understanding the strategies adopted by companies for their information systems

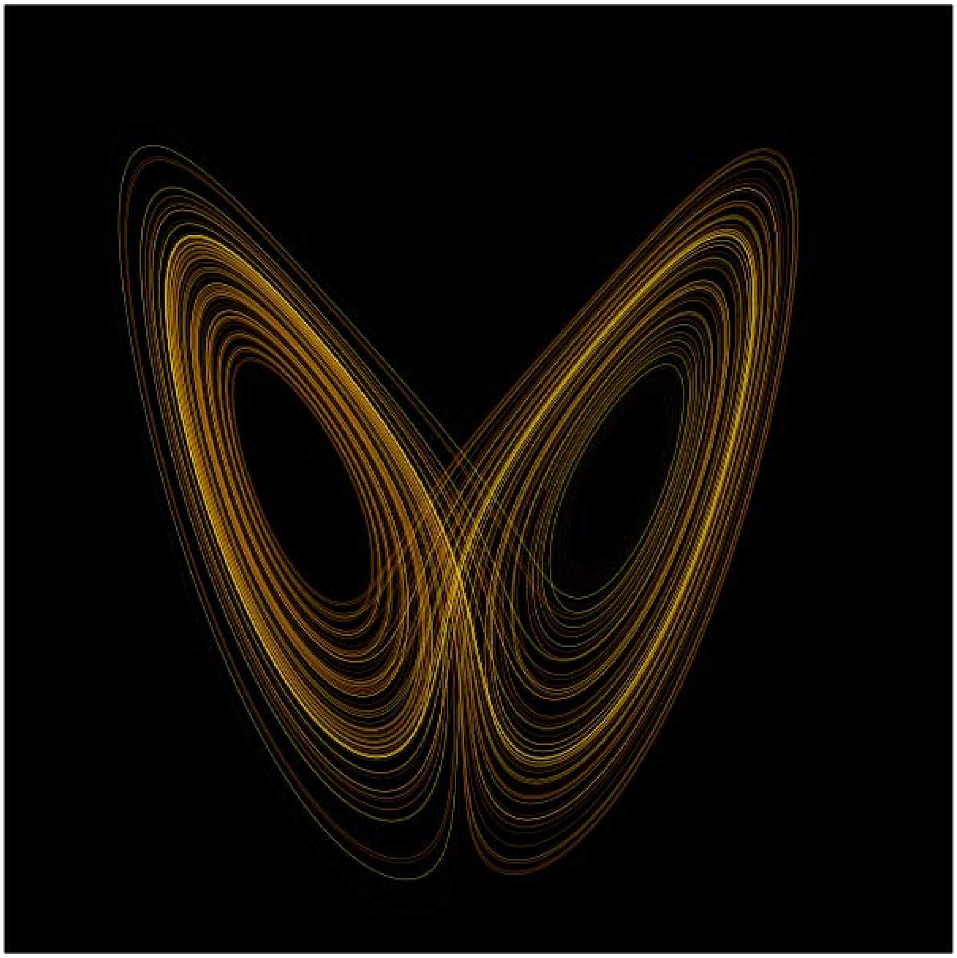
Management science is also concerned with so-called ”soft-operational analysis”, which concerns methods for strategic planning, strategic decision support, and Problem

Structuring Methods (PSM). In dealing with these sorts of challenges mathematical modeling and simulation are not appropriate or will not suffice. Therefore, during the past 30 years, a number of non-quantified modeling methods have been developed. These include:

- stakeholder based approaches including metagame analysis and drama theory
- morphological analysis and various forms of influence diagrams.
- approaches using cognitive mapping
- the Strategic Choice Approach
- robustness analysis

Chapter 4

Dynamical System



The Lorenz attractor is an example of a non-linear dynamical system. Studying this system helped give rise to Chaos theory.

A **dynamical system** is a concept in mathematics where a fixed rule describes the time dependence of a point in a geometrical space. Examples include the mathematical models that describe the swinging of a clock pendulum, the flow of water in a pipe, and the number of fish each spring in a lake.

At any given time a dynamical system has a *state* given by a set of real numbers (a vector) which can be represented by a point in an appropriate *state space* (a geometrical manifold). Small changes in the state of the system correspond to small changes in the numbers. The *evolution rule* of the dynamical system is a fixed rule that describes what future states follow from the current state. The rule is deterministic; in other words, for a given time interval only one future state follows from the current state.

Overview

The concept of a dynamical system has its origins in Newtonian mechanics. There, as in other natural sciences and engineering disciplines, the evolution rule of dynamical systems is given implicitly by a relation that gives the state of the system only a short time into the future. (The relation is either a differential equation, difference equation or other time scale.) To determine the state for all future times requires iterating the relation many times—each advancing time a small step. The iteration procedure is referred to as *solving the system* or *integrating the system*. Once the system can be solved, given an initial point it is possible to determine all its future points, a collection known as a *trajectory* or *orbit*.

Before the advent of fast computing machines, solving a dynamical system required sophisticated mathematical techniques and could be accomplished only for a small class of dynamical systems. Numerical methods implemented on electronic computing machines have simplified the task of determining the orbits of a dynamical system.

For simple dynamical systems, knowing the trajectory is often sufficient, but most dynamical systems are too complicated to be understood in terms of individual trajectories. The difficulties arise because:

- The systems studied may only be known approximately—the parameters of the system may not be known precisely or terms may be missing from the equations. The approximations used bring into question the validity or relevance of numerical solutions. To address these questions several notions of stability have been introduced in the study of dynamical systems, such as Lyapunov stability or structural stability. The stability of the dynamical system implies that there is a class of models or initial conditions for which the trajectories would be equivalent. The operation for comparing orbits to establish their equivalence changes with the different notions of stability.
- The type of trajectory may be more important than one particular trajectory. Some trajectories may be periodic, whereas others may wander through many different states of the system. Applications often require enumerating these classes or maintaining the system within one class. Classifying all possible trajectories has

led to the qualitative study of dynamical systems, that is, properties that do not change under coordinate changes. Linear dynamical systems and systems that have two numbers describing a state are examples of dynamical systems where the possible classes of orbits are understood.

- The behavior of trajectories as a function of a parameter may be what is needed for an application. As a parameter is varied, the dynamical systems may have bifurcation points where the qualitative behavior of the dynamical system changes. For example, it may go from having only periodic motions to apparently erratic behavior, as in the transition to turbulence of a fluid.
- The trajectories of the system may appear erratic, as if random. In these cases it may be necessary to compute averages using one very long trajectory or many different trajectories. The averages are well defined for ergodic systems and a more detailed understanding has been worked out for hyperbolic systems. Understanding the probabilistic aspects of dynamical systems has helped establish the foundations of statistical mechanics and of chaos.

It was in the work of Poincaré that these dynamical systems themes developed.

Basic definitions

A dynamical system is a manifold *M* called the phase (or state) space endowed with a family of smooth evolution functions Φ^t that for any element of $t \in T$, the time, map a point of the phase space back into the phase space. The notion of smoothness changes with applications and the type of manifold. There are several choices for the set *T*. When *T* is taken to be the reals, the dynamical system is called a *flow*; and if *T* is restricted to the non-negative reals, then the dynamical system is a *semi-flow*. When *T* is taken to be the integers, it is a *cascade* or a *map*; and the restriction to the non-negative integers is a *semi-cascade*.

Examples

The evolution function Φ^t is often the solution of a *differential equation of motion*

$$\dot{x} = v(x) \, .$$

The equation gives the time derivative, represented by the dot, of a trajectory *x(t)* on the phase space starting at some point x_0. The *vector field v(x)* is a smooth function that at every point of the phase space *M* provides the velocity vector of the dynamical system at that point. (These vectors are not vectors in the phase space *M*, but in the tangent space T_xM of the point *x*.) Given a smooth Φ^t, an autonomous vector field can be derived from it.

There is no need for higher order derivatives in the equation, nor for time dependence in *v(x)* because these can be eliminated by considering systems of higher dimensions. Other types of differential equations can be used to define the evolution rule:

$$G(x, \dot{x}) = 0$$

is an example of an equation that arises from the modeling of mechanical systems with complicated constraints.

The differential equations determining the evolution function Φ^t are often ordinary differential equations: in this case the phase space *M* is a finite dimensional manifold. Many of the concepts in dynamical systems can be extended to infinite-dimensional manifolds—those that are locally Banach spaces—in which case the differential equations are partial differential equations. In the late 20th century the dynamical system perspective to partial differential equations started gaining popularity.

Further examples

- Logistic map
- Dyadic transformation
- Tent map
- Double pendulum
- Arnold's cat map
- Horseshoe map
- Baker's map is an example of a chaotic piecewise linear map
- Billiards and outer billiards
- Hénon map
- Lorenz system
- Circle map
- Rössler map
- List of chaotic maps
- Swinging Atwood's machine
- Quadratic map simulation system
- Bouncing ball dynamics

Linear dynamical systems

Linear dynamical systems can be solved in terms of simple functions and the behavior of all orbits classified. In a linear system the phase space is the N-dimensional Euclidean space, so any point in phase space can be represented by a vector with N numbers. The analysis of linear systems is possible because they satisfy a superposition principle: if *u(t)* and *w(t)* satisfy the differential equation for the vector field (but not necessarily the initial condition), then so will *u(t)* + *w(t)*.

Flows

For a flow, the vector field *Φ(x)* is a linear function of the position in the phase space, that is,

$$\phi(x) = Ax + b\,,$$

with A a matrix, b a vector of numbers and x the position vector. The solution to this system can be found by using the superposition principle (linearity). The case $b \neq 0$ with $A = 0$ is just a straight line in the direction of b:

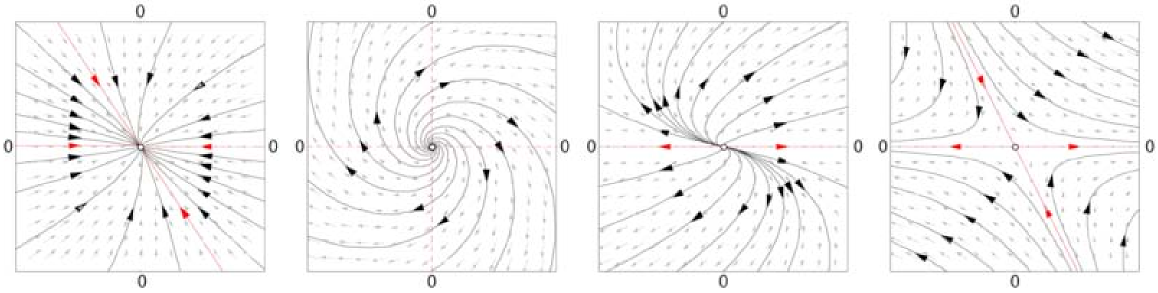
$$\Phi^t(x_1) = x_1 + bt \, .$$

When b is zero and $A \neq 0$ the origin is an equilibrium (or singular) point of the flow, that is, if $x_0 = 0$, then the orbit remains there. For other initial conditions, the equation of motion is given by the exponential of a matrix: for an initial point x_0,

$$\Phi^t(x_0) = e^{tA} x_0 \, .$$

When $b = 0$, the eigenvalues of A determine the structure of the phase space. From the eigenvalues and the eigenvectors of A it is possible to determine if an initial point will converge or diverge to the equilibrium point at the origin.

The distance between two different initial conditions in the case $A \neq 0$ will change exponentially in most cases, either converging exponentially fast towards a point, or diverging exponentially fast. Linear systems display sensitive dependence on initial conditions in the case of divergence. For nonlinear systems this is one of the (necessary but not sufficient) conditions for chaotic behavior.



Linear vector fields and a few trajectories.

Maps

A discrete-time, affine dynamical system has the form

$$x_{n+1} = Ax_n + b \, ,$$

with A a matrix and b a vector. As in the continuous case, the change of coordinates $x \rightarrow x + (1 - A)^{-1}b$ removes the term b from the equation. In the new coordinate system, the origin is a fixed point of the map and the solutions are of the linear system $A^n x_0$. The solutions for the map are no longer curves, but points that hop in the phase space. The orbits are organized in curves, or fibers, which are collections of points that map into themselves under the action of the map.

As in the continuous case, the eigenvalues and eigenvectors of *A* determine the structure of phase space. For example, if u_1 is an eigenvector of *A*, with a real eigenvalue smaller than one, then the straight lines given by the points along $\alpha\ u_1$, with $\alpha \in \mathbf{R}$, is an invariant curve of the map. Points in this straight line run into the fixed point.

There are also many other discrete dynamical systems.

Local dynamics

The qualitative properties of dynamical systems do not change under a smooth change of coordinates (this is sometimes taken as a definition of qualitative): a *singular point* of the vector field (a point where $v(x) = 0$) will remain a singular point under smooth transformations; a *periodic orbit* is a loop in phase space and smooth deformations of the phase space cannot alter it being a loop. It is in the neighborhood of singular points and periodic orbits that the structure of a phase space of a dynamical system can be well understood. In the qualitative study of dynamical systems, the approach is to show that there is a change of coordinates (usually unspecified, but computable) that makes the dynamical system as simple as possible.

Rectification

A flow in most small patches of the phase space can be made very simple. If *y* is a point where the vector field $v(y) \neq 0$, then there is a change of coordinates for a region around *y* where the vector field becomes a series of parallel vectors of the same magnitude. This is known as the rectification theorem.

The rectification theorem says that away from singular points the dynamics of a point in a small patch is a straight line. The patch can sometimes be enlarged by stitching several patches together, and when this works out in the whole phase space *M* the dynamical system is *integrable*. In most cases the patch cannot be extended to the entire phase space. There may be singular points in the vector field (where $v(x) = 0$); or the patches may become smaller and smaller as some point is approached. The more subtle reason is a global constraint, where the trajectory starts out in a patch, and after visiting a series of other patches comes back to the original one. If the next time the orbit loops around phase space in a different way, then it is impossible to rectify the vector field in the whole series of patches.

Near periodic orbits

In general, in the neighborhood of a periodic orbit the rectification theorem cannot be used. Poincaré developed an approach that transforms the analysis near a periodic orbit to the analysis of a map. Pick a point x_0 in the orbit γ and consider the points in phase space in that neighborhood that are perpendicular to $v(x_0)$. These points are a Poincaré section $S(\gamma, x_0)$, of the orbit. The flow now defines a map, the Poincaré map $F : S \to S$, for points starting in *S* and returning to *S*. Not all these points will take the same amount of time to come back, but the times will be close to the time it takes x_0.

The intersection of the periodic orbit with the Poincaré section is a fixed point of the Poincaré map F. By a translation, the point can be assumed to be at $x = 0$. The Taylor series of the map is $F(x) = J \cdot x + O(x^2)$, so a change of coordinates h can only be expected to simplify F to its linear part

$$h^{-1} \circ F \circ h(x) = J \cdot x \,.$$

This is known as the conjugation equation. Finding conditions for this equation to hold has been one of the major tasks of research in dynamical systems. Poincaré first approached it assuming all functions to be analytic and in the process discovered the non-resonant condition. If $\lambda_1,\ldots,\lambda_\nu$ are the eigenvalues of J they will be resonant if one eigenvalue is an integer linear combination of two or more of the others. As terms of the form $\lambda_i - \sum$ (multiples of other eigenvalues) occurs in the denominator of the terms for the function h, the non-resonant condition is also known as the small divisor problem.

Conjugation results

The results on the existence of a solution to the conjugation equation depend on the eigenvalues of J and the degree of smoothness required from h. As J does not need to have any special symmetries, its eigenvalues will typically be complex numbers. When the eigenvalues of J are not in the unit circle, the dynamics near the fixed point x_0 of F is called *hyperbolic* and when the eigenvalues are on the unit circle and complex, the dynamics is called *elliptic*.

In the hyperbolic case the Hartman-Grobman theorem gives the conditions for the existence of a continuous function that maps the neighborhood of the fixed point of the map to the linear map $J \cdot x$. The hyperbolic case is also *structurally stable*. Small changes in the vector field will only produce small changes in the Poincaré map and these small changes will reflect in small changes in the position of the eigenvalues of J in the complex plane, implying that the map is still hyperbolic.

The Kolmogorov-Arnold-Moser (KAM) theorem gives the behavior near an elliptic point.

Bifurcation theory

When the evolution map Φ^t (or the vector field it is derived from) depends on a parameter μ, the structure of the phase space will also depend on this parameter. Small changes may produce no qualitative changes in the phase space until a special value μ_0 is reached. At this point the phase space changes qualitatively and the dynamical system is said to have gone through a bifurcation.

Bifurcation theory considers a structure in phase space (typically a fixed point, a periodic orbit, or an invariant torus) and studies its behavior as a function of the parameter μ. At the bifurcation point the structure may change its stability, split into new structures, or merge with other structures. By using Taylor series approximations of the maps and an understanding of the differences that may be eliminated by a change of coordinates, it is possible to catalog the bifurcations of dynamical systems.

The bifurcations of a hyperbolic fixed point x_0 of a system family F_μ can be characterized by the eigenvalues of the first derivative of the system $DF_\mu(x_0)$ computed at the bifurcation point. For a map, the bifurcation will occur when there are eigenvalues of DF_μ on the unit circle. For a flow, it will occur when there are eigenvalues on the imaginary axis.

Some bifurcations can lead to very complicated structures in phase space. For example, the Ruelle–Takens scenario describes how a periodic orbit bifurcates into a torus and the torus into a strange attractor. In another example, Feigenbaum period-doubling describes how a stable periodic orbit goes through a series of period-doubling bifurcations.

Ergodic systems

In many dynamical systems it is possible to choose the coordinates of the system so that the volume (really a ν-dimensional volume) in phase space is invariant. This happens for mechanical systems derived from Newton's laws as long as the coordinates are the position and the momentum and the volume is measured in units of (position) × (momentum). The flow takes points of a subset A into the points Φ^t(A) and invariance of the phase space means that

$$\mathrm{vol}(A) = \mathrm{vol}(\Phi^t(A)).$$

In the Hamiltonian formalism, given a coordinate it is possible to derive the appropriate (generalized) momentum such that the associated volume is preserved by the flow. The volume is said to be computed by the Liouville measure.

In a Hamiltonian system not all possible configurations of position and momentum can be reached from an initial condition. Because of energy conservation, only the states with the same energy as the initial condition are accessible. The states with the same energy form an energy shell Ω, a sub-manifold of the phase space. The volume of the energy shell, computed using the Liouville measure, is preserved under evolution.

For systems where the volume is preserved by the flow, Poincaré discovered the recurrence theorem: Assume the phase space has a finite Liouville volume and let F be a phase space volume-preserving map and A a subset of the phase space. Then almost every point of A returns to A infinitely often. The Poincaré recurrence theorem was used by Zermelo to object to Boltzmann's derivation of the increase in entropy in a dynamical system of colliding atoms.

One of the questions raised by Boltzmann's work was the possible equality between time averages and space averages, what he called the ergodic hypothesis. The hypothesis states that the length of time a typical trajectory spends in a region A is vol(A)/vol(Ω).

The ergodic hypothesis turned out not to be the essential property needed for the development of statistical mechanics and a series of other ergodic-like properties were introduced to capture the relevant aspects of physical systems. Koopman approached the study of ergodic systems by the use of functional analysis. An observable *a* is a function that to each point of the phase space associates a number (say instantaneous pressure, or average height). The value of an observable can be computed at another time by using the evolution function φ^{t}. This introduces an operator U^{t}, the transfer operator,

$$(U^t a)(x) = a(\Phi^{-t}(x))\,.$$

By studying the spectral properties of the linear operator *U* it becomes possible to classify the ergodic properties of Φ^{t}. In using the Koopman approach of considering the action of the flow on an observable function, the finite-dimensional nonlinear problem involving Φ^{t} gets mapped into an infinite-dimensional linear problem involving *U*.

The Liouville measure restricted to the energy surface Ω is the basis for the averages computed in equilibrium statistical mechanics. An average in time along a trajectory is equivalent to an average in space computed with the Boltzmann factor $\exp(-\beta H)$. This idea has been generalized by Sinai, Bowen, and Ruelle (SRB) to a larger class of dynamical systems that includes dissipative systems. SRB measures replace the Boltzmann factor and they are defined on attractors of chaotic systems.

Nonlinear dynamical systems and chaos

Simple nonlinear dynamical systems and even piecewise linear systems can exhibit a completely unpredictable behavior, which might seem to be random. (Remember that we are speaking of completely deterministic systems!). This seemingly unpredictable behavior has been called *chaos*. Hyperbolic systems are precisely defined dynamical systems that exhibit the properties ascribed to chaotic systems. In hyperbolic systems the tangent space perpendicular to a trajectory can be well separated into two parts: one with the points that converge towards the orbit (the *stable manifold*) and another of the points that diverge from the orbit (the *unstable manifold*).

This branch of mathematics deals with the long-term qualitative behavior of dynamical systems. Here, the focus is not on finding precise solutions to the equations defining the dynamical system (which is often hopeless), but rather to answer questions like "Will the system settle down to a steady state in the long term, and if so, what are the possible attractors?" or "Does the long-term behavior of the system depend on its initial condition?"

Note that the chaotic behavior of complicated systems is not the issue. Meteorology has been known for years to involve complicated—even chaotic—behavior. Chaos theory has been so surprising because chaos can be found within almost trivial systems. The logistic map is only a second-degree polynomial; the horseshoe map is piecewise linear.

Geometrical definition

A dynamical system is the tuple $\langle \mathcal{M}, f, \mathcal{T} \rangle$, with $\mathcal{M}$a manifold (locally a Banach space or Euclidean space), $\mathcal{T}$ the domain for time (non-negative reals, the integers, ...) and *f* an evolution rule t→f^t (with $t \in \mathcal{T}$) such that f^t is a diffeomorphism of the manifold to itself. So, f is a mapping of the time-domain $\mathcal{T}$ into the space of diffeomorphisms of the manifold to itself. In other terms, f(t) is a diffeomorphism, for every time t in the domain $\mathcal{T}$.

Measure theoretical definition

A dynamical system may be defined formally, as a measure-preserving transformation of a sigma-algebra, the quadruplet (*X*,Σ,μ,τ). Here, *X* is a set, and Σ is a sigma-algebra on *X*, so that the pair (*X*,Σ) is a measurable space. μ is a finite measure on the sigma-algebra, so that the triplet (*X*,Σ,μ) is a probability space. A map $\tau : X \to X$is said to be Σ-measurable if and only if, for every $\sigma \in \Sigma$, one has $\tau^{-1}\sigma \in \Sigma$. A map τ is said to **preserve the measure** if and only if, for every $\sigma \in \Sigma$, one has $\mu(\tau^{-1}\sigma) = \mu(\sigma)$. Combining the above, a map τ is said to be a **measure-preserving transformation of *X*** , if it is a map from *X* to itself, it is Σ-measurable, and is measure-preserving. The quadruple (*X*,Σ,μ,τ), for such a τ, is then defined to be a **dynamical system**.

The map τ embodies the time evolution of the dynamical system. Thus, for discrete dynamical systems the iterates $\tau^n = \tau \circ \tau \circ \ldots \circ \tau$for integer *n* are studied. For continuous dynamical systems, the map τ is understood to be a finite time evolution map and the construction is more complicated.

Examples of dynamical systems

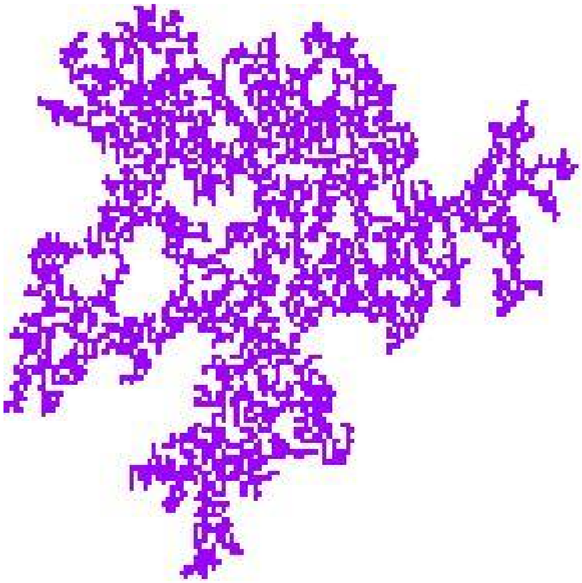
Internal links

- Arnold's cat map
- Baker's map is an example of a chaotic piecewise linear map
- Circle map
- Double pendulum
- Billiards and Outer Billiards
- Henon map
- Horseshoe map
- Irrational rotation
- List of chaotic maps
- Logistic map
- Lorenz system
- Rossler map

Chapter 5

Biological Engineering and Environmental Engineering

Biological engineering



Modeling of the spread of disease using Cellular Automata and Nearest Neighbor Interactions

Biological engineering, **biotechnological engineering** or **bioengineering** (including **biological systems engineering**) is the application of concepts and methods of physics, chemistry, and mathematics to solve problems in life sciences, using engineering's own analytical and synthetic methodologies. In this context, while traditional engineering applies physical and mathematical sciences to analyze, design and manufacture inanimate tools, structures and processes, bioengineering uses the same sciences to study many aspects of living organisms. Usually it is used to analyze and solve problems related to human health.

Biological engineering is a science-based discipline founded upon the biological sciences in the same way that chemical engineering, electrical engineering, and mechanical engineering are based upon chemistry, electricity and magnetism and statics, respectively.

Biological engineering can be differentiated from its roots of pure biology or classical engineering in the following way. Biological studies often follow a reductionist approach in viewing a system on its smallest possible scale which naturally leads toward tools such as functional genomics. Engineering approaches, using classical design perspectives, are constructionist, building new devices, approaches, and technologies from component concepts. Biological engineering utilizes both of these methods in concert relying on reductionist approaches to define the fundamental units which are then commingled to generate something new. Although engineered biological systems have been used to manipulate information, construct materials, process chemicals, produce energy, provide food, and help maintain or enhance human health and our environment, our ability to quickly and reliably engineer biological systems that behave as expected remains less well developed than our mastery over mechanical and electrical systems.

The differentiation between biological engineering and overlap with Biomedical Engineering can be unclear, as many universities now use the terms "bioengineering" and "biomedical engineering" interchangeably . Some contend that Biological Engineering (like biotechnology) has a broader base which spans molecular methods (tends to emphasize the using of biological substances - applying engineering principles to molecular biology, biochemistry, microbiology, pharmacology, protein chemistry, cytology, immunology, neurobiology and neuroscience, cellular and tissue based methods (including devices and sensors), whole organisms (plants, animals), and up increasing length scales to ecosystems. Neither biological engineering nor biomedical engineering is wholly contained within the other, as there are non-biological products for medical needs and biological products for non-medical needs.

ABET , the U.S.-based accreditation board for engineering B.S. programs, makes a distinction between Biomedical Engineering and Biological Engineering; however, the differences are quite small. Biomedical engineers must have life science courses that include human physiology and have experience in performing measurements on living systems while biological engineers must have life science courses (which may or may not include physiology) and experience in making measurements not specifically on living systems. Foundational engineering courses are often the same and include thermodynamics, fluid and mechanical dynamics, kinetics, electronics, and materials properties.

The word bioengineering was coined by British scientist and broadcaster Heinz Wolff in 1954. The term bioengineering is also used to describe the use of vegetation in civil engineering construction. The term bioengineering may also be applied to environmental modifications such as surface soil protection, slope stabilisation, watercourse and shoreline protection, windbreaks, vegetation barriers including noise barriers and visual screens, and the ecological enhancement of an area. The first biological engineering program was created at Mississippi State University in 1967, making it the first biological engineering curriculum in the United States. More recent programs have been launched at MIT and Utah State University .

Biological Engineers or *bioengineers* are engineers who use the principles of biology and the tools of engineering to create usable, tangible products. Biological Engineering employs knowledge and expertise from a number of pure and applied sciences, such as mass and heat transfer, kinetics, biocatalysts, biomechanics, bioinformatics, separation and purification processes, bioreactor design, surface science, fluid mechanics, thermodynamics, and polymer science. It is used in the design of medical devices, diagnostic equipment, biocompatible materials, renewable bioenergy, ecological engineering, and other areas that improve the living standards of societies.

In general, biological engineers attempt to either mimic biological systems in order to create products or modify and control biological systems so that they can replace, augment, or sustain chemical and mechanical processes. Bioengineers can apply their expertise to other applications of engineering and biotechnology, including genetic modification of plants and microorganisms, bioprocess engineering, and biocatalysis.

Because other engineering disciplines also address living organisms (e.g., prosthetics in mechanical engineering), the term biological engineering can be applied more broadly to include agricultural engineering and biotechnology. In fact, many old agricultural engineering departments in universities over the world have rebranded themselves as **agricultural and biological engineering** or **agricultural and biosystems engineering**. Biological engineering is also called bioengineering by some colleges and Biomedical engineering is called Bioengineering by others, and is a rapidly developing field with fluid categorization. The Main Fields of Bioengineering may be categorised as:

- **Bioprocess Engineering**: Bioprocess Design, Biocatalysis, Bioseparation, Bioinformatics, Bioenergy
- **Genetic Engineering**: Synthetic Biology, Horizontal gene transfer.
- **Cellular Engineering**: Cell Engineering, Tissue Culture Engineering, Metabolic engineering.
- **Biomedical Engineering**: Biomedical technology, Biomedical Diagnostics, Biomedical Therapy, Biomechanics, Biomaterials.

Environmental engineering

Environmental engineering is the application of science and engineering principles to improve the environment (air, water, and/or land resources), to provide healthy water, air, and land for human habitation and for other organisms, and to remediate polluted sites.

Environmental engineering involves waste water management and air pollution control, recycling, waste disposal, radiation protection, industrial hygiene, environmental sustainability, and public health issues as well as a knowledge of environmental engineering law. It also includes studies on the environmental impact of proposed construction projects.

Environmental engineers conduct hazardous-waste management studies to evaluate the significance of such hazards, advise on treatment and containment, and develop regulations to prevent mishaps. Environmental engineers also design municipal water supply and industrial wastewater treatment systems as well as address local and worldwide environmental issues such as the effects of acid rain, global warming, ozone depletion, water pollution and air pollution from automobile exhausts and industrial sources. At many universities, Environmental Engineering programs follow either the Department of Civil Engineering or The Department of Chemical Engineering at Engineering faculties. Environmental "civil" engineers focus on hydrology, water resources management, bioremediation, and water treatment plant design. Environmental "chemical" engineers, on the other hand, focus on environmental chemistry, advanced air and water treatment technologies and separation processes.

Additionally, engineers are more frequently obtaining specialized training in law (J.D.) and are utilizing their technical expertise in the practices of Environmental engineering law.. About four percent of environmental engineers go on to obtain Board Certification in their specialty area(s) of environmental engineering (Board Certified Environmental Engineer or BCEE)**.**

Most jurisdictions also impose licensing and registration requirements.

Development of environmental engineering

Ever since people first recognized that their health and well-being were related to the quality of their environment, they have applied thoughtful principles to attempt to improve the quality of their environment. The ancient Harappan civilization utilized early sewers in some cities. The Romans constructed aqueducts to prevent drought and to create a clean, healthful water supply for the metropolis of Rome. In the 15th century, Bavaria created laws restricting the development and degradation of alpine country that constituted the region's water supply

The field emerged as a separate environmental discipline during the middle third of the

20th century in response to widespread public concern about water and pollution and increasingly extensive environmental quality degradation. However, its roots extend back to early efforts in public health engineering. Modern environmental engineering began in London in the mid-19th century when Joseph Bazalgette designed the first major sewerage system that reduced the incidence of waterborne diseases such as cholera. The introduction of drinking water treatment and sewage treatment in industrialized countries reduced waterborne diseases from leading causes of death to rarities.

In many cases, as societies grew, actions that were intended to achieve benefits for those societies had longer-term impacts which reduced other environmental qualities. One example is the widespread application of DDT to control agricultural pests in the years following World War II. While the agricultural benefits were outstanding and crop yields increased dramatically, thus reducing world hunger substantially, and malaria was controlled better than it ever had been, numerous species were brought to the verge of extinction due to the impact of the DDT on their reproductive cycles. The story of DDT as vividly told in Rachel Carson's "Silent Spring" is considered to be the birth of the modern environmental movement and the development of the modern field of "environmental engineering."

Conservation movements and laws restricting public actions that would harm the environment have been developed by various societies for millennia. Notable examples are the laws decreeing the construction of sewers in London and Paris in the 19th century and the creation of the U.S. national park system in the early 20th century.

Scope of Environmental Engineering

Briefly speaking, the main task of environmental engineers is to protect public health by protecting (from further degradation), preserving (the present condition of), and enhancing the environment.

Environmental engineering is the application of science and engineering principles to the environment. Some consider environmental engineering to include the development of sustainable processes. There are several divisions of the field of environmental engineering.

Environmental impact assessment and mitigation

In this division, engineers and scientists use a systemic identification and evaluation process to assess the potential impacts of a proposed project , plans, programs, policies, or legislative actions upon the physical-chemical, biological, cultural, and socioeconomic components on environmental conditions. They apply scientific and engineering principles to evaluate if there are likely to be any adverse impacts to water quality, air quality, habitat quality, flora and fauna, agricultural capacity, traffic impacts, social impacts, ecological impacts, noise impacts, visual(landscape) impacts, etc. If impacts are expected, they then develop mitigation measures to limit or prevent such impacts. An example of a mitigation measure would be the creation of wetlands in a nearby location to mitigate the filling in of wetlands necessary for a road development if it is not possible to reroute the road.

The practice of environmental assessment was intitiated on January 1, 1970, the effective date of the National Environmental Policy Act (NEPA) in the United States. Since that time, more than 100 developing and developed nations either have planned specific analogous laws or have adopted procedure used elsewhere. NEPA is applicable to all federal agencies in the United States.

Water supply and treatment

Engineers and scientists work to secure water supplies for potable and agricultural use. They evaluate the water balance within a watershed and determine the available water supply, the water needed for various needs in that watershed, the seasonal cycles of water movement through the watershed and they develop systems to store, treat, and convey water for various uses. Water is treated to achieve water quality objectives for the end uses. In the case of potable water supply, water is treated to minimize the risk of infectious disease transmission, the risk of non-infectious illness, and to create a palatable water flavor. Water distribution systems are designed and built to provide adequate water pressure and flow rates to meet various end-user needs such as domestic use, fire suppression, and irrigation.

Wastewater conveyance and treatment

Water pollution

Most urban and many rural areas no longer discharge human waste directly to the land through outhouse, septic, and/or honey bucket systems, but rather deposit such waste into water and convey it from households via sewer systems. Engineers and scientists develop collection and treatment systems to carry this waste material away from where people live and produce the waste and discharge it into the environment. In developed countries, substantial resources are applied to the treatment and detoxification of this waste before it is discharged into a river, lake, or ocean system. Developing nations are striving to obtain the resources to develop such systems so that they can improve water quality in their surface waters and reduce the risk of water-borne infectious disease.

Sewage treatment plant, Australia.

There are numerous wastewater treatment technologies. A wastewater treatment train can consist of a primary clarifier system to remove solid and floating materials, a secondary treatment system consisting of an aeration basin followed by flocculation and sedimentation or an activated sludge system and a secondary clarifier, a tertiary biological nitrogen removal system, and a final disinfection process. The aeration basin/activated sludge system removes organic material by growing bacteria (activated sludge). The secondary clarifier removes the activated sludge from the water. The tertiary system, although not always included due to costs, is becoming more prevalent to remove nitrogen and phosphorus and to disinfect the water before discharge to a surface water stream or ocean outfall.

Air quality management

Engineers apply scientific and engineering principles to the design of manufacturing and combustion processes to reduce air pollutant emissions to acceptable levels. Scrubbers, electrostatic precipitators, catalytic converters, and various other processes are utilized to remove particulate matter, nitrogen oxides, sulfur oxides, volatile organic compounds (VOC), reactive organic gases (ROG) and other air pollutants from flue gases and other sources prior to allowing their emission to the atmosphere.

Scientists have developed air pollution dispersion models to evaluate the concentration of a pollutant at a receptor or the impact on overall air quality from vehicle exhausts and industrial flue gas stack emissions.

To some extent, this field overlaps the desire to decrease carbon dioxide and other greenhouse gas emissions from combustion processes**.**

Other applications

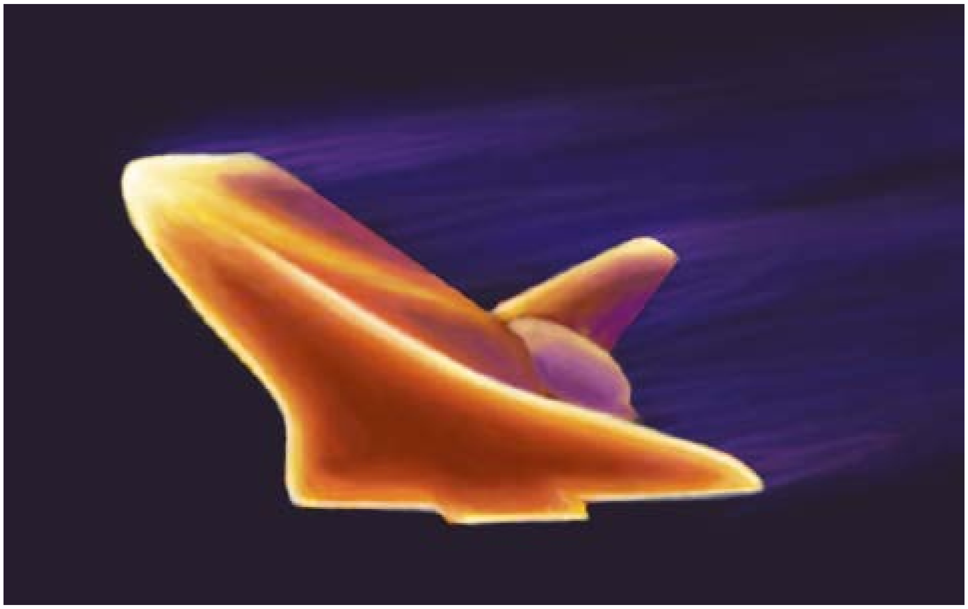
- Environmental policy and regulation development
- Contaminated land management and site remediation
- Environment, Health and Safety
- Hazardous waste management
- Natural resource management
- Noise pollution
- Risk assessment
- Solid waste management

Prominent environmental engineers

- Robert A. Gearheart
- Paul V. Roberts
- Abel Wolman

Chapter 6

Materials Science



Simulation of the outside of the Space Shuttle as it heats up to over 1,500 °C (2,730 °F) during re-entry into the Earth's atmosphere

Materials science is an interdisciplinary field applying the properties of matter to various areas of science and engineering. This scientific field investigates the relationship between the structure of materials at atomic or molecular scales and their macroscopic properties. It incorporates elements of applied physics and chemistry. With significant media attention focused on nanoscience and nanotechnology in recent years, materials science has been propelled to the forefront at many universities. It is also an important part of forensic engineering and failure analysis. Materials science also deals with *fundamental properties* and *characteristics* of materials.

History

The material of choice of a given era is often its defining point. Phrases such as Stone Age, Bronze Age, and Steel Age are good examples. Originally deriving from the manufacture of ceramics and its putative derivative metallurgy, materials science is one of the oldest forms of engineering and applied science. Modern materials science evolved directly from metallurgy, which itself evolved from mining and (likely) ceramics and the use of fire. A major breakthrough in the understanding of materials occurred in the late 19th century, when the American scientist Josiah Willard Gibbs demonstrated that the thermodynamic properties related to atomic structure in various phases are related to the physical properties of a material. Important elements of modern materials science are a product of the space race: the understanding and engineering of the metallic alloys, and silica and carbon materials, used in the construction of space vehicles enabling the exploration of space. Materials science has driven, and been driven by, the development of revolutionary technologies such as plastics, semiconductors, and biomaterials.

Before the 1960s (and in some cases decades after), many *materials science* departments were named *metallurgy* departments, from a 19th and early 20th century emphasis on metals. The field has since broadened to include every class of materials, including: ceramics, polymers, semiconductors, magnetic materials, medical implant materials and biological materials (materiomics).

Fundamentals

The basis of materials science involves relating the desired properties and relative performance of a material in a certain application to the structure of the atoms and phases in that material through characterization. The major determinants of the structure of a material and thus of its properties are its constituent chemical elements and the way in which it has been processed into its final form. These characteristics, taken together and related through the laws of thermodynamics, govern a material’s microstructure, and thus its properties.

The manufacture of a perfect crystal of a material is currently physically impossible. Instead materials scientists manipulate the defects in crystalline materials such as precipitates, grain boundaries (Hall-Petch relationship), interstitial atoms, vacancies or substitutional atoms, to create materials with the desired properties.

Not all materials have a regular crystal structure. Polymers display varying degrees of crystallinity, and many are completely non-crystalline. Glasses, some ceramics, and many natural materials are amorphous, not possessing any long-range order in their atomic arrangements. The study of polymers combines elements of chemical and statistical thermodynamics to give thermodynamic, as well as mechanical, descriptions of physical properties.

In addition to industrial interest, materials science has gradually developed into a field which provides tests for condensed matter or solid state theories. New physics emerge because of the diverse new material properties which need to be explained.

Materials in industry

Radical materials advances can drive the creation of new products or even new industries, but stable industries also employ materials scientists to make incremental improvements and troubleshoot issues with currently used materials. Industrial applications of materials science include materials design, cost-benefit tradeoffs in industrial production of materials, processing techniques (casting, rolling, welding, ion implantation, crystal growth, thin-film deposition, sintering, glassblowing, etc.), and analytical techniques (characterization techniques such as electron microscopy, x-ray diffraction, calorimetry, nuclear microscopy (HEFIB), Rutherford backscattering, neutron diffraction, small-angle X-ray scattering (SAXS), etc.

Besides material characterization, the material scientist/engineer also deals with the extraction of materials and their conversion into useful forms. Thus ingot casting, foundry techniques, blast furnace extraction, and electrolytic extraction are all part of the required knowledge of a metallurgist/engineer. Often the presence, absence or variation of minute quantities of secondary elements and compounds in a bulk material will have a great impact on the final properties of the materials produced, for instance, steels are classified based on 1/10 and 1/100 weight percentages of the carbon and other alloying elements they contain. Thus, the extraction and purification techniques employed in the extraction of iron in the blast furnace will have an impact of the quality of steel that may be produced.

The overlap between physics and materials science has led to the offshoot field of *materials physics*, which is concerned with the physical properties of materials. The approach is generally more macroscopic and applied than in condensed matter physics.

Metal alloys

The study of metal alloys is a significant part of materials science. Of all the metallic alloys in use today, the alloys of iron (steel, stainless steel, cast iron, tool steel, alloy steels) make up the largest proportion both by quantity and commercial value. Iron alloyed with various proportions of carbon gives low, mid and high carbon steels. For the steels, the hardness and tensile strength of the steel is directly related to the amount of carbon present, with increasing carbon levels also leading to lower ductility and toughness. The addition of silicon and graphitization will produce cast irons (although some cast irons are made precisely with no graphitization). The addition of chromium, nickel and molybdenum to carbon steels (more than 10%) gives us stainless steels.

Other significant metallic alloys are those of aluminium, titanium, copper and magnesium. Copper alloys have been known for a long time (since the Bronze Age), while the alloys of the other three metals have been relatively recently developed. Due to the chemical reactivity of these metals, the electrolytic extraction processes required were only developed relatively recently. The alloys of aluminium, titanium and magnesium are also known and valued for their high strength-to-weight ratios and, in the case of magnesium, their ability to provide electromagnetic shielding. These materials are ideal for situations where high strength-to-weight ratios are more important than bulk cost, such as in the aerospace industry and certain automotive engineering applications.

Polymers

Polymers are also an important part of materials science. Polymers are the raw materials (the resins) used to make what we commonly call plastics. Plastics are really the final product, created after one or more polymers or additives have been added to a resin during processing, which is then shaped into a final form. Polymers which have been around, and which are in current widespread use, include polyethylene, polypropylene, PVC, polystyrene, nylons, polyesters, acrylics, polyurethanes, and polycarbonates. Plastics are generally classified as "commodity", "specialty" and "engineering" plastics.

PVC (polyvinyl-chloride) is widely used, inexpensive, and annual production quantities are large. It lends itself to an incredible array of applications, from artificial leather to electrical insulation and cabling, packaging and containers. Its fabrication and processing are simple and well-established. The versatility of PVC is due to the wide range of plasticisers and other additives that it accepts. The term "additives" in polymer science refers to the chemicals and compounds added to the polymer base to modify its material properties.

Polycarbonate would be normally considered an engineering plastic (other examples include PEEK, ABS). Engineering plastics are valued for their superior strengths and other special material properties. They are usually not used for disposable applications, unlike commodity plastics.

Specialty plastics are materials with unique characteristics, such as ultra-high strength, electrical conductivity, electro-fluorescence, high thermal stability, etc.

The dividing lines between the various types of plastics is not based on material but rather on their properties and applications. For instance, polyethylene (PE) is a cheap, low friction polymer commonly used to make disposable shopping bags and trash bags, and is considered a commodity plastic, whereas Medium-Density Polyethylene MDPE is used for underground gas and water pipes, and another variety called Ultra-high Molecular Weight Polyethylene UHMWPE is an engineering plastic which is used extensively as the glide rails for industrial equipment and the low-friction socket in implanted hip joints.

Ceramics and glasses

Composite materials

Another application of material science in industry is the making of composite materials. Composite materials are structured materials composed of two or more macroscopic phases. Applications range from structural elements such as steel-reinforced concrete, to the thermally insulative tiles which play a key and integral role in NASA's Space Shuttle thermal protection system which is used to protect the surface of the shuttle from the heat of re-entry into the Earth's atmosphere. One example is Reinforced Carbon-Carbon (RCC), The light gray material which withstands re-entry temperatures up to 1510 °C (2750 °F) and protects the Space Shuttle's wing leading edges and nose cap. RCC is a laminated composite material made from graphite rayon cloth and impregnated with a phenolic resin. After curing at high temperature in an autoclave, the laminate is pyrolized to convert the resin to carbon, impregnated with furfural alcohol in a vacuum chamber, and cured/pyrolized to convert the furfural alcohol to carbon. In order to provide oxidation resistance for reuse capability, the outer layers of the RCC are converted to silicon carbide.

Other examples can be seen in the "plastic" casings of television sets, cell-phones and so on. These plastic casings are usually a composite material made up of a thermoplastic matrix such as acrylonitrile-butadiene-styrene (ABS) in which calcium carbonate chalk, talc, glass fibres or carbon fibres have been added for added strength, bulk, or electro-static dispersion. These additions may be referred to as reinforcing fibres, or dispersants, depending on their purpose.

Classes of materials

Materials science encompasses various classes of materials, each of which may constitute a separate field. Materials are sometimes classified by the type of bonding present between the atoms:

1. Ionic crystals
2. Covalent crystals
3. Metals
4. Intermetallics
5. Semiconductors
6. Polymers
7. Composite materials
8. Vitreous materials

Overview

- Nanotechnology – rigorously, the study of materials where the effects of quantum confinement, the Gibbs-Thomson effect, or any other effect only present at the nanoscale is the defining property of the material; but more commonly, it is the creation and study of materials whose defining structural properties are anywhere from less than a nanometer to one hundred nanometers in scale, such as molecularly engineered materials.

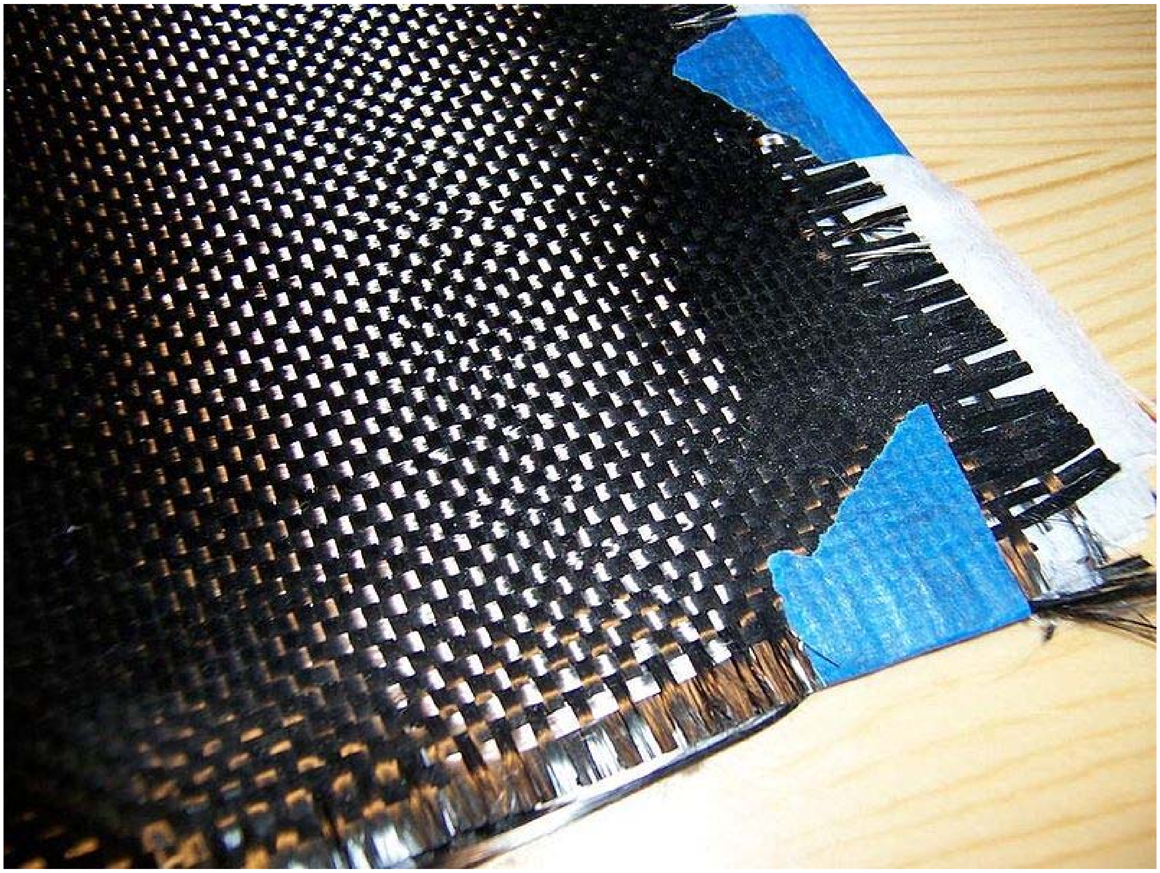
- Microtechnology - study of materials and processes and their interaction, allowing microfabrication of structures of micrometric dimensions, such as MicroElectroMechanical Systems (MEMS).
- Crystallography – the study of how atoms in a solid fill space, the defects associated with crystal structures such as grain boundaries and dislocations, and the characterization of these structures and their relation to physical properties.
- Materials Characterization – such as diffraction with x-rays, electrons, or neutrons, and various forms of spectroscopy and chemical analysis such as Raman spectroscopy, energy-dispersive spectroscopy (EDS), chromatography, thermal analysis, electron microscope analysis, etc., in order to understand and define the properties of materials.

Si_3N_4 ceramic bearing parts

- Metallurgy – the study of metals and their alloys, including their extraction, microstructure and processing.
- Biomaterials – materials that are derived from and/or used with biological systems.

- Electronic and magnetic materials – materials such as semiconductors used to create integrated circuits, storage media, sensors, and other devices.
- Tribology – the study of the wear of materials due to friction and other factors.
- Surface science/Catalysis – interactions and structures between solid-gas solid-liquid or solid-solid interfaces.
- Ceramography – the study of the microstructures of high-temperature materials and refractories, including structural ceramics such as RCC, polycrystalline silicon carbide and transformation toughened ceramics

Some practitioners often consider rheology a sub-field of materials science, because it can cover any material that flows. However, modern rheology typically deals with non-Newtonian fluid dynamics, so it is often considered a sub-field of continuum mechanics.



A cloth of woven carbon fiber filaments is commonly used for reinforcement in composite materials.

- Glass Science – any non-crystalline material including inorganic glasses, vitreous metals and non-oxide glasses.
- Forensic engineering – the study of how products fail, and the vital role of the materials of construction
- Forensic materials engineering – the study of material failure, and the light it sheds on how engineers specify materials in their product

- Textile Reinforced Materials - materials in the form of ceramic or concrete are reinforced with a primarily woven or non-woven textile structure to impose high strength with comparatively more flexibility to withstand vibrations and sudden jerks.

Primary topics

- Thermodynamics, statistical mechanics, and physical chemistry, for phase equilibrium conditions, phase diagrams of materials systems (multi-phase, multi-component, reacting and non-reacting systems)
- Phase transformation kinetics, for the kinetics of phase transformations (with particular emphasis on solid-solid phase transitions)
- Transport phenomena for the transport of heat, mass, and momentum in materials processing.
- Crystallography, quantum chemistry or quantum physics, for the structure (symmetry and defects) and bonding in materials (e.g., ionic, metallic, covalent, and van der Waals bonding)
- Mechanical behavior of materials, to understand the mechanical properties of materials, defects and their propagation, and their behavior under static, dynamic, and cyclic loads
- Electronic properties of materials, and solid-state physics, for the understanding of the electronic, thermal, magnetic, and optical properties of materials
- Diffraction and wave mechanics, for the science behind characterization systems, e.g., transmission electron microscopy (TEM)

Household items made of various kinds of plastic.

- Polymer properties, synthesis, and characterization, for a specialized understanding of how polymers behave, how they are made, and how they are characterized; exciting applications of polymers include liquid crystal displays (LCDs, the displays found in most cell-phones, cameras, and iPods), novel photovoltaic devices based on semiconductor polymers (which, unlike the traditional silicon solar panels, are flexible and cheap to manufacture, albeit with lower efficiency), and membranes for room-temperature fuel cells (as proton exchange membranes) and filtration systems in the environmental and biomedical fields
- Biomaterials, physiology, biomechanics, biochemistry, for a specialized understanding of how materials integrate into biological systems, e.g., through materiomics
- Semiconductor materials and semiconductor devices, for a specialized understanding of the advanced processes used in industry (e.g. crystal growth techniques, thin-film deposition, ion implantation, photolithography), their properties, and their integration in electronic devices
- Alloying, corrosion, and thermal or mechanical processing, for a specialized treatment of metallurgical materials—with applications ranging from aerospace and industrial equipment to the civil industries

Professional organizations

- Materials Research Society, MRS
- European Materials Research Society, EMRS
- ASM International
- The Minerals, Metals, & Materials Society, TMS
- Materials Australia
- American Ceramic Society, ACerS
- NACE International
- The American Institute of Mining, Metallurgical, and Petroleum Engineers, AIME
- Society for the Advancement of Material and Process Engineering, SAMPE
- The Institute of Materials, Minerals and Mining, IOM^3
- Alpha Sigma Mu, ΑΣΜ
- Central European Institute of Technology, CEITEC

Chapter 7

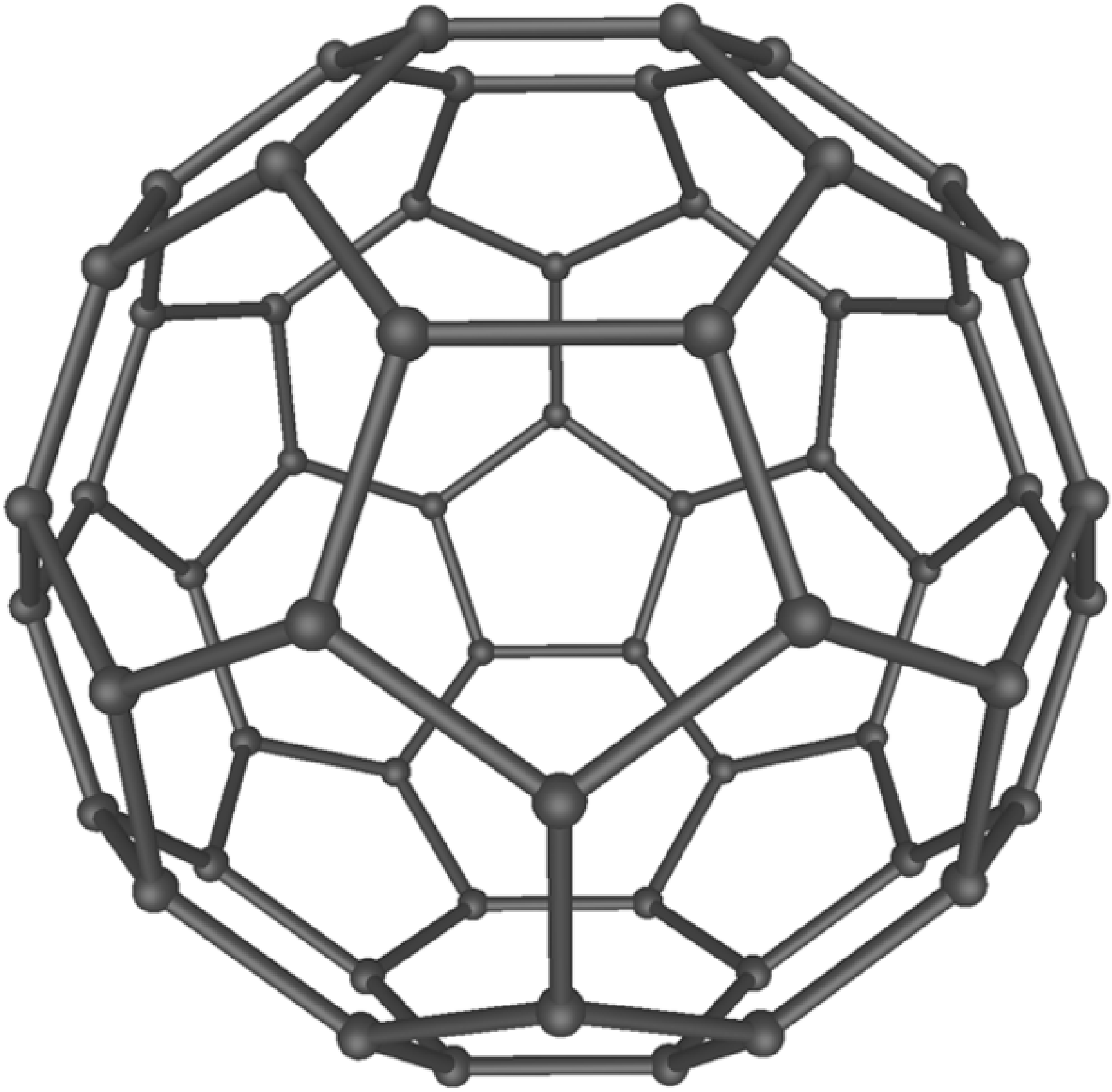
Nanotechnology

Nanotechnology (sometimes shortened to "**nanotech**") is the study of manipulating matter on an atomic and molecular scale. Generally, nanotechnology deals with structures sized between 1 to 100 nanometre in at least one dimension, and involves developing materials or devices possessing at least one dimension within that size. Quantum mechanical effects are very important at this scale, which is in the quantum realm.

Nanotechnology is very diverse, ranging from extensions of conventional device physics to completely new approaches based upon molecular self-assembly, from developing new materials with dimensions on the nanoscale to investigating whether we can directly control matter on the atomic scale.

There is much debate on the future implications of nanotechnology. Nanotechnology may be able to create many new materials and devices with a vast range of applications, such as in medicine, electronics, biomaterials and energy production. On the other hand, nanotechnology raises many of the same issues as any new technology, including concerns about the toxicity and environmental impact of nanomaterials, and their potential effects on global economics, as well as speculation about various doomsday scenarios. These concerns have led to a debate among advocacy groups and governments on whether special regulation of nanotechnology is warranted.

Origins



Buckminsterfullerene C_{60}, also known as the buckyball, is a representative member of the carbon structures known as fullerenes. Members of the fullerene family are a major subject of research falling under the nanotechnology umbrella.

The first use of the concepts found in 'nano-technology' (but pre-dating use of that name) was in "There's Plenty of Room at the Bottom", a talk given by physicist Richard Feynman at an American Physical Society meeting at California Institute of Technology (Caltech) on December 29, 1959. Feynman described a process by which the ability to manipulate individual atoms and molecules might be developed, using one set of precise tools to build and operate another proportionally smaller set, and so on down to the needed scale. In the course of this, he noted, scaling issues would arise from the changing magnitude of various physical phenomena: gravity would become less important, surface tension and van der Waals attraction would become increasingly more significant, etc. This basic idea appeared plausible, and exponential assembly enhances it with parallelism to produce a useful quantity of end products. The term "nanotechnology" was defined by Tokyo University of Science Professor Norio Taniguchi in a 1974 paper as follows: "'Nano-technology' mainly consists of the processing of, separation, consolidation, and deformation of materials by one atom or by one molecule." In the 1980s the basic idea of this definition was explored in much more depth by Dr. K. Eric Drexler, who promoted the technological significance of nano-scale phenomena and devices through speeches and the books *Engines of Creation: The Coming Era of Nanotechnology* (1986) and *Nanosystems: Molecular Machinery, Manufacturing, and Computation*, and so the term acquired its current sense. *Engines of Creation* is considered the first book on the topic of nanotechnology. Nanotechnology and nanoscience got started in the early 1980s with two major developments; the birth of cluster science and the invention of the scanning tunneling microscope (STM). This development led to the discovery of fullerenes in 1985 and carbon nanotubes a few years later. In another development, the synthesis and properties of semiconductor nanocrystals was studied; this led to a fast increasing number of metal and metal oxide nanoparticles and quantum dots. The atomic force microscope (AFM or SFM) was invented six years after the STM was invented. In 2000, the United States National Nanotechnology Initiative was founded to coordinate Federal nanotechnology research and development and is evaluated by the President's Council of Advisors on Science and Technology.

Fundamental concepts

Nanotechnology is the engineering of functional systems at the molecular scale. This covers both current work and concepts that are more advanced. In its original sense, nanotechnology refers to the projected ability to construct items from the bottom up, using techniques and tools being developed today to make complete, high performance products.

One nanometer (nm) is one billionth, or 10^{-9}, of a meter. By comparison, typical carbon-carbon bond lengths, or the spacing between these atoms in a molecule, are in the range 0.12–0.15 nm, and a DNA double-helix has a diameter around 2 nm. On the other hand, the smallest cellular life-forms, the bacteria of the genus Mycoplasma, are around 200 nm in length. By convention, nanotechnology is taken as the scale range 1 to 100 nm following the definition used by the National Nanotechnology Initiative in the US. The lower limit is set by the size of atoms (hydrogen has the smallest atoms, which are approximately a quarter of a nm diameter) since nanotechnology must build its devices from atoms and molecules. The upper limit is more or less arbitrary but is around the size that phenomena not observed in larger structures start to become apparent and can be made use of in the nano device. These new phenomena make nanotechnology distinct from devices which are merely miniaturised versions of an equivalent macroscopic device; such devices are on a larger scale and come under the description of microtechnology.

To put that scale in another context, the comparative size of a nanometer to a meter is the same as that of a marble to the size of the earth. Or another way of putting it: a nanometer is the amount an average man's beard grows in the time it takes him to raise the razor to his face.

Two main approaches are used in nanotechnology. In the "bottom-up" approach, materials and devices are built from molecular components which assemble themselves chemically by principles of molecular recognition. In the "top-down" approach, nano-objects are constructed from larger entities without atomic-level control.

Areas of physics such as nanoelectronics, nanomechanics, nanophotonics and nanoionics have evolved during the last few decades to provide a basic scientific foundation of nanotechnology.

Larger to smaller: a materials perspective

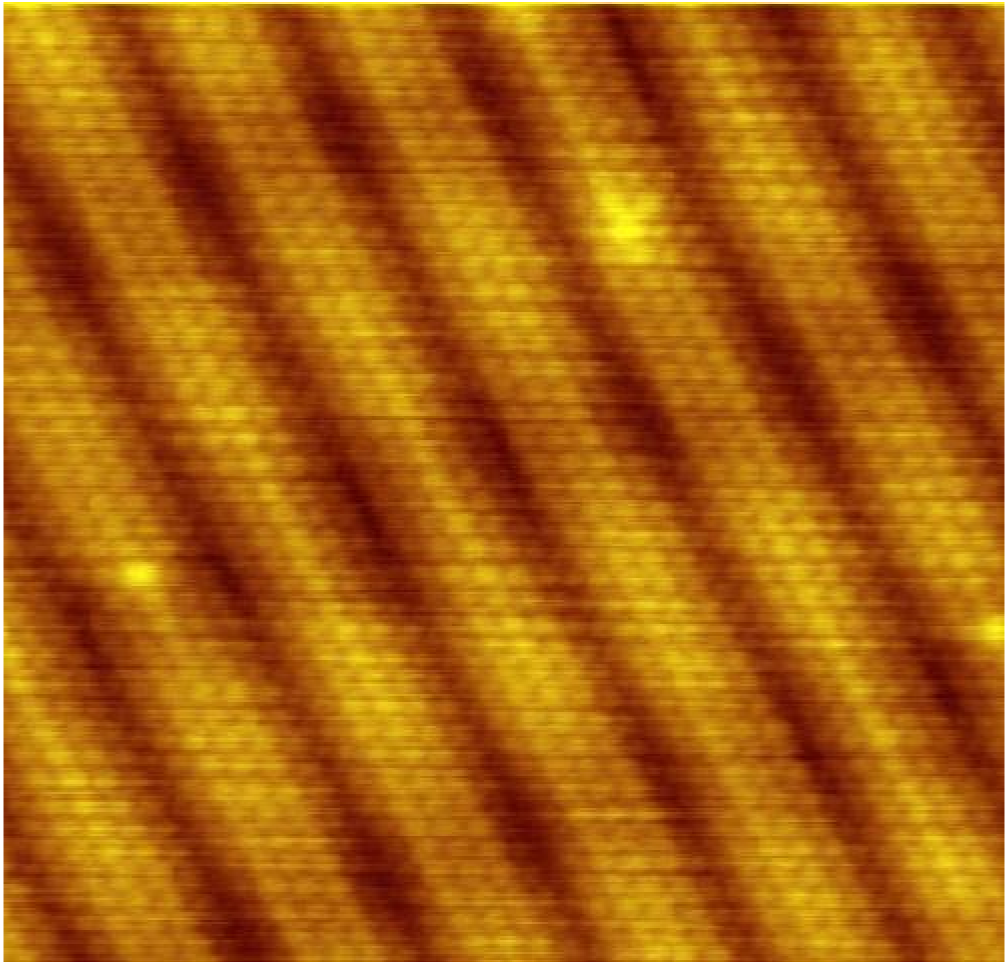


Image of reconstruction on a clean Gold(100) surface, as visualized using scanning tunneling microscopy. The positions of the individual atoms composing the surface are visible.

A number of physical phenomena become pronounced as the size of the system decreases. These include statistical mechanical effects, as well as quantum mechanical effects, for example the “quantum size effect” where the electronic properties of solids are altered with great reductions in particle size. This effect does not come into play by going from macro to micro dimensions. However, quantum effects become dominant when the nanometer size range is reached, typically at distances of 100 nanometers or less, the so called quantum realm. Additionally, a number of physical (mechanical, electrical, optical, etc.) properties change when compared to macroscopic systems. One example is the increase in surface area to volume ratio altering mechanical, thermal and catalytic properties of materials. Diffusion and reactions at nanoscale, nanostructures materials and nanodevices with fast ion transport are generally referred to nanoionics. *Mechanical* properties of nanosystems are of interest in the nanomechanics research. The catalytic activity of nanomaterials also opens potential risks in their interaction with biomaterials.

Materials reduced to the nanoscale can show different properties compared to what they exhibit on a macroscale, enabling unique applications. For instance, opaque substances become transparent (copper); stable materials turn combustible (aluminum); insoluble materials become soluble (gold). A material such as gold, which is chemically inert at normal scales, can serve as a potent chemical catalyst at nanoscales. Much of the fascination with nanotechnology stems from these quantum and surface phenomena that matter exhibits at the nanoscale.

Simple to complex: a molecular perspective

Modern synthetic chemistry has reached the point where it is possible to prepare small molecules to almost any structure. These methods are used today to manufacture a wide variety of useful chemicals such as pharmaceuticals or commercial polymers. This ability raises the question of extending this kind of control to the next-larger level, seeking methods to assemble these single molecules into supramolecular assemblies consisting of many molecules arranged in a well defined manner.

These approaches utilize the concepts of molecular self-assembly and/or supramolecular chemistry to automatically arrange themselves into some useful conformation through a bottom-up approach. The concept of molecular recognition is especially important: molecules can be designed so that a specific configuration or arrangement is favored due to non-covalent intermolecular forces. The Watson–Crick basepairing rules are a direct result of this, as is the specificity of an enzyme being targeted to a single substrate, or the specific folding of the protein itself. Thus, two or more components can be designed to be complementary and mutually attractive so that they make a more complex and useful whole.

Such bottom-up approaches should be capable of producing devices in parallel and be much cheaper than top-down methods, but could potentially be overwhelmed as the size and complexity of the desired assembly increases. Most useful structures require complex and thermodynamically unlikely arrangements of atoms. Nevertheless, there are many examples of self-assembly based on molecular recognition in biology, most notably Watson–Crick basepairing and enzyme-substrate interactions. The challenge for nanotechnology is whether these principles can be used to engineer new constructs in addition to natural ones.

Molecular nanotechnology: a long-term view

Molecular nanotechnology, sometimes called molecular manufacturing, describes engineered nanosystems (nanoscale machines) operating on the molecular scale. Molecular nanotechnology is especially associated with the molecular assembler, a machine that can produce a desired structure or device atom-by-atom using the principles of mechanosynthesis. Manufacturing in the context of productive nanosystems is not related to, and should be clearly distinguished from, the conventional technologies used to manufacture nanomaterials such as carbon nanotubes and nanoparticles.

When the term "nanotechnology" was independently coined and popularized by Eric Drexler (who at the time was unaware of an earlier usage by Norio Taniguchi) it referred to a future manufacturing technology based on molecular machine systems. The premise was that molecular scale biological analogies of traditional machine components demonstrated molecular machines were possible: by the countless examples found in biology, it is known that sophisticated, stochastically optimised biological machines can be produced.

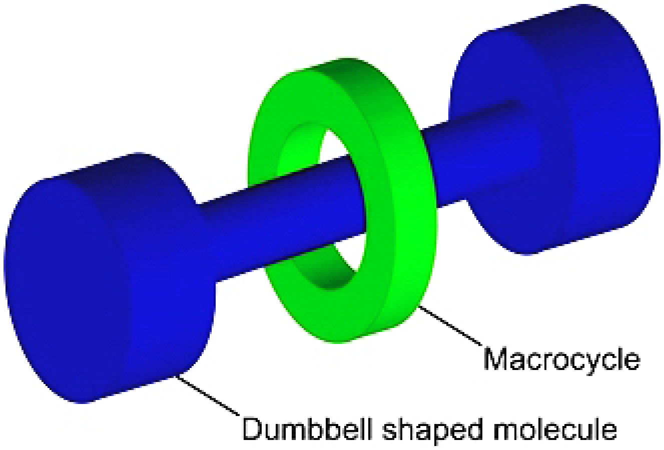
It is hoped that developments in nanotechnology will make possible their construction by some other means, perhaps using biomimetic principles. However, Drexler and other researchers have proposed that advanced nanotechnology, although perhaps initially implemented by biomimetic means, ultimately could be based on mechanical engineering principles, namely, a manufacturing technology based on the mechanical functionality of these components (such as gears, bearings, motors, and structural members) that would enable programmable, positional assembly to atomic specification. The physics and engineering performance of exemplar designs were analyzed in Drexler's book *Nanosystems*.

In general it is very difficult to assemble devices on the atomic scale, as all one has to position atoms on other atoms of comparable size and stickiness. Another view, put forth by Carlo Montemagno, is that future nanosystems will be hybrids of silicon technology and biological molecular machines. Yet another view, put forward by the late Richard Smalley, is that mechanosynthesis is impossible due to the difficulties in mechanically manipulating individual molecules.

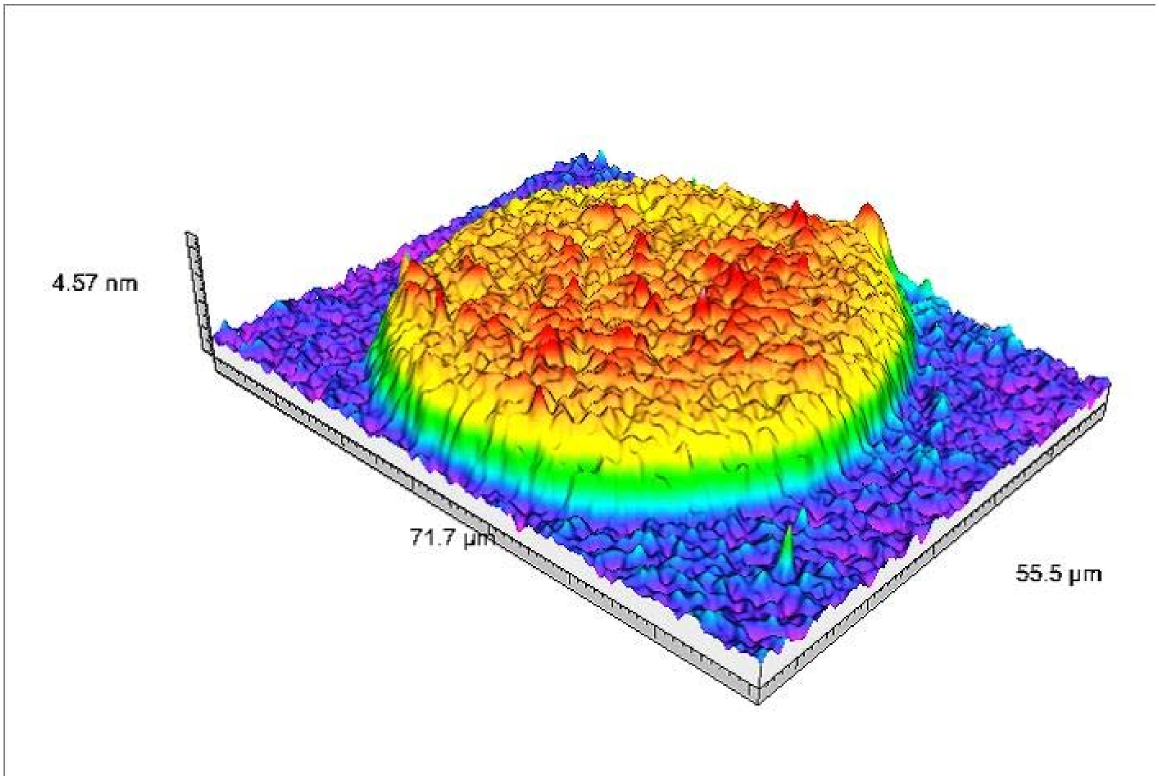
This led to an exchange of letters in the ACS publication Chemical & Engineering News in 2003. Though biology clearly demonstrates that molecular machine systems are possible, non-biological molecular machines are today only in their infancy. Leaders in research on non-biological molecular machines are Dr. Alex Zettl and his colleagues at Lawrence Berkeley Laboratories and UC Berkeley. They have constructed at least three distinct molecular devices whose motion is controlled from the desktop with changing voltage: a nanotube nanomotor, a molecular actuator, and a nanoelectromechanical relaxation oscillator.

An experiment indicating that positional molecular assembly is possible was performed by Ho and Lee at Cornell University in 1999. They used a scanning tunneling microscope to move an individual carbon monoxide molecule (CO) to an individual iron atom (Fe) sitting on a flat silver crystal, and chemically bound the CO to the Fe by applying a voltage.

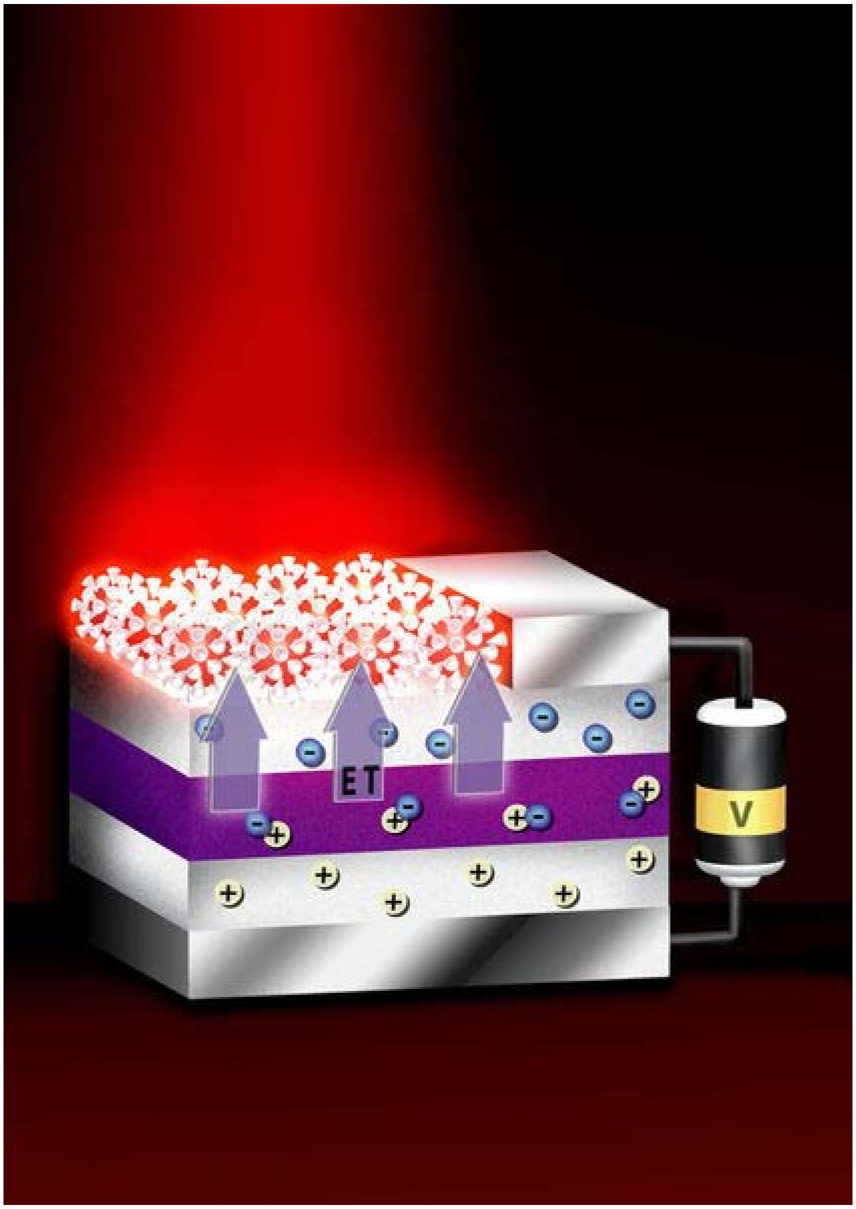
Current research



Graphical representation of a rotaxane, useful as a molecular switch.



Sarfus image of a DNA biochip elaborated by bottom-up approach.



This device transfers energy from nano-thin layers of quantum wells to nanocrystals above them, causing the nanocrystals to emit visible light.

Nanomaterials

The nanomaterials field includes subfields which develop or study materials having unique properties arising from their nanoscale dimensions.

- Interface and colloid science has given rise to many materials which may be useful in nanotechnology, such as carbon nanotubes and other fullerenes, and various nanoparticles and nanorods. Nanomaterials with fast ion transport are related also to nanoionics and nanoelectronics.

- Nanoscale materials can also be used for bulk applications; most present commercial applications of nanotechnology are of this flavor.
- Progress has been made in using these materials for medical applications.
- Nanoscale materials are sometimes used in solar cells which combats the cost of traditional Silicon solar cells
- Development of applications incorporating semiconductor nanoparticles to be used in the next generation of products, such as display technology, lighting, solar cells and biological imaging.

Bottom-up approaches

These seek to arrange smaller components into more complex assemblies.

- DNA nanotechnology utilizes the specificity of Watson–Crick basepairing to construct well-defined structures out of DNA and other nucleic acids.
- Approaches from the field of "classical" chemical synthesis also aim at designing molecules with well-defined shape (e.g. bis-peptides).
- More generally, molecular self-assembly seeks to use concepts of supramolecular chemistry, and molecular recognition in particular, to cause single-molecule components to automatically arrange themselves into some useful conformation.
- Atomic force microscope tips can be used as a nanoscale "write head" to deposit a chemical upon a surface in a desired pattern in a process called dip pen nanolithography. This technique fits into the larger subfield of nanolithography.

Top-down approaches

These seek to create smaller devices by using larger ones to direct their assembly.

- Many technologies that descended from conventional solid-state silicon methods for fabricating microprocessors are now capable of creating features smaller than 100 nm, falling under the definition of nanotechnology. Giant magnetoresistance-based hard drives already on the market fit this description, as do atomic layer deposition (ALD) techniques. Peter Grünberg and Albert Fert received the Nobel Prize in Physics in 2007 for their discovery of Giant magnetoresistance and contributions to the field of spintronics.
- Solid-state techniques can also be used to create devices known as nanoelectromechanical systems or NEMS, which are related to microelectromechanical systems or MEMS.
- Focused ion beams can directly remove material, or even deposit material when suitable pre-cursor gasses are applied at the same time. For example, this technique is used routinely to create sub-100 nm sections of material for analysis in Transmission electron microscopy.
- Atomic force microscope tips can be used as a nanoscale "write head" to deposit a resist, which is then followed by a etching process to remove material in a top-down method.

Functional approaches

These seek to develop components of a desired functionality without regard to how they might be assembled.

- Molecular electronics seeks to develop molecules with useful electronic properties. These could then be used as single-molecule components in a nanoelectronic device.
- Synthetic chemical methods can also be used to create synthetic molecular motors, such as in a so-called nanocar.

Biomimetic approaches

- Bionics or biomimicry seeks to apply biological methods and systems found in nature, to the study and design of engineering systems and modern technology. Biomineralization is one example of the systems studied.

- Bionanotechnology the use of biomolecules for applications in nanotechnology, including use of viruses.

Speculative

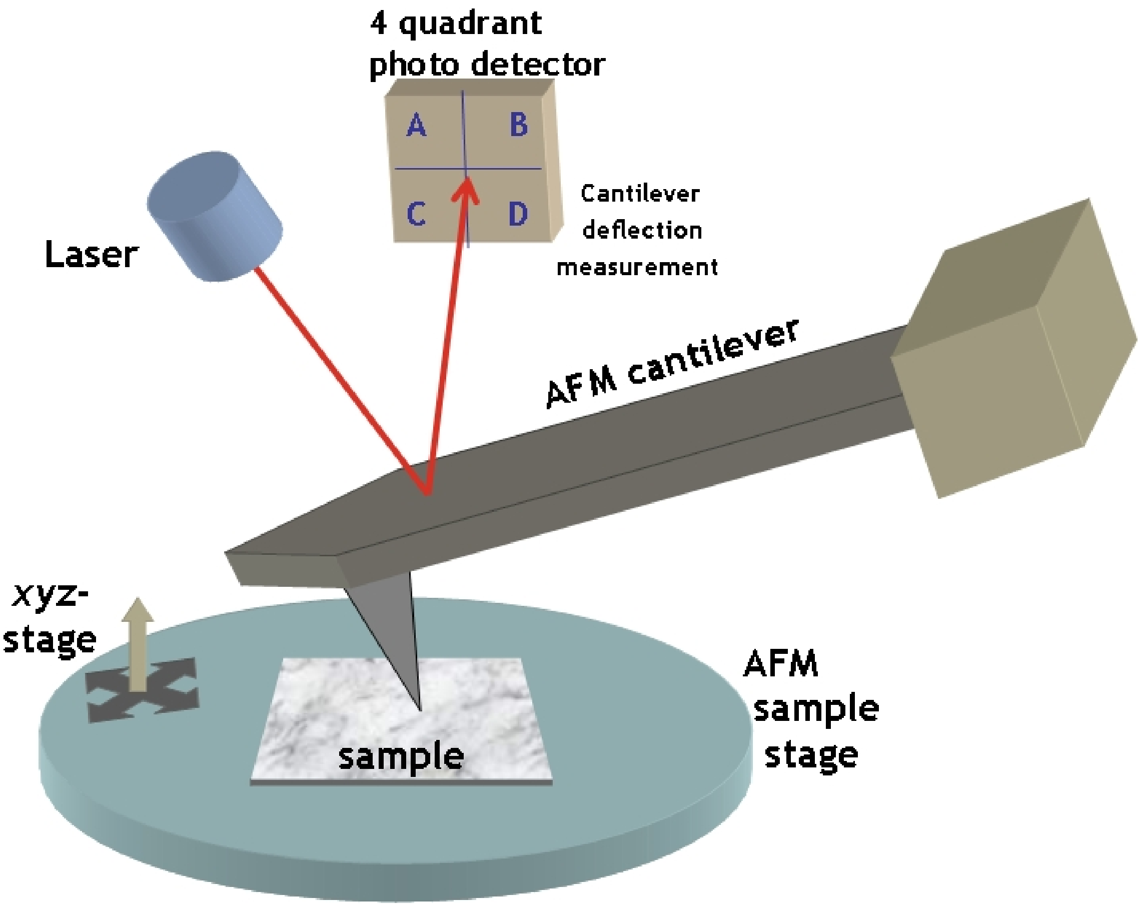
These subfields seek to anticipate what inventions nanotechnology might yield, or attempt to propose an agenda along which inquiry might progress. These often take a big-picture view of nanotechnology, with more emphasis on its societal implications than the details of how such inventions could actually be created.

- Molecular nanotechnology is a proposed approach which involves manipulating single molecules in finely controlled, deterministic ways. This is more theoretical than the other subfields and is beyond current capabilities.
- Nanorobotics centers on self-sufficient machines of some functionality operating at the nanoscale. There are hopes for applying nanorobots in medicine, but it may not be easy to do such a thing because of several drawbacks of such devices. Nevertheless, progress on innovative materials and methodologies has been demonstrated with some patents granted about new nanomanufacturing devices for future commercial applications, which also progressively helps in the development towards nanorobots with the use of embedded nanobioelectronics concepts.
- Productive nanosystems are "systems of nanosystems" which will be complex nanosystems that produce atomically precise parts for other nanosystems, not necessarily using novel nanoscale-emergent properties, but well-understood fundamentals of manufacturing. Because of the discrete (i.e. atomic) nature of matter and the possibility of exponential growth, this stage is seen as the basis of another industrial revolution. Mihail Roco, one of the architects of the USA's National Nanotechnology Initiative, has proposed four states of nanotechnology that seem to parallel the technical progress of the Industrial Revolution,

progressing from passive nanostructures to active nanodevices to complex nanomachines and ultimately to productive nanosystems.

- Programmable matter seeks to design materials whose properties can be easily, reversibly and externally controlled though a fusion of information science and materials science.
- Due to the popularity and media exposure of the term nanotechnology, the words picotechnology and femtotechnology have been coined in analogy to it, although these are only used rarely and informally.

Tools and techniques



Typical AFM setup. A microfabricated cantilever with a sharp tip is deflected by features on a sample surface, much like in a phonograph but on a much smaller scale. A laser beam reflects off the backside of the cantilever into a set of photodetectors, allowing the deflection to be measured and assembled into an image of the surface.

There are several important modern developments. The atomic force microscope (AFM) and the Scanning Tunneling Microscope (STM) are two early versions of scanning probes that launched nanotechnology. There are other types of scanning probe microscopy, all flowing from the ideas of the scanning confocal microscope developed by Marvin Minsky in 1961 and the scanning acoustic microscope (SAM) developed by

Calvin Quate and coworkers in the 1970s, that made it possible to see structures at the nanoscale. The tip of a scanning probe can also be used to manipulate nanostructures (a process called positional assembly). Feature-oriented scanning-positioning methodology suggested by Rostislav Lapshin appears to be a promising way to implement these nanomanipulations in automatic mode. However, this is still a slow process because of low scanning velocity of the microscope. Various techniques of nanolithography such as optical lithography, X-ray lithography dip pen nanolithography, electron beam lithography or nanoimprint lithography were also developed. Lithography is a top-down fabrication technique where a bulk material is reduced in size to nanoscale pattern.

Another group of nanotechnological techniques include those used for fabrication of nanowires, those used in semiconductor fabrication such as deep ultraviolet lithography, electron beam lithography, focused ion beam machining, nanoimprint lithography, atomic layer deposition, and molecular vapor deposition, and further including molecular self-assembly techniques such as those employing di-block copolymers. However, all of these techniques preceded the nanotech era, and are extensions in the development of scientific advancements rather than techniques which were devised with the sole purpose of creating nanotechnology and which were results of nanotechnology research.

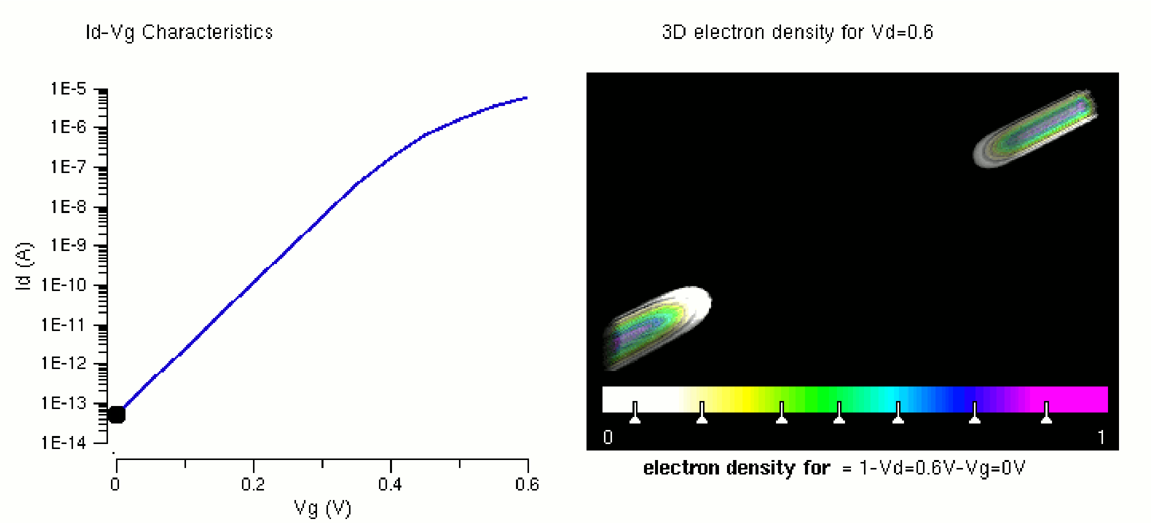
The top-down approach anticipates nanodevices that must be built piece by piece in stages, much as manufactured items are made. Scanning probe microscopy is an important technique both for characterization and synthesis of nanomaterials. Atomic force microscopes and scanning tunneling microscopes can be used to look at surfaces and to move atoms around. By designing different tips for these microscopes, they can be used for carving out structures on surfaces and to help guide self-assembling structures. By using, for example, feature-oriented scanning-positioning approach, atoms can be moved around on a surface with scanning probe microscopy techniques. At present, it is expensive and time-consuming for mass production but very suitable for laboratory experimentation.

In contrast, bottom-up techniques build or grow larger structures atom by atom or molecule by molecule. These techniques include chemical synthesis, self-assembly and positional assembly. Dual polarisation interferometry is one tool suitable for characterisation of self assembled thin films. Another variation of the bottom-up approach is molecular beam epitaxy or MBE. Researchers at Bell Telephone Laboratories like John R. Arthur. Alfred Y. Cho, and Art C. Gossard developed and implemented MBE as a research tool in the late 1960s and 1970s. Samples made by MBE were key to the discovery of the fractional quantum Hall effect for which the 1998 Nobel Prize in Physics was awarded. MBE allows scientists to lay down atomically precise layers of atoms and, in the process, build up complex structures. Important for research on semiconductors, MBE is also widely used to make samples and devices for the newly emerging field of spintronics.

However, new therapeutic products, based on responsive nanomaterials, such as the ultradeformable, stress-sensitive Transfersome vesicles, are under development and already approved for human use in some countries.

Applications

As of August 21, 2008, the Project on Emerging Nanotechnologies estimates that over 800 manufacturer-identified nanotech products are publicly available, with new ones hitting the market at a pace of 3–4 per week. The project lists all of the products in a publicly accessible online. Most applications are limited to the use of "first generation" passive nanomaterials which includes titanium dioxide in sunscreen, cosmetics and some food products; Carbon allotropes used to produce gecko tape; silver in food packaging, clothing, disinfectants and household appliances; zinc oxide in sunscreens and cosmetics, surface coatings, paints and outdoor furniture varnishes; and cerium oxide as a fuel catalyst.



One of the major applications of nanotechnology is in the area of nanoelectronics with MOSFET's being made of small nanowires ~10 nm in length. Here is a simulation of such a nanowire.

The National Science Foundation (a major distributor for nanotechnology research in the United States) funded researcher David Berube to study the field of nanotechnology. His findings are published in the monograph Nano-Hype: The Truth Behind the Nanotechnology Buzz. This study concludes that much of what is sold as “nanotechnology” is in fact a recasting of straightforward materials science, which is leading to a “nanotech industry built solely on selling nanotubes, nanowires, and the like” which will “end up with a few suppliers selling low margin products in huge volumes." Further applications which require actual manipulation or arrangement of nanoscale components await further research. Though technologies branded with the term 'nano' are sometimes little related to and fall far short of the most ambitious and transformative technological goals of the sort in molecular manufacturing proposals, the term still connotes such ideas. According to Berube, there may be a danger that a "nano bubble" will form, or is forming already, from the use of the term by scientists and entrepreneurs to garner funding, regardless of interest in the transformative possibilities of more ambitious and far-sighted work.

Implications

Because of the far-ranging claims that have been made about potential applications of nanotechnology, a number of serious concerns have been raised about what effects these will have on our society if realized, and what action if any is appropriate to mitigate these risks.

There are possible dangers that arise with the development of nanotechnology. The Center for Responsible Nanotechnology suggests that new developments could result, among other things, in untraceable weapons of mass destruction, networked cameras for use by the government, and weapons developments fast enough to destabilize arms races ("Nanotechnology Basics").

One area of concern is the effect that industrial-scale manufacturing and use of nanomaterials would have on human health and the environment, as suggested by nanotoxicology research. Groups such as the Center for Responsible Nanotechnology have advocated that nanotechnology should be specially regulated by governments for these reasons. Others counter that overregulation would stifle scientific research and the development of innovations which could greatly benefit mankind.

Other experts, including director of the Woodrow Wilson Center's Project on Emerging Nanotechnologies David Rejeski, have testified that successful commercialization depends on adequate oversight, risk research strategy, and public engagement. Berkeley, California is currently the only city in the United States to regulate nanotechnology; Cambridge, Massachusetts in 2008 considered enacting a similar law, but ultimately rejected this.

Health and environmental concerns

Some of the recently developed nanoparticle products may have unintended consequences. Researchers have discovered that silver nanoparticles used in socks only to reduce foot odor are being released in the wash with possible negative consequences. Silver nanoparticles, which are bacteriostatic, may then destroy beneficial bacteria which are important for breaking down organic matter in waste treatment plants or farms.

A study at the University of Rochester found that when rats breathed in nanoparticles, the particles settled in the brain and lungs, which led to significant increases in biomarkers for inflammation and stress response. A study in China indicated that nanoparticles induce skin aging through oxidative stress in hairless mice.

A two-year study at UCLA's School of Public Health found lab mice consuming nano-titanium dioxide showed DNA and chromosome damage to a degree "linked to all the big killers of man, namely cancer, heart disease, neurological disease and aging".

A major study published more recently in Nature Nanotechnology suggests some forms of carbon nanotubes – a poster child for the “nanotechnology revolution” – could be as harmful as asbestos if inhaled in sufficient quantities. Anthony Seaton of the Institute of Occupational Medicine in Edinburgh, Scotland, who contributed to the article on carbon nanotubes said "We know that some of them probably have the potential to cause mesothelioma. So those sorts of materials need to be handled very carefully." In the absence of specific nano-regulation forthcoming from governments, Paull and Lyons (2008) have called for an exclusion of engineered nanoparticles from organic food. A newspaper article reports that workers in a paint factory developed serious lung disease and nanoparticles were found in their lungs.

Regulation

Calls for tighter regulation of nanotechnology have occurred alongside a growing debate related to the human health and safety risks associated with nanotechnology. Furthermore, there is significant debate about who is responsible for the regulation of nanotechnology. While some non-nanotechnology specific regulatory agencies currently cover some products and processes (to varying degrees) – by "bolting on" nanotechnology to existing regulations – there are clear gaps in these regimes. In "Nanotechnology Oversight: An Agenda for the Next Administration," former EPA deputy administrator J. Clarence (Terry) Davies lays out a clear regulatory roadmap for the next presidential administration and describes the immediate and longer term steps necessary to deal with the current shortcomings of nanotechnology oversight.

Stakeholders concerned by the lack of a regulatory framework to assess and control risks associated with the release of nanoparticles and nanotubes have drawn parallels with bovine spongiform encephalopathy ('mad cow's disease), thalidomide, genetically modified food, nuclear energy, reproductive technologies, biotechnology, and asbestosis. Dr. Andrew Maynard, chief science advisor to the Woodrow Wilson Center's Project on Emerging Nanotechnologies, concludes (among others) that there is insufficient funding for human health and safety research, and as a result there is currently limited understanding of the human health and safety risks associated with nanotechnology. As a result, some academics have called for stricter application of the precautionary principle, with delayed marketing approval, enhanced labelling and additional safety data development requirements in relation to certain forms of nanotechnology.

The Royal Society report identified a risk of nanoparticles or nanotubes being released during disposal, destruction and recycling, and recommended that "manufacturers of products that fall under extended producer responsibility regimes such as end-of-life regulations publish procedures outlining how these materials will be managed to minimize possible human and environmental exposure" (p.xiii). Reflecting the challenges for ensuring responsible life cycle regulation, the Institute for Food and Agricultural Standards has proposed standards for nanotechnology research and development should be integrated across consumer, worker and environmental standards. They also propose that NGOs and other citizen groups play a meaningful role in the development of these standards.

Chapter 8

Optics

Optics includes study of dispersion of light

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

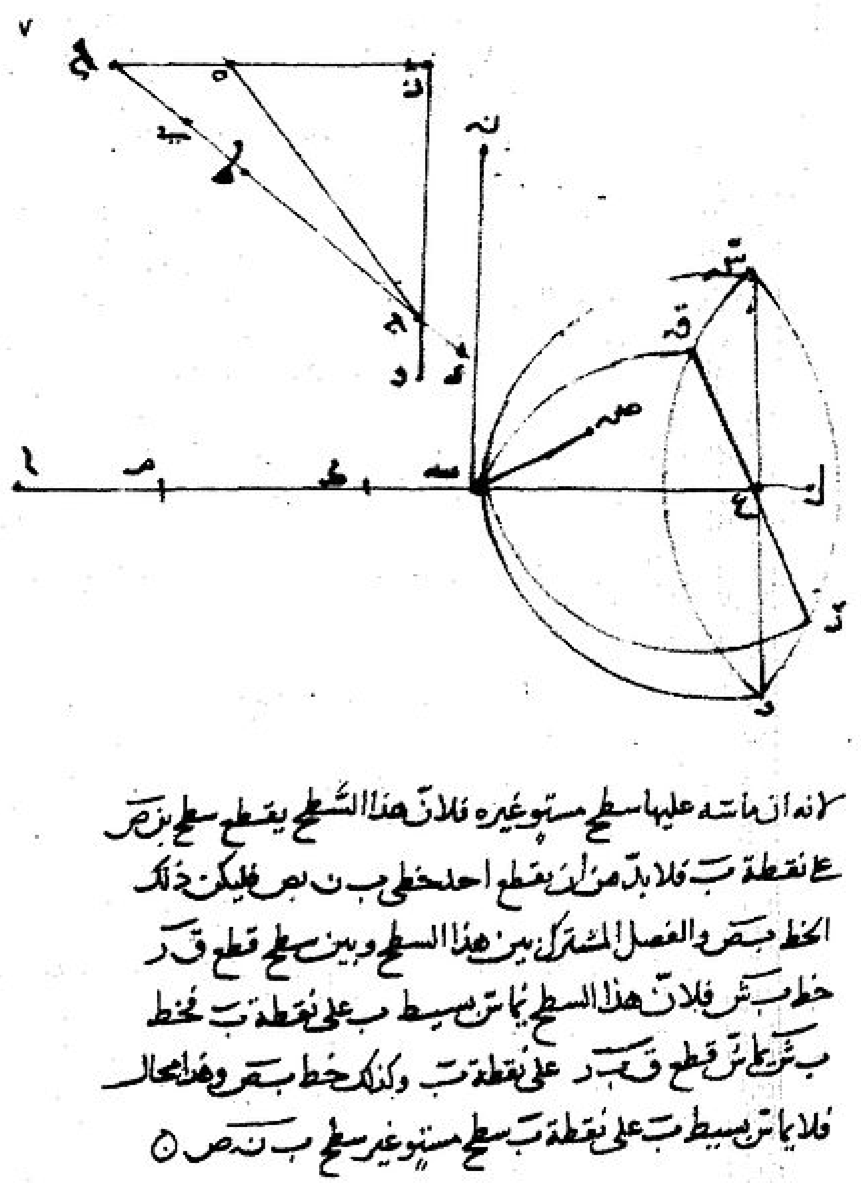
Most optical phenomena can be accounted for using the classical electromagnetic description of light. Complete electromagnetic descriptions of light are, however, often difficult to apply in practice. Practical optics is usually done using simplified models. The most common of these, geometric optics,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

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

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

History

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



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

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

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

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

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

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

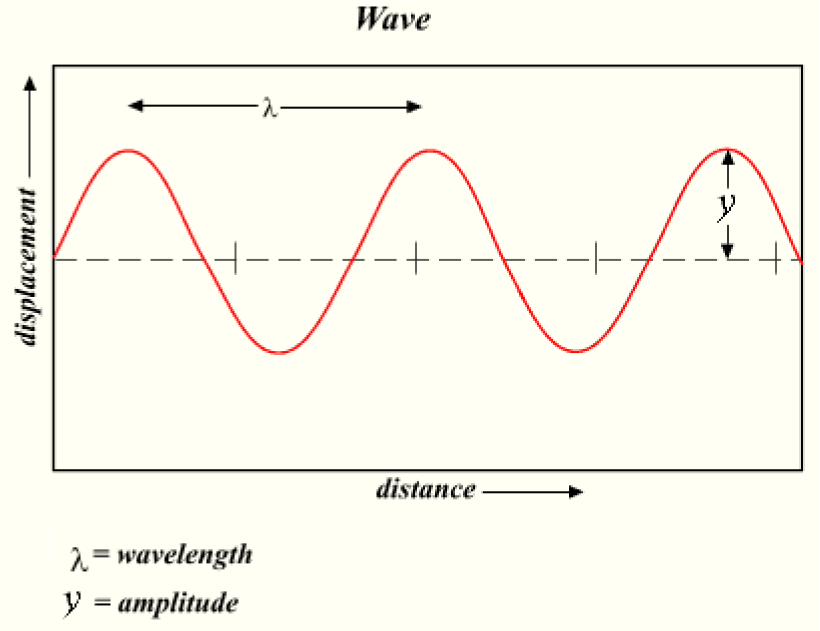
Optical theory progressed in the mid-17th century with treatises written by philosopher René Descartes, which explained a variety of optical phenomena including reflection and refraction by assuming that light was emitted by objects which produced it. This differed substantively from the ancient Greek 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:

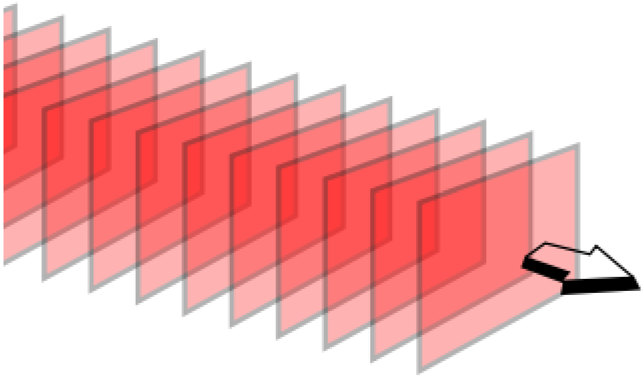
$$n = c / v.$$

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

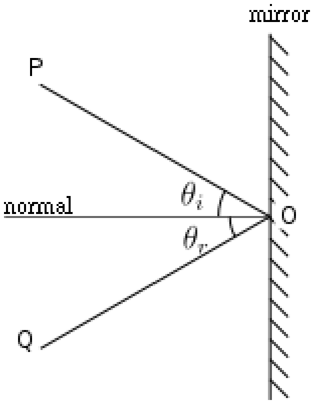


Diagram of specular reflection

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

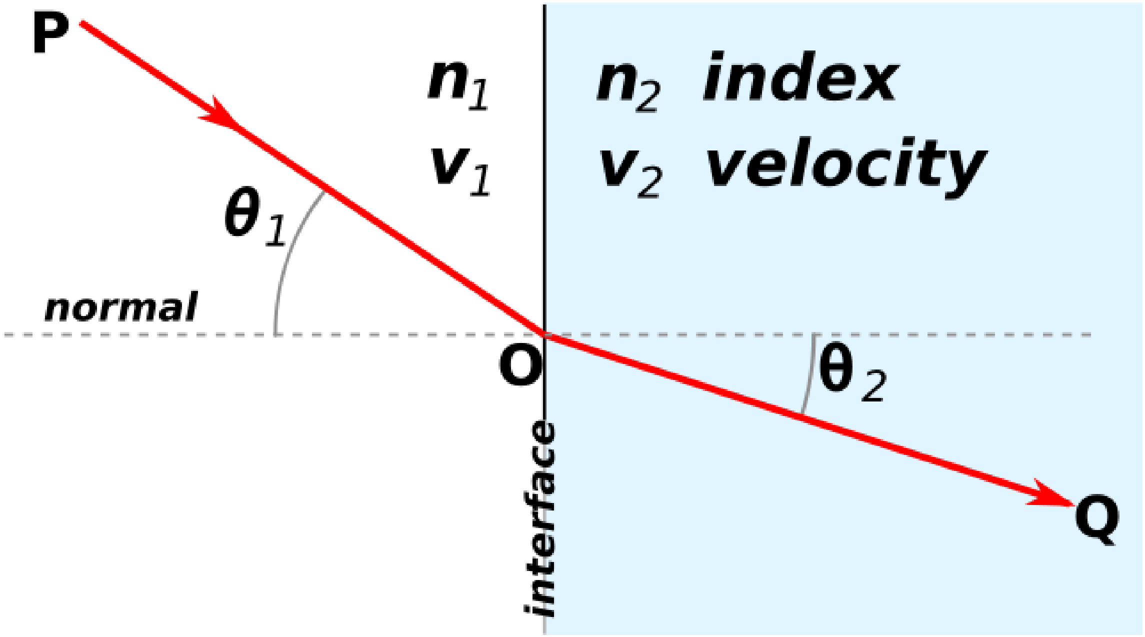


Illustration of Snell's Law for the case $n_1 < n_2$, such as air/water interface

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:

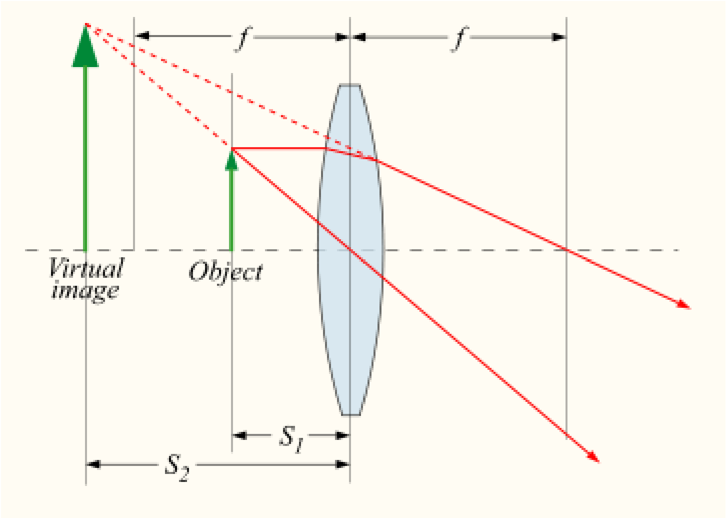
$$v_1 \sin\theta_2 \ = v_2 \sin\theta_1$$

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.

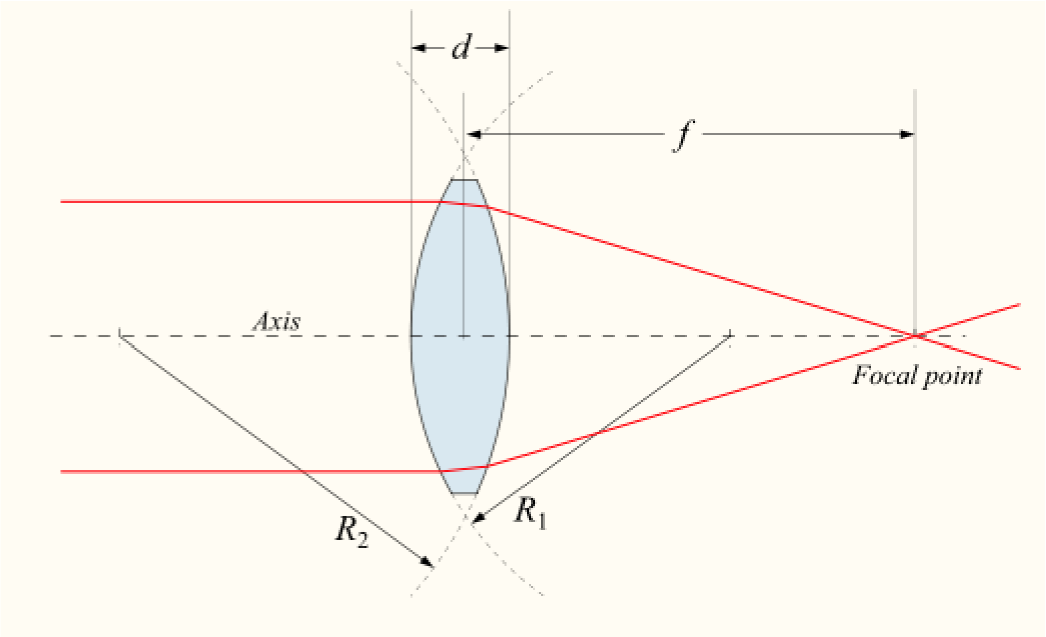


A ray tracing diagram for a converging lens.

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.

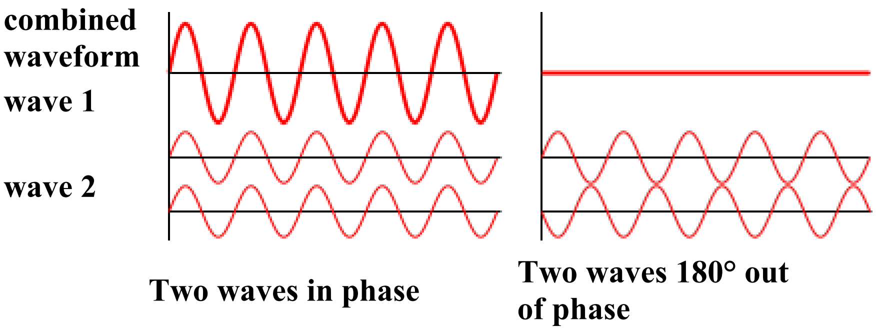
Complex models based on physical optics can account for the propagation of any wavefront through an optical system, including predicting the wavelength, amplitude, and phase of the wave. Additionally, all of the results from geometrical optics can be recovered using the techniques of Fourier optics which 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. See below for an illustration of this effect.



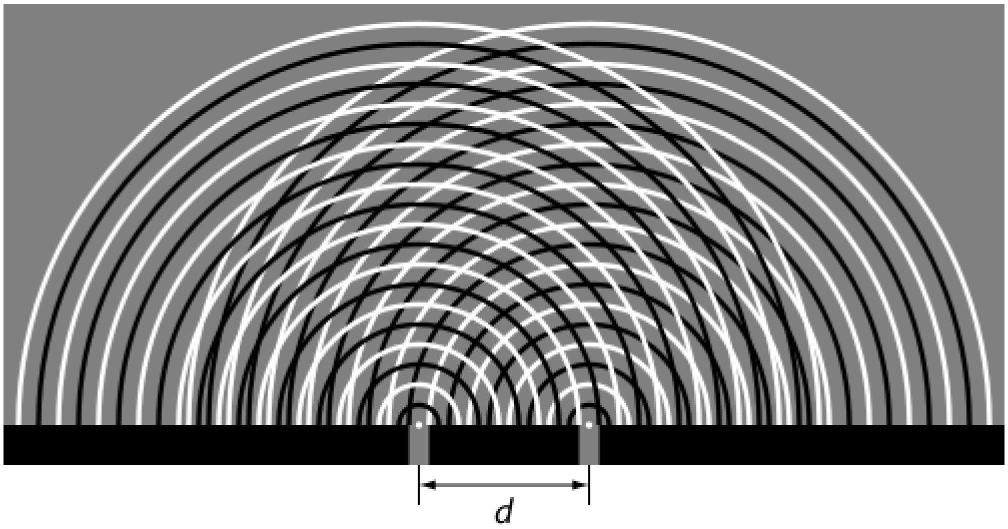
When oil or fuel is spilled, colorful patterns are formed by thin-film interference.

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.

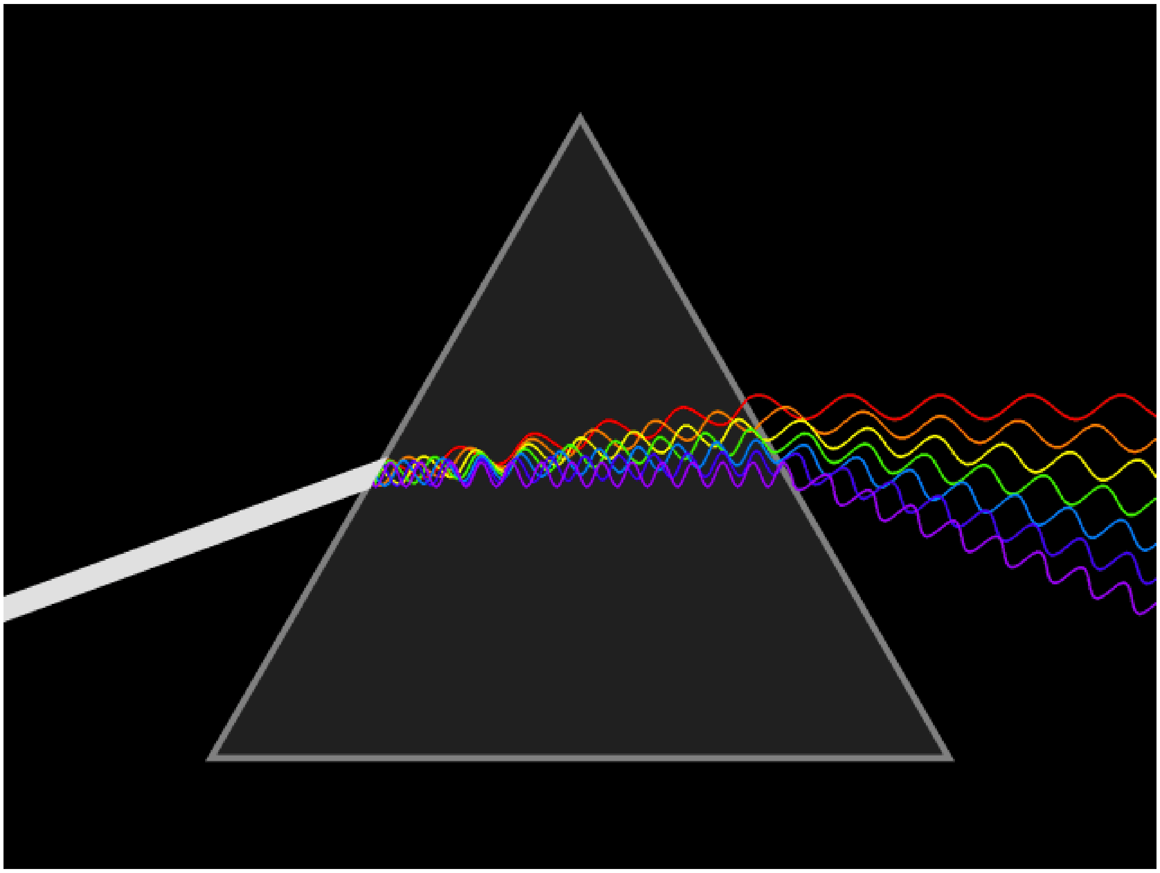
Diffraction effects limit the ability for an optical detector to optically resolve separate light sources. In general, light that is passing through an aperture will experience diffraction and the best images that can be created (as described in diffraction-limited optics) appear as a central spot with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

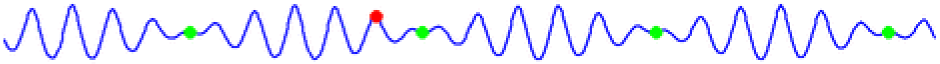


High frequency (blue) light is deflected the most, and low frequency (red) the least.

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

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

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



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

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

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

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

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

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

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

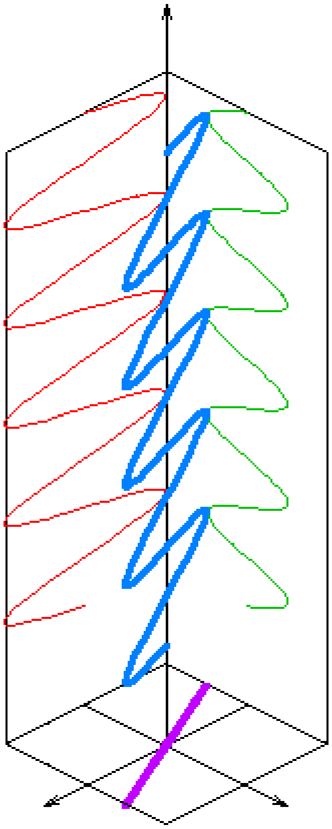
If *D* is less than zero, the medium is said to have *positive dispersion* or normal dispersion. If *D* is greater than zero, the medium has *negative dispersion*. If a light pulse is propagated through a normally dispersive medium, the result is the higher frequency components slow down more than the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

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

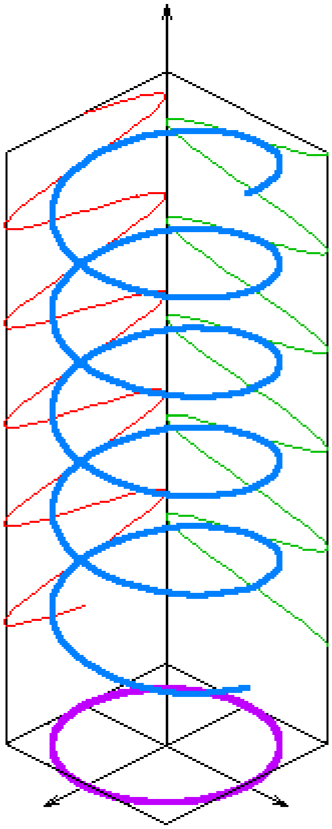
Polarization

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

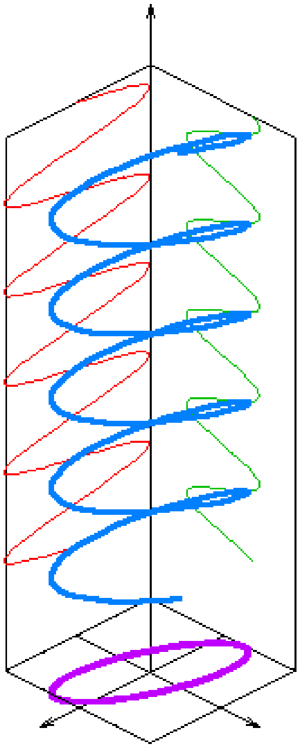
The typical way to consider polarization is to keep track of the orientation of the electric field vector as the electromagnetic wave propagates. The electric field vector of a plane wave may be arbitrarily divided into two perpendicular components labeled *x* and *y* (with **z** indicating the direction of travel). The shape traced out in the x-y plane by the electric field vector is a Lissajous figure that describes the *polarization state*. The following figures show some examples of the evolution of the electric field vector (blue), with time (the vertical axes), at a particular point in space, along with its *x* and *y* components (red/left and green/right), and the path traced by the vector in the plane (purple): The same evolution would occur when looking at the electric field at a particular time while evolving the point in space, along the direction opposite to propagation.



Linear



Circular



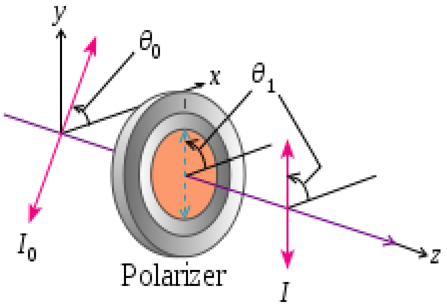
Elliptical polarization

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

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

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

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



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

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

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

where

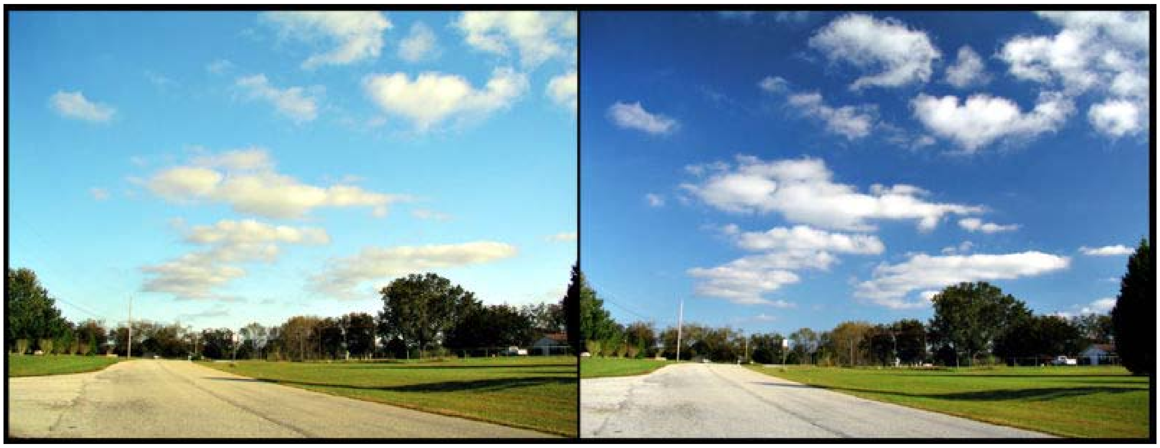
I_0 is the initial intensity,
and θ_i is the angle between the light's initial polarization direction and the axis of the polarizer.

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

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

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

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



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

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

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

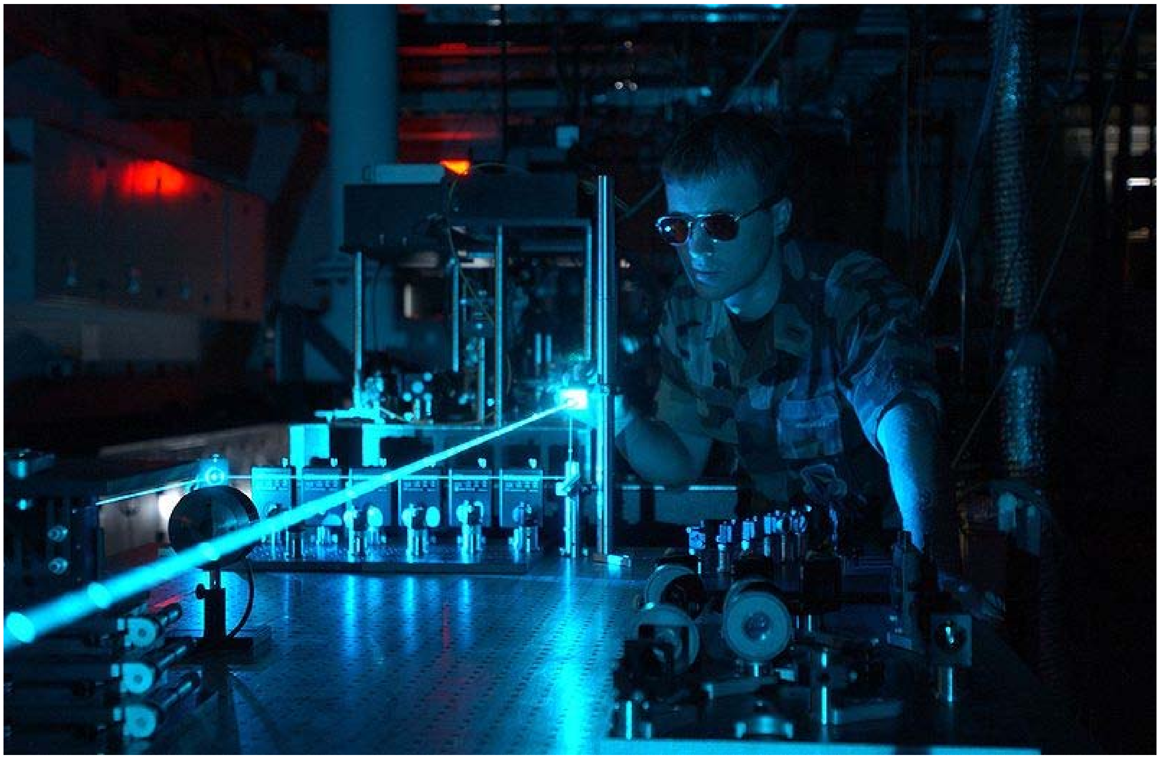
Modern optics

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

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

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

Lasers



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

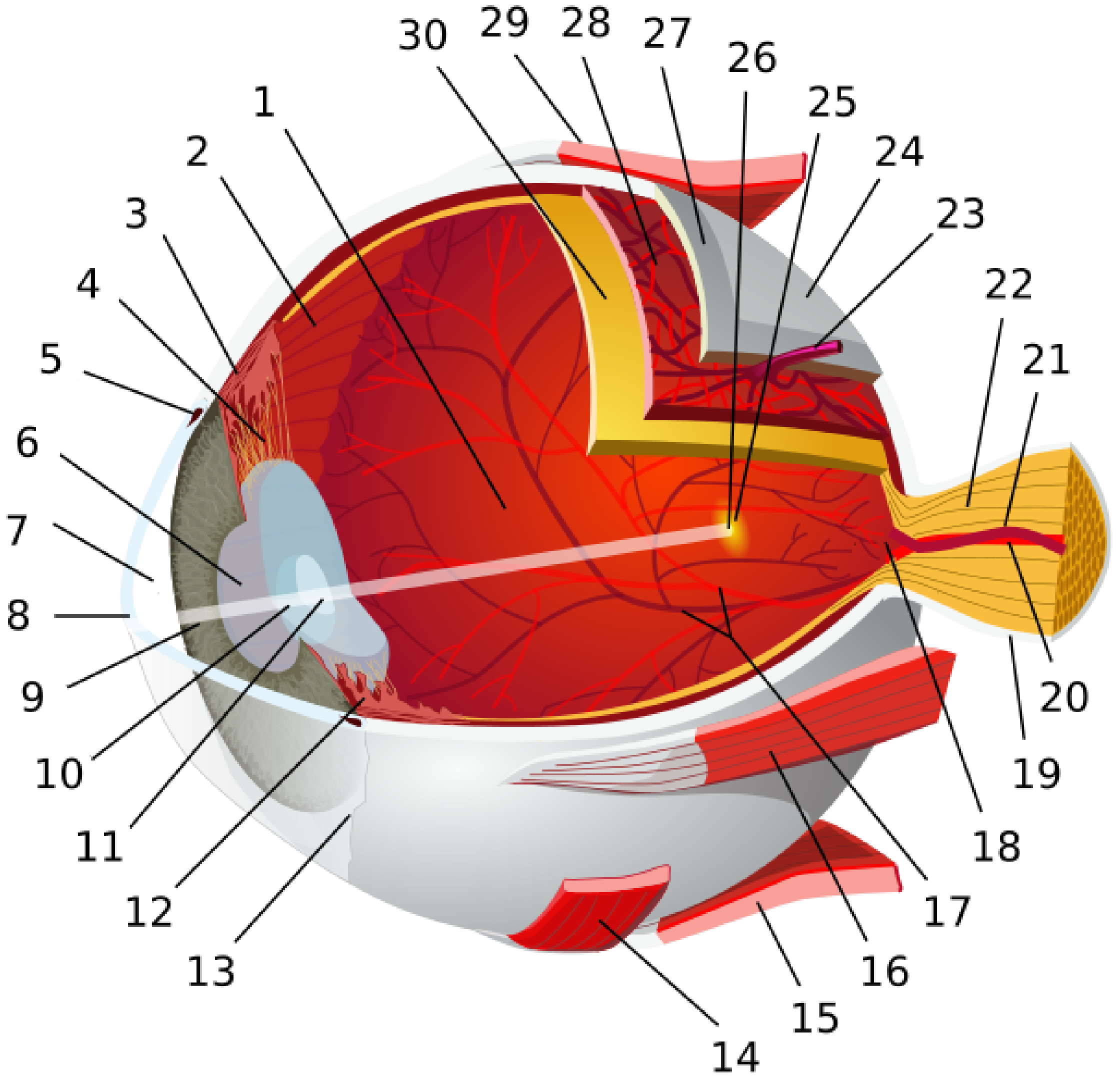
A laser is a device that emits light (electromagnetic radiation) through a process called *stimulated emission*. The term *laser* is an acronym for *Light Amplification by Stimulated Emission of Radiation*. Laser light is usually spatially coherent, which means that the light either is emitted in a narrow,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

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

Applications

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

Human eye



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

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

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

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

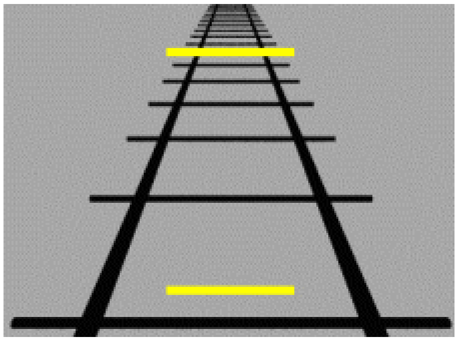
Ciliary muscles around the lens allow the eye's focus to be adjusted. This process is known as accommodation. The near point and far point define the nearest and farthest distances from the eye at which an object can be brought into sharp focus. For a person with normal vision, the far point is located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

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

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

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

Visual effects



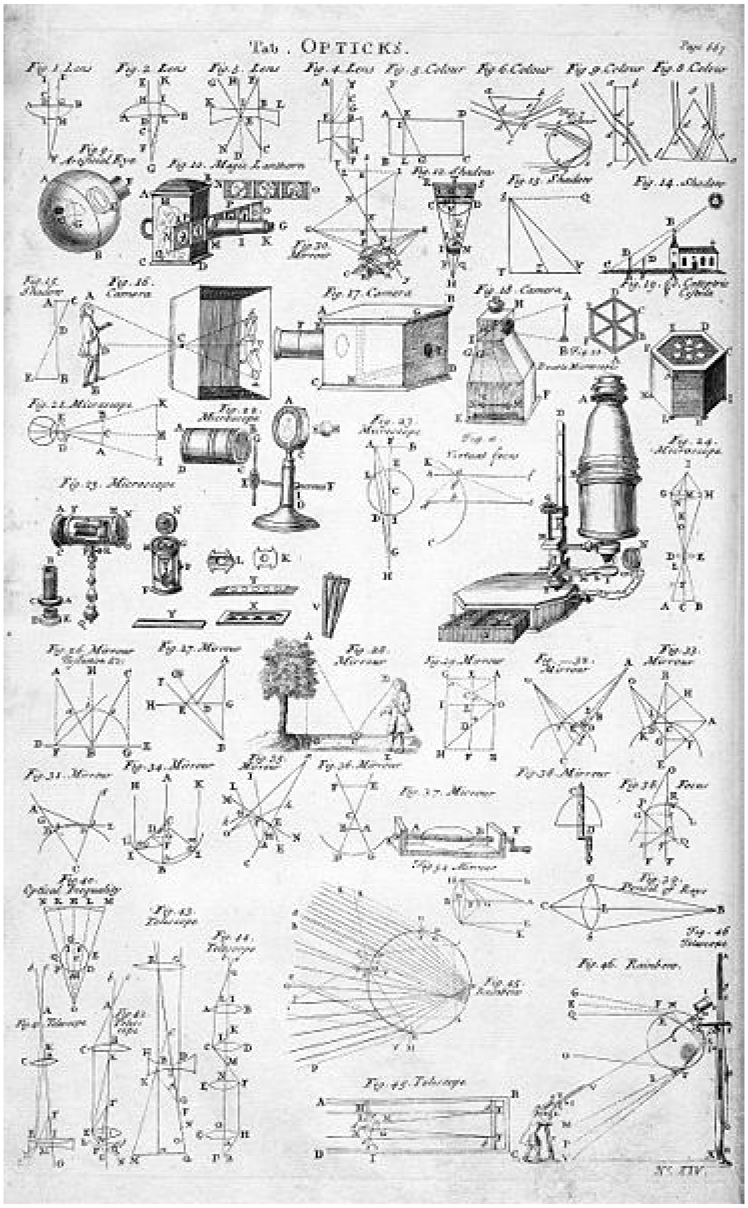
The Ponzo Illusion relies on the fact that parallel lines appear to converge as they approach infinity.

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

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

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

Optical instruments



Illustrations of various optical instruments from the 1728 *Cyclopaedia*

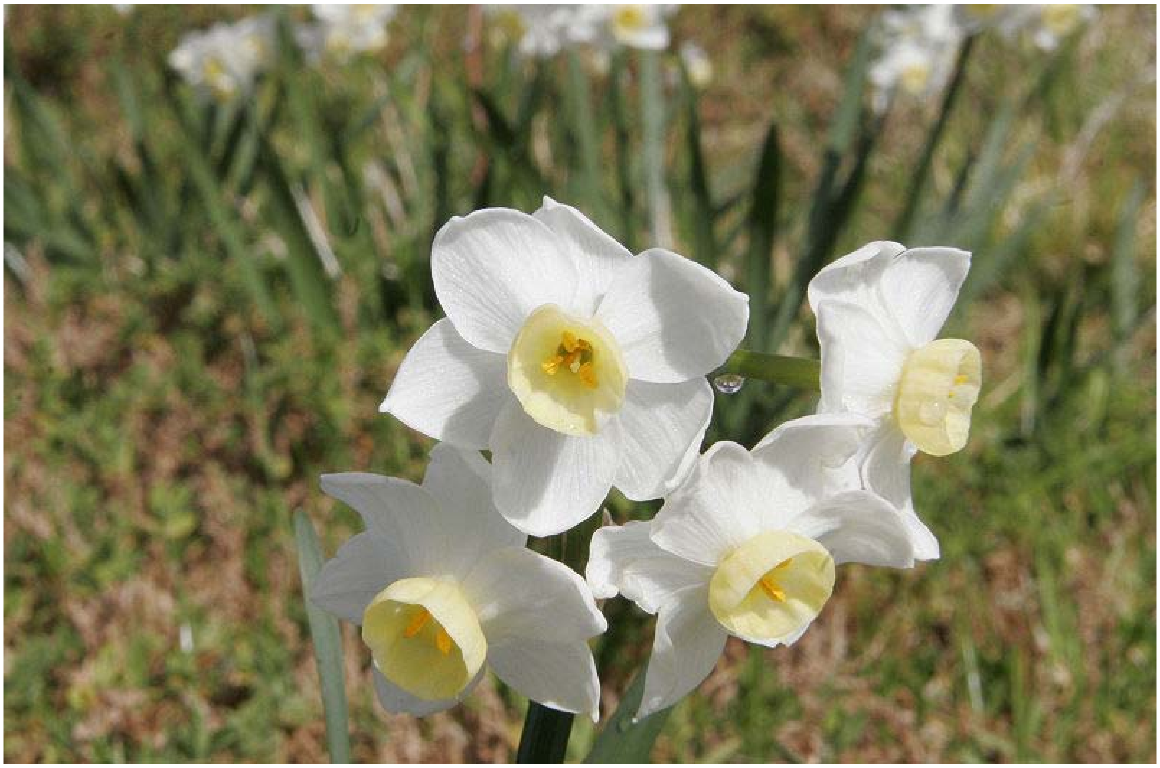
Single lenses have a variety of applications including photographic lenses, corrective lenses, and magnifying glasses while single mirrors are used in parabolic reflectors and rear-view mirrors. Combining a number of mirrors, prisms, and lenses produces compound optical instruments which have practical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

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

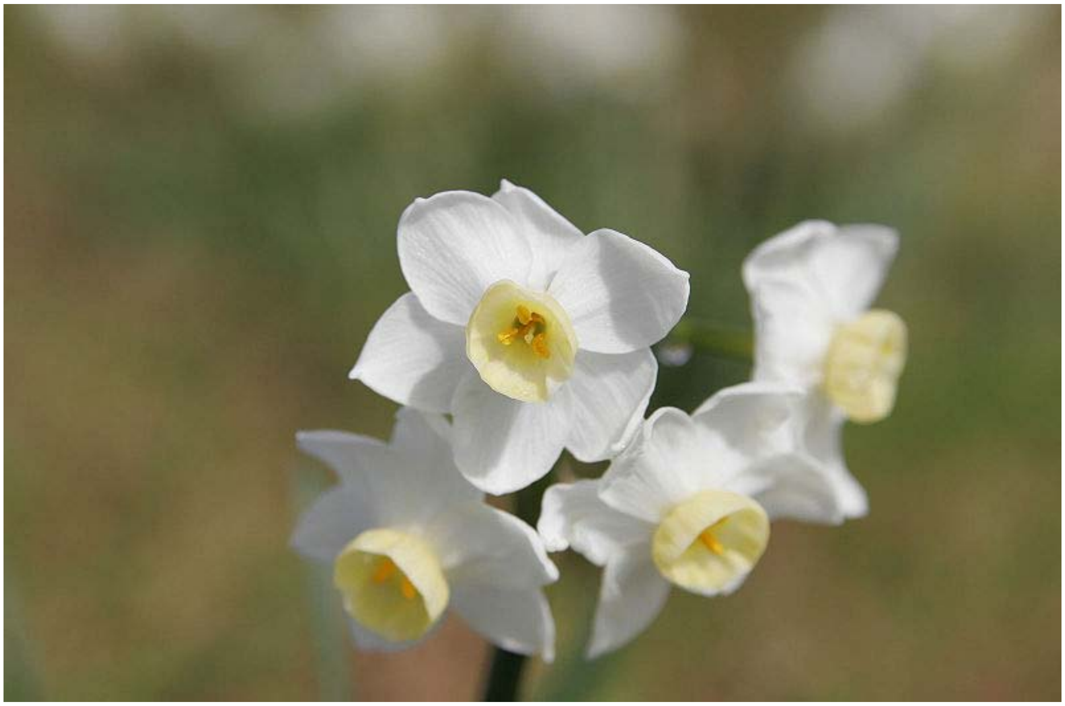
The first telescopes, called *refracting telescopes* were also developed with a single objective and eyepiece lens. In contrast to the microscope, the objective lens of the telescope was designed with a large focal length to avoid optical aberrations. The objective focuses an image of a distant object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

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

Photography



Photograph taken with aperture *f*/32



Photograph taken with aperture *f*/5

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

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

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

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

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

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

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

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

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

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

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

Atmospheric optics

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

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

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

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

Chapter 9

Geophysics

Geophysics is the physics of the Earth and its environment in space. Its subjects include the shape of the Earth, its gravitational and magnetic fields, the dynamics of the Earth as a whole and of its component parts, the Earth's internal structure, composition and tectonics, the generation of magmas, volcanism and rock formation, the hydrological cycle including snow and ice, all aspects of the oceans, the atmosphere, ionosphere, magnetosphere and solar-terrestrial relations, and analogous problems associated with the Moon and other planets.

Geophysics is also applied to societal needs, such as mineral resources, mitigation of natural hazards and environmental protection. Geophysical survey data are used to analyze potential petroleum reservoirs and mineral deposits, to locate groundwater, to locate archaeological finds, to find the thicknesses of glaciers and soils, and for environmental remediation.

History

Replica of Zhang Heng's seismoscope.

Ancient and classical eras

The magnetic compass existed in China back as far as the fourth century BC. It was used as much for feng shui as for navigation on land. It was not until good steel needles could be forged that compasses were used for navigation at sea; before that, they could not retain their magnetism for long. The first mention of a compass in Europe was in 1190.

In circa 240 BC, Erastothenes of Cyrene deduced that the Earth was round and measured the circumference of the Earth, using trigonometry and the angle of the Sun at more than one latitude in Egypt. He developed a system of latitude and longitude and measured the tilt of the Earth's axis.

Perhaps the earliest contribution to seismology was the invention of a seismoscope by the prolific inventor Zhang Heng in 132 CE. This instrument was designed to drop a bronze ball from the mouth of a dragon into the mouth of a toad. By looking at which of eight toads had the ball, one could determine the direction of the earthquake. It was 1571 years before the first design for a seismoscope was published in Europe, by Jean de la Hautefeuille. It was never built.

Beginnings of modern science

One of the publications that marked the beginning of modern science was William Gilbert's *De Magnete* (1600), a report of a series of meticulous experiments in magnetism. Gilbert deduced that compasses point north because the Earth itself is magnetic.

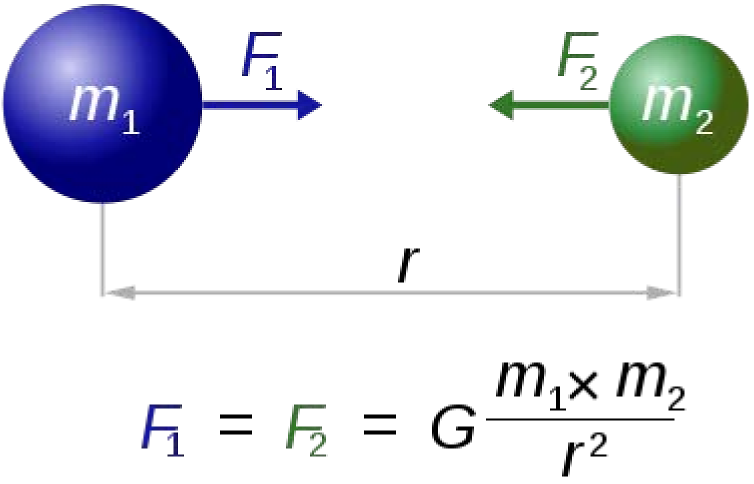
In 1687 Isaac Newton published his *Principia*, which not only laid the foundations for classical mechanics and gravitation but also explained a variety of geophysical phenomena such as the tides and the precession of the equinox.

The first seismometer, an instrument capable of keeping a continuous record of seismic activity, was built by James Forbes in 1844.

Physical phenomena

Geophysics is a highly interdisciplinary subject and geophysicists contribute to every area of the Earth sciences. To provide a clearer idea of what constitutes geophysics, here we, describes phenomena that are studied in physics and how they relate to the Earth and its surroundings.

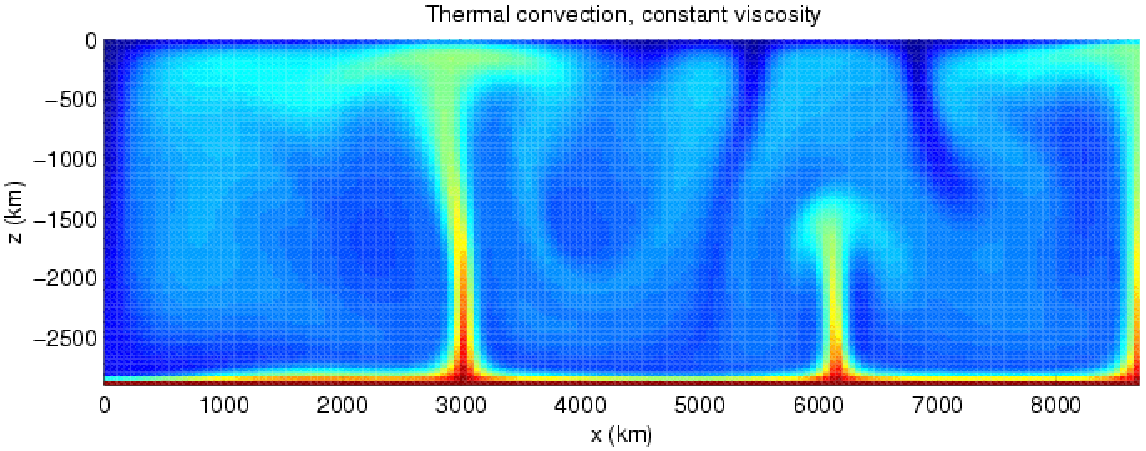
Gravity



The mechanism of Newton's law of universal gravitation.

The gravitational pull of the Moon and Sun give rise to two high tides and two low tides every lunar day, or every 24 hours and 50 minutes. Therefore, there is a gap of 12 hours and 25 minutes between every high tide and between every low tide. Gravitational forces make rocks press down on deeper rocks, increasing their density as the depth increases. Measurements of gravitational acceleration and gravitational potential at the Earth's surface and above it can be used to look for mineral deposits. They also reflect the dynamics of tectonic plates. The geopotential surface called the geoid is one definition of the shape of the Earth. The geoid would be the global mean sea level if the oceans were in equilibrium and could be extended through the continents (such as with very narrow canals).

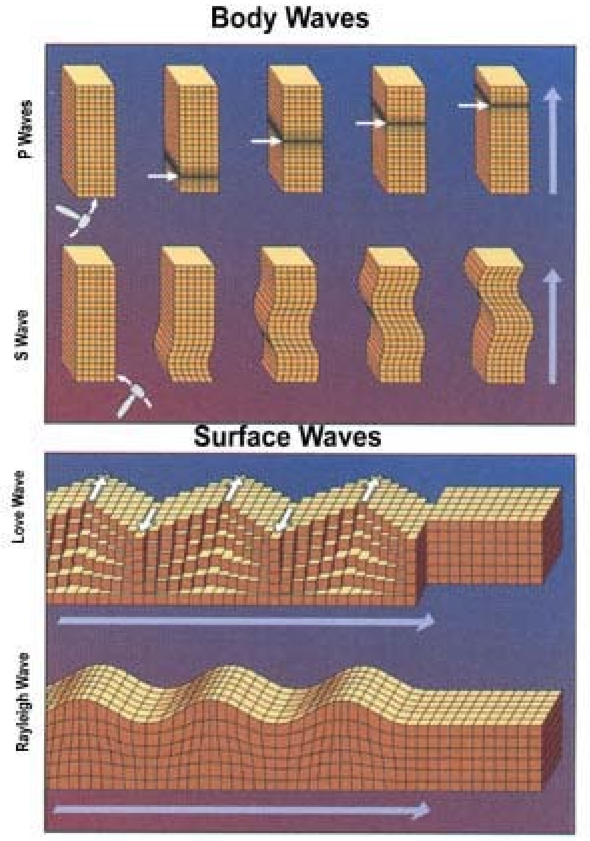
Heat flow



A model of thermal convection in the Earth's mantle.

The Earth is cooling, and the resulting heat flow generates the Earth's magnetic field through the geodynamo and plate tectonics through mantle convection. The main sources of heat are the primordial heat and radioactivity, although there are also contributions from phase transitions. Heat is mostly carried to the surface by thermal convection, although there are two thermal boundary layers - the core-mantle boundary and the lithosphere - in which heat is transported by conduction. Some heat is carried up from the bottom of the mantle by mantle plumes. The heat flow at the Earth's surface is about 4.2 × 10^{13} W, and it is a potential source of geothermal energy.

Vibrations



Body waves and surface waves.

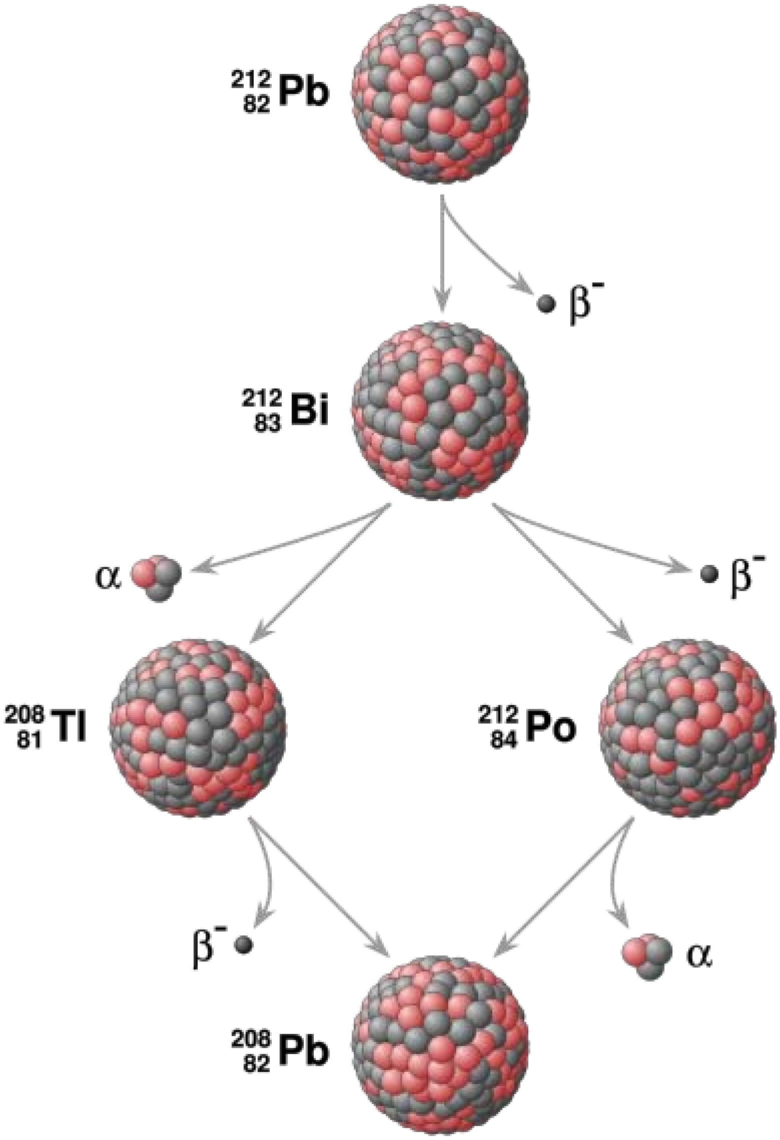
Seismic waves are vibrations that travel through the Earth's interior or along its surface. The entire Earth can also oscillate in forms that are called normal modes. One such mode is the "breathing mode", a uniform expansion and contraction of the Earth.

Ground motions from waves or normal modes are measured using seismographs. If the waves come from a localized source such as an earthquake or explosion, measurements at more than one location can be used to locate the source. The locations of earthquakes provide information on plate tectonics and mantle convection.

Seismic waves can also provide information on the region that the waves travel through. If the density or composition of the rock changes suddenly, some of the waves are reflected. Reflections can provide information on near-surface structure. Changes in the travel direction, called refraction, can be used to infer the deep structure of the Earth.

Earthquakes pose a risk to humans. Understanding their mechanisms, which depend on the type of earthquake (e.g., intraplate or deep focus), can lead to better estimates of earthquake risk and improvements in earthquake engineering.

Radioactivity



Example of a radioactive decay chain

Radioactive decay, in addition to being the main source of heat in the Earth, is an invaluable tool for geochronology. Unstable isotopes decay at predictable rates, and the decay rates of different isotopes cover several orders of magnitude, so radioactive decay can be used to accurately date both recent events and events in past geologic eras.

Electricity

Although we mainly notice electricity during thunderstorms, there is always a downward electric field near the surface that averages 120 V m^{-1}. Relative to the solid Earth, the atmosphere has a net positive charge due to bombardment by cosmic rays. A current of about 1800 A flows in the global circuit. It flows downward from the ionosphere over most of the Earth and back upwards through thunderstorms. The flow is manifested by lightning below the clouds and sprites above.

A variety of electric methods are used in geophysical survey. Some measure spontaneous potential, a potential that arises in the ground because of man-made or natural disturbances. Telluric currents flow in Earth and the Oceans. They have two causes: electromagnetic induction by the time-varying, external-origin geomagnetic field and motion of conducting bodies (such as seawater) across the Earth's permanent magnetic field. The distribution of telluric current density can be used to detect variations in electrical resistivity of underground structures. Geophysicists can also provide the electric current themselves.

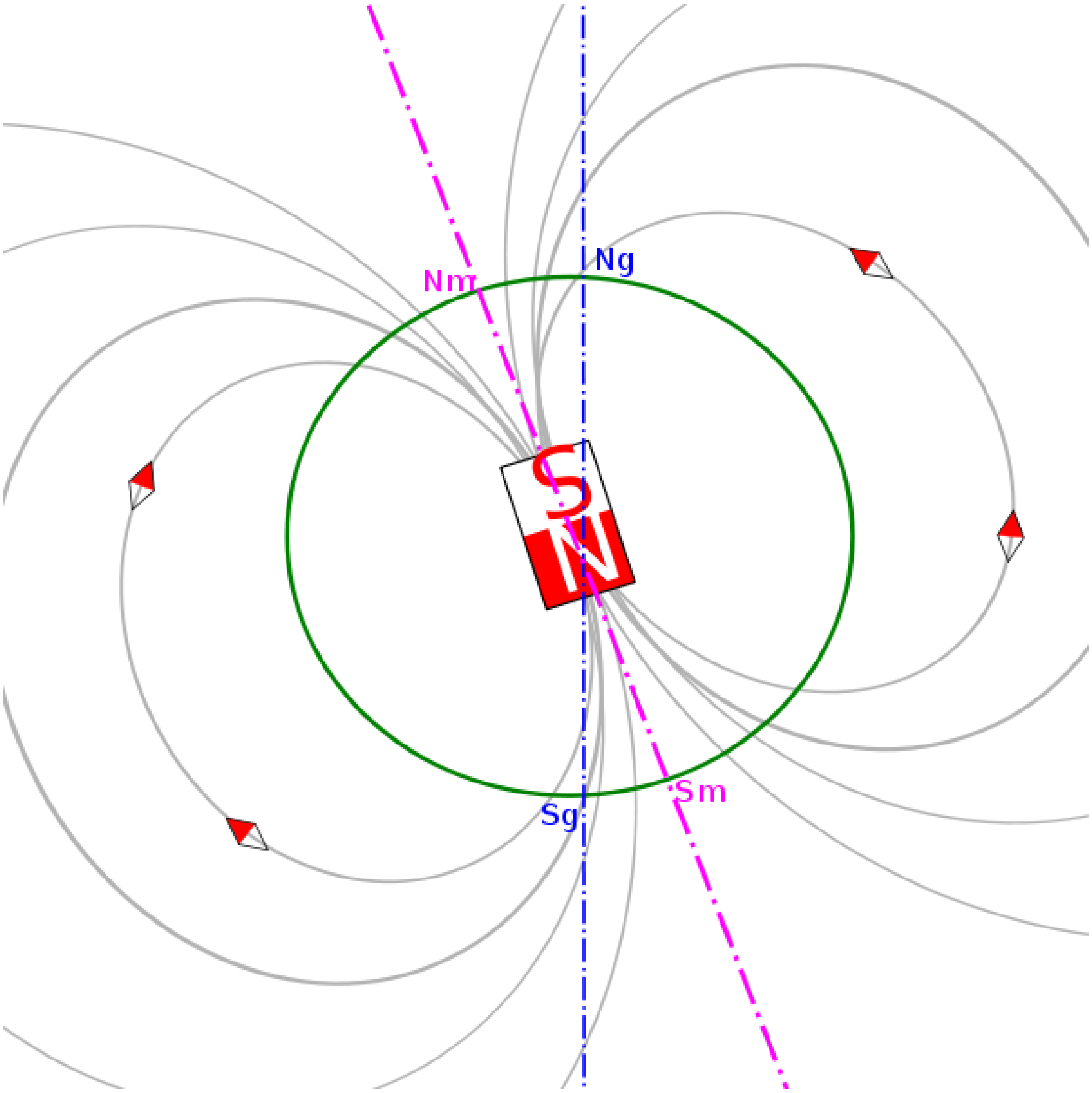
Electromagnetic waves

Electromagnetic waves occur in the ionosphere and magnetosphere as well as the Earth's outer core. dawn chorus is caused by high-energy electrons that get caught in the Van Allen radiation belt. Whistlers are produced by lightning strikes. Hiss may be generated by both. Electromagnetic waves may also be generated by earthquakes.

In the Earth's outer core, electric currents in the highly conductive liquid iron create magnetic fields by electromagnetic induction. Alfvén waves are magnetohydrodynamic waves in the magnetosphere or the Earth's core. In the core, they probably have little observable effect on the geomagnetic field, but slower waves such as magnetic Rossby waves may be one source of geomagnetic secular variation.

Electromagnetic methods that are used for geophysical survey include transient electromagnetics and magnetotellurics.

Magnetism



The variation between magnetic north and "true" north.

The Earth's magnetic field protects the Earth from the deadly Solar wind and has long been used for navigation. It originates in the fluid motions of the Earth's outer core. The magnetic field in the upper atmosphere gives rise to the auroras.

The Earth's field is roughly like a tilted dipole, but it changes over time (a phenomenon called geomagnetic secular variation). Mostly the geomagnetic pole stays near the geographic pole, but at random intervals averaging a million years or so, the polarity of the Earth's field reverses. These geomagnetic reversals are recorded in rocks and their signature can be seen in striped magnetic anomalies on the seafloor. These stripes provide quantitative information on seafloor spreading, a part of plate tectonics. In addition, the magnetization in rocks can be used to measure the motion of continents.

Fluid dynamics

Fluid motions occur in the magnetosphere, atmosphere, ocean, mantle and core. Even the mantle, though it has an enormous viscosity, flows like a fluid over long time intervals. This flow is reflected in phenomena such as isostasy and post-glacial rebound. The mantle flow drives plate tectonics and the flow in the Earth's core drives the geodynamo.

Geophysical fluid dynamics is a primary tool in physical oceanography and meteorology. The rotation of the Earth has profound effects on the Earth's fluid dynamics, often due to the Coriolis effect. In the atmosphere it gives rise to large-scale patterns like Rossby waves and determines the basic circulation patterns of storms. In the ocean they drive large-scale circulation patterns as well as Kelvin waves and Ekman spirals at the ocean surface. In the Earth's core, the circulation of the molten iron is structured by Taylor columns.

Waves and other phenomena in the magnetosphere can be modeled using magnetohydrodynamics.

Condensed matter physics

The physical properties of minerals must be understood to infer the composition of the Earths' interior from seismology, the geothermal gradient and other sources of information. Mineral physicists study the elastic properties of minerals as well as their high-pressure phase diagrams, melting points and equations of state at high pressure. Studies of creep determine how rocks that are brittle at the surface can flow deep down. These properties determine the rheology that determines the geodynamics.

Water is a very complex substance and its unique properties are essential for life. Its physical properties shape the hydrosphere and are an essential part of the water cycle and climate. Its thermodynamic properties determine evaporation and the thermal gradient in the atmosphere. The many types of precipitation involve a complex mixture of processes such as coalescence, supercooling and supersaturation. Some of the precipitated water becomes groundwater, and groundwater flow includes phenomena such as percolation, while the conductivity of water makes electrical and electromagnetic methods useful for tracking groundwater flow. Physical properties of water such as salinity have a large effect on its motion in the oceans.

The many phases of ice form the cryosphere and come in forms like ice sheets, glaciers, sea ice, freshwater ice, snow, and frozen ground (or permafrost).

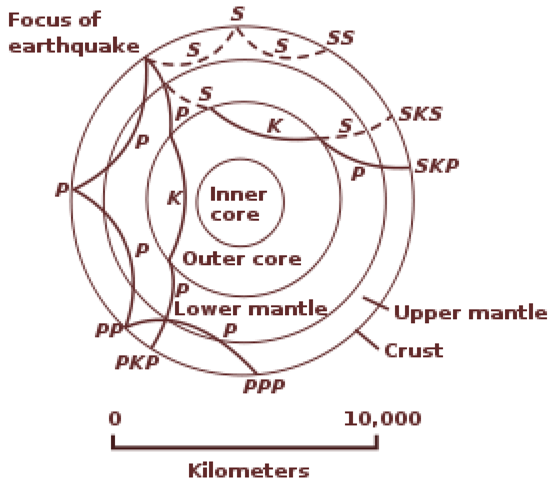
Regions of the Earth

Size and form of the Earth

The Earth is roughly spherical, but it bulges towards the Equator, so it is roughly in the shape of an ellipsoid. This bulge is due to its rotation and is nearly consistent with an

Earth in hydrostatic equilibrium. The detailed shape of the Earth, however, is also affected by the distribution of continents and ocean basins, and to some extent by the dynamics of the plates.

Structure of the Earth



Mapping the interior of the Earth with earthquake waves.

Evidence from seismology, heat flow at the surface, and mineral physics is combined with the Earth's mass and moment of inertia to infer models of the Earth's interior - its composition, density, temperature, pressure. The Earth's mass is $M = 5.975 \times 10^{24}$ kg and its mean radius is $R = 6371$ km , so its mean specific gravity is $< \rho > = 5.515$. This is substantially higher than the typical specific gravity (2.7–3.3) of rocks at the surface. Its moment of inertia is 0.33 $M\,R^2$, whereas it would be 0.4 $M\,R^2$ if the earth was a sphere of constant density. Both lines of evidence point to a concentration of mass near the center. However, the density of the rock will increase with depth because of the increasing pressure. To determine how large this effect is, the Adams–Williamson equation is used to determine how density increases with pressure. The conclusion is that pressure alone cannot account for the increase in density. Instead, we know that the Earth's core is composed of an alloy of iron and other minerals.

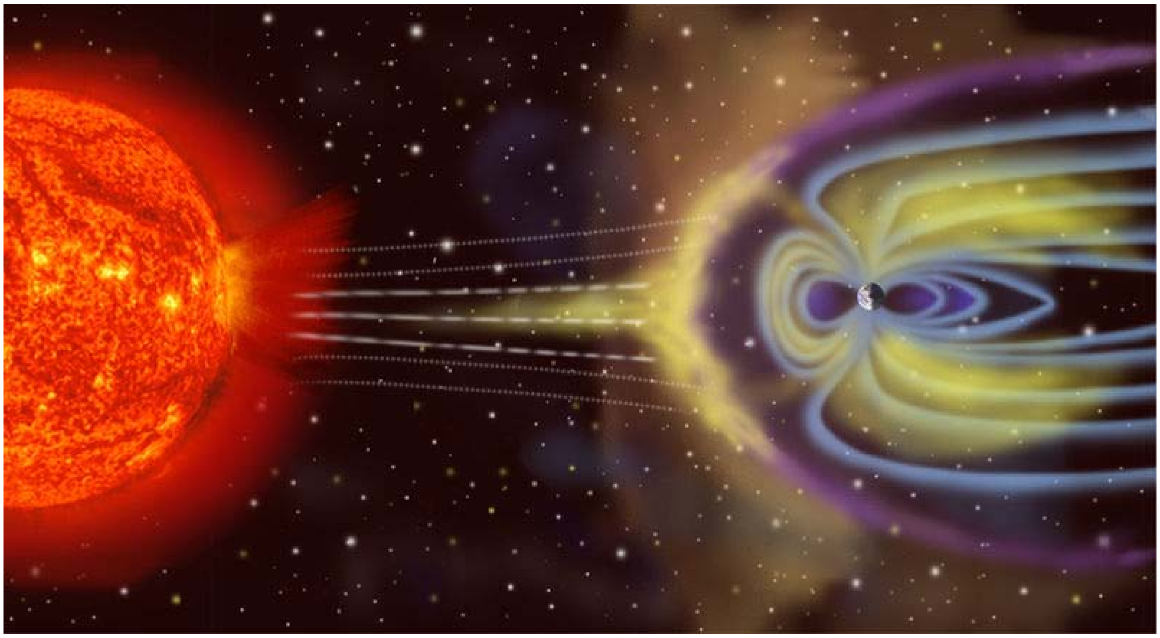
Reconstructions of seismic waves in the deep interior of the Earth show that there are no S-waves in the outer core. This indicates that the outer core is liquid, because liquids cannot support shear. The outer core is liquid, and the motion of this highly conductive fluid generates the Earth's field. The inner core, however, is solid because of the enormous pressure.

Reconstruction of seismic reflections in the deep interior indicate some major discontinuities in seismic velocities that demarcate the major zones of the Earth: inner core, outer core, mantle, lithosphere and crust. The mantle itself is divided into the upper mantle, transition zone, lower mantle and D'' layer. Between the crust and the mantle is the Mohorovičić discontinuity.

The seismic model of the Earth does not by itself determine the composition of the layers. For a complete model of the Earth, mineral physics is needed to interpret seismic velocities in terms of composition. The mineral properties are temperature-dependent, so the geotherm must also be determined. This requires physical theory for thermal conduction and convection and the heat contribution of radioactive elements. The main model for the radial structure of the interior of the Earth is the Preliminary Reference Earth Model (PREM). Some parts of this model have been updated by recent findings in mineral physics and supplemented by seismic tomography. The mantle is mainly composed of silicates, and the boundaries between layers of the mantle are probably due to phase transitions.

The mantle acts as a solid for seismic waves, but under high pressures and temperatures it deforms so that over millions of years it acts like a liquid. This makes plate tectonics possible. Geodynamics is the study of the fluid flow in the mantle and core.

The magnetosphere



The solar wind is deflected by the magnetosphere (not to scale)

If a planet's magnetic field is strong enough, its interaction with the solar wind forms a magnetosphere around a planet. Early space probes mapped out the gross dimensions of the terrestrial magnetic field, which extends about 10 Earth radii towards the Sun. The solar wind, a stream of charged particles, streams out and around the terrestrial magnetic field, and continues behind the magnetic tail, hundreds of Earth radii downstream. Inside the magnetosphere, there are relatively dense regions of solar wind particles, the Van Allen radiation belts.

Other fields and related disciplines

Fields

- Geodesy, measurement of the Earth: GPS, vertical and horizontal motions of the Earth's surface, navigation, the study of the Earth's gravitational field, and the size and form of the Earth
- The study of large-scale motions of the Earth's surface and interior, including:

    - Tectonophysics, the study of the physical processes that cause and result from plate tectonics
    - Geodynamics, the study of modes of transport deformation within the Earth: rock deformation, mantle flow and convection, heat flow, lithosphere dynamics

- Geomagnetism, the study of the Earth's magnetic field, including its origin, telluric currents driven by the magnetic field, the Van Allen belts, and the interaction between the magnetosphere and the solar wind. This field is associated with paleomagnetism, or the measurement of the orientation of the Earth's magnetic field over the geologic past.
- Seismology, the study of the structure and composition of the Earth through seismic waves, and of surface deformations during earthquakes and seismic hazards
- Mathematical geophysics, The development and applications of mathematical methods and techniques for the solution of geophysical problems.
- Geophysical surveying:

    - Exploration and engineering geophysics, using surface methods to detect or infer the presence and position of concentrations of ore minerals and hydrocarbons
    - Archaeological geophysics, for archaeological imaging or mapping
    - Environmental and Engineering Geophysics, for locating underground storage tanks (USTs) or utilities, Unexploded ordnance (UXO), delineating landfills, locating voids or potential subsidence, finding depth to, P-wave or S-wave velocity of, or rippability of bedrock, or the pathway of groundwater movement
    - Shallow seismology is used in exploration geophysics (to find oil and gas) and for environmental characterization of the subsurface

Related disciplines

- Volcanology, the study of volcanoes, volcanic features (hot springs, geysers, fumaroles), volcanic rock, and heat flow related to volcanoes
- Atmospheric sciences, which includes:

    - Atmospheric electricity and the ionosphere

    - Aeronomy, the study of the physical structure and chemistry of the atmosphere.
    - Meteorology and Climatology, which both involve studies of the weather.

- The study of water on the Earth, hydrology, physical oceanography and glaciology
- Geological and geophysical engineering and Engineering geology, applying geophysics to the engineering design of facilities including roads, tunnels, and mines
- The study of the rocks and minerals, including petrophysics and aspects of mineralogy such as physical mineralogy and crystal structure

Methods of geophysics

Space probes

Space probes made it possible to collect data from not only the visible light region, but in other areas of the electromagnetic spectrum. The planets can be characterized by their force fields: gravity and their magnetic fields, which are studied through geophysics and space physics.

Measuring the changes in acceleration experienced by spacecraft as they orbit has allowed fine details of the gravity fields of the planets to be mapped. For example, in the 1970s, the gravity field disturbances above lunar maria were measured through lunar orbiters, which lead to the discovery of concentrations of mass, mascons, beneath the Imbrium, Serenitatis, Crisium, Nectaris and Humorum basins.

In 2002, NASA launched the Gravity Recovery and Climate Experiment, wherein two twin satellites map variations in Earth's gravity field by making measurements of the distance between the two satellites using GPS and a microwave ranging system. Gravity variations detected by GRACE include those caused by changes in ocean currents; runoff and ground water depletion; melting ice sheets and glaciers.

Chapter 10

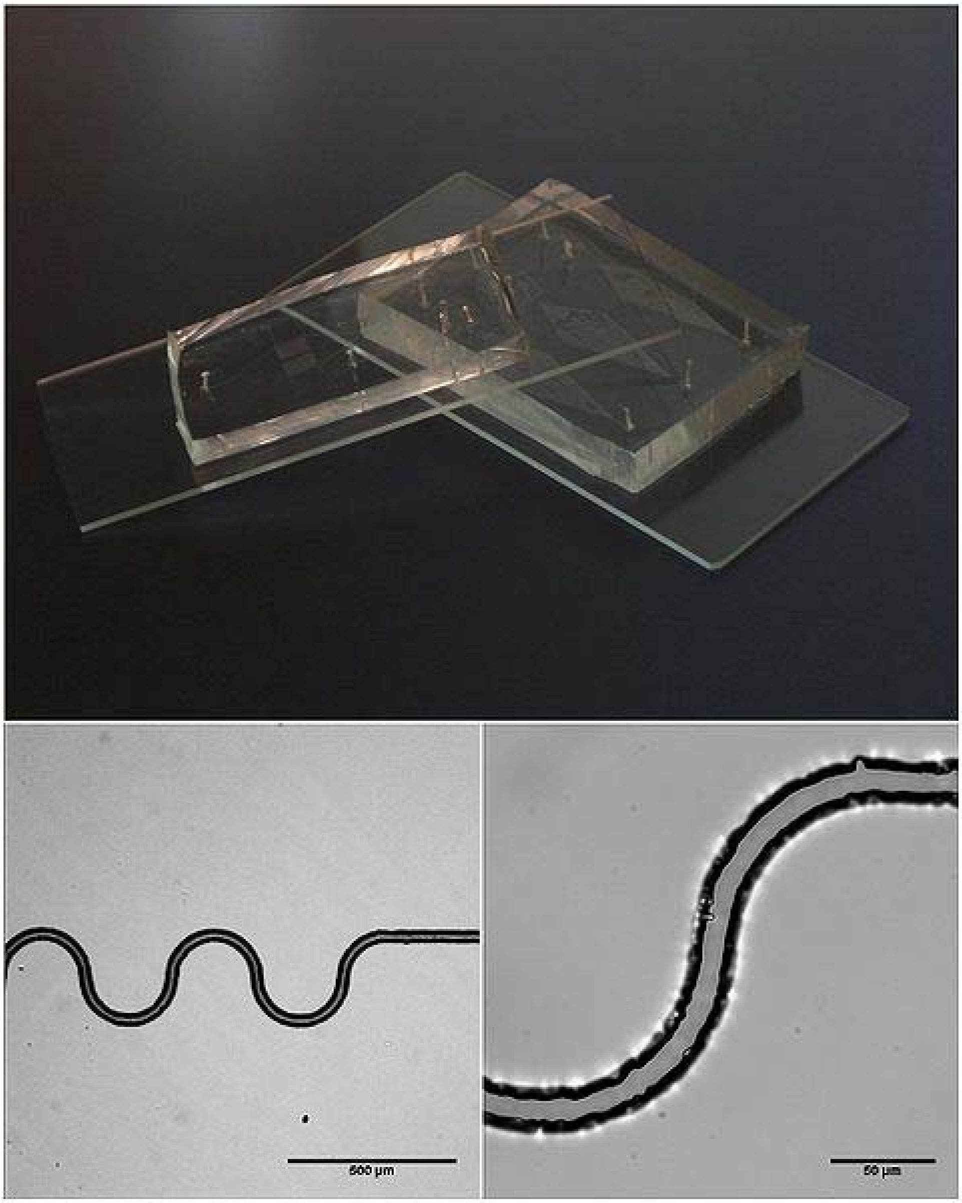
Microfluidics

Microfluidics deals with the behavior, precise control and manipulation of fluids that are geometrically constrained to a small, typically sub-millimeter, scale. Typically, **micro** means one of the following features:

- small volumes (nl, pl, fl)
- small size
- low energy consumption
- effects of the micro domain

It is a multidisciplinary field intersecting engineering, physics, chemistry, microtechnology and biotechnology, with practical applications to the design of systems in which such small volumes of fluids will be used. Microfluidics emerged in the beginning of the 1980s and is used in the development of inkjet printheads, DNA chips, lab-on-a-chip technology, micro-propulsion, and micro-thermal technologies.

Microscale behavior of fluids



Silicone rubber and glass microfluidic devices. Top: a photograph of the devices. Bottom: DIC micrographs of a serpentine channel ~15 µm wide.

The behavior of fluids at the microscale can differ from 'macrofluidic' behavior in that factors such as surface tension, energy dissipation, and fluidic resistance start to dominate the system. Microfluidics studies how these behaviors change, and how they can be worked around, or exploited for new uses.

At small scales (channel diameters of around 100 nanometers to several hundred micrometers) some interesting and sometimes unintuitive properties appear. In particular, the Reynolds number (which compares the effect of momentum of a fluid to the effect of viscosity) can become very low. A key consequence of this is that fluids, when side-by-side, do not necessarily mix in the traditional sense; molecular transport between them must often be through diffusion.

High specificity of chemical and physical properties (concentration, pH, temperature, shear force, etc.) can also be ensured resulting in more uniform reaction conditions and higher grade products in single and multi-step reactions.

Effects of micro domain

- laminar flow
- surface tension
- electrowetting
- fast thermal relaxation
- electrical surface charges
- diffusion

Key application areas

Microfluidic structures include micropneumatic systems, i.e. microsystems for the handling of off-chip fluids (liquid pumps, gas valves, etc.), and microfluidic structures for the on-chip handling of nano- and picolitre volumes. To date, the most successful commercial application of microfluidics is the inkjet printhead. Significant research has been applied to the application of microfluidics for the production of industrially relevant quantities of material.

Advances in microfluidics technology are revolutionizing molecular biology procedures for enzymatic analysis (e.g., glucose and lactate assays), DNA analysis (e.g., polymerase chain reaction and high-throughput sequencing), and proteomics. The basic idea of microfluidic biochips is to integrate assay operations such as detection, as well as sample pre-treatment and sample preparation on one chip.

An emerging application area for biochips is clinical pathology, especially the immediate point-of-care diagnosis of diseases. In addition, microfluidics-based devices, capable of continuous sampling and real-time testing of air/water samples for biochemical toxins and other dangerous pathogens, can serve as an always-on "bio-smoke alarm" for early warning.

Continuous-flow microfluidics

These technologies are based on the manipulation of continuous liquid flow through microfabricated channels. Actuation of liquid flow is implemented either by external pressure sources, external mechanical pumps, integrated mechanical micropumps, or by combinations of capillary forces and electrokinetic mechanisms. Continuous-flow microfluidic operation is the mainstream approach because it is easy to implement and less sensitive to protein fouling problems. Continuous-flow devices are adequate for many well-defined and simple biochemical applications, and for certain tasks such as chemical separation, but they are less suitable for tasks requiring a high degree of flexibility or ineffect fluid manipulations. These closed-channel systems are inherently difficult to integrate and scale because the parameters that govern flow field vary along the flow path making the fluid flow at any one location dependent on the properties of the entire system. Permanently-etched microstructures also lead to limited reconfigurability and poor fault tolerance capability.

Process monitoring capabilities in continuous-flow systems can be achieved with highly sensitive microfluidic flow sensors based on MEMS technology which offer resolutions down to the nanoliter range.

Digital (droplet-based) microfluidics

Alternatives to the above closed-channel continuous-flow systems include novel open structures, where discrete, independently controllable droplets are manipulated on a substrate using electrowetting. Following the analogy of digital microelectronics, this approach is referred to as digital microfluidics. Le Pesant et al. pioneered the use of electrocapillary forces to move droplets on a digital track. The "fluid transistor" pioneered by Cytonix also played a role. The technology was subsequently commercialized by Duke University. By using discrete unit-volume droplets, a microfluidic function can be reduced to a set of repeated basic operations, i.e., moving one unit of fluid over one unit of distance. This "digitization" method facilitates the use of a hierarchical and cell-based approach for microfluidic biochip design. Therefore, digital microfluidics offers a flexible and scalable system architecture as well as high fault-tolerance capability. Moreover, because each droplet can be controlled independently, these systems also have dynamic reconfigurability, whereby groups of unit cells in a microfluidic array can be reconfigured to change their functionality during the concurrent execution of a set of bioassays. Although droplets are manipulated in confined microfluidic channels, since the control on droplets is not independent, it should not be confused as "digital microfluidics". One common actuation method for digital microfluidics is electrowetting-on-dielectric (EWOD). Many lab-on-a-chip applications have been demonstrated within the digital microfluidics paradigm using electrowetting. However, recently other techniques for droplet manipulation have also been demonstrated using Surface Acoustic Waves, optoelectrowetting, mechanical actuation, etc.

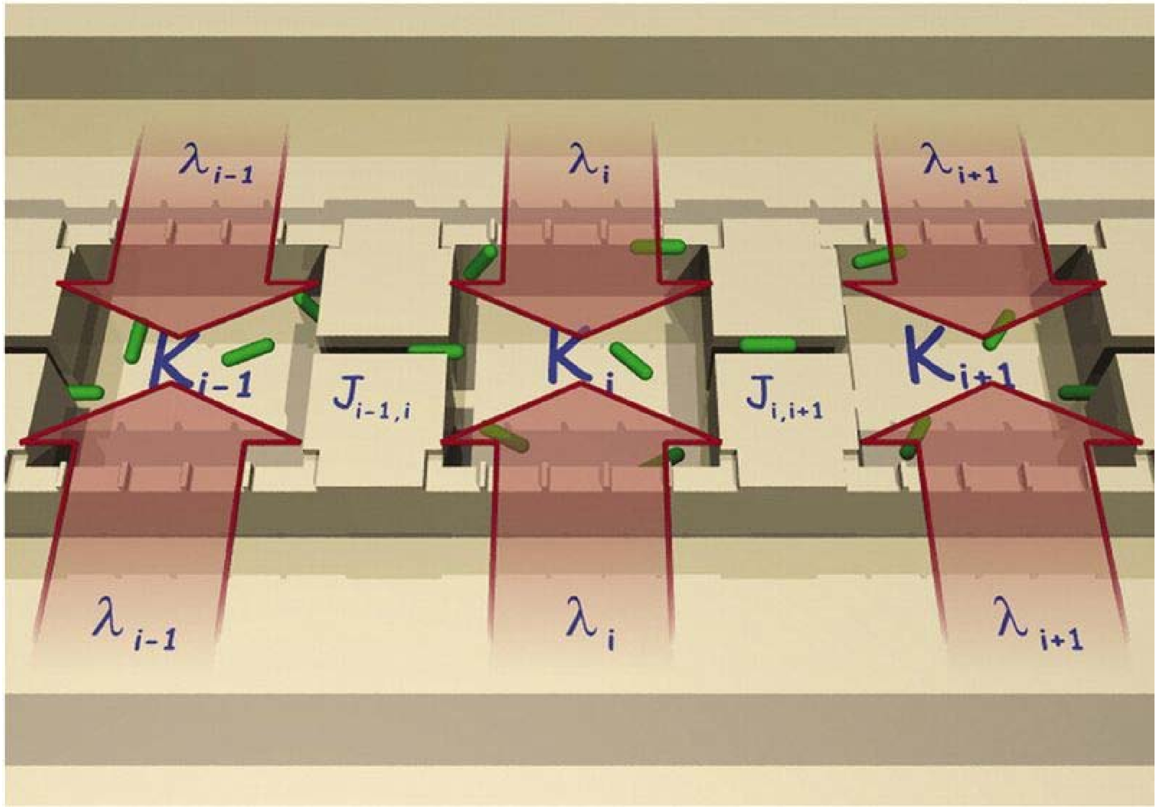
DNA chips (microarrays)

Early biochips were based on the idea of a DNA microarray, e.g., the GeneChip DNAarray from Affymetrix, which is a piece of glass, plastic or silicon substrate on which pieces of DNA (probes) are affixed in a microscopic array. Similar to a DNA microarray, a protein array is a miniature array where a multitude of different capture agents, most frequently monoclonal antibodies, are deposited on a chip surface; they are used to determine the presence and/or amount of proteins in biological samples, e.g., blood. A drawback of DNA and protein arrays is that they are neither reconfigurable nor scalable after manufacture. Digital microfluidics has been described as a means for carrying out Digital PCR.

Molecular biology

In addition to microarrays biochips have been designed for two-dimensional electrophoresis, transcriptome analysis, and PCR amplification. Other applications include various electrophoresis and liquid chromatography applications for proteins and DNA, cell separation, in particular blood cell separation, protein analysis, cell manipulation and analysis including cell viability analysis and microorganism capturing.

Evolutionary biology



Three Micro Habitat Patches MHPs connected by dispersal corridors (indicated here as $J_{i,j}$) into a 1D lattice. The ecosystem service (of habitat renewal) to each MHP represented here as λ_i (red arrows). Each MHP can also hold different carrying capacity K_i for its supporting local population of bacterial cells (depicted in green).

By combining microfluidics with landscape ecology and nanofluidics, a nano/micro fabricated fluidic landscape can be constructed by building local patches of bacterial habitat and connecting them by dispersal corridors. The resulting landscapes can be used as physical implementations of an adaptive landscape, by generating a spatial mosaic of patches of opportunity distributed in space and time. The patchy nature of these fluidic landscapes allows for the study of adapting bacterial cells in a metapopulation system.

The evolutionary ecology of these bacterial systems in these synthetic ecosystems allows for using biophysics to address questions in evolutionary biology.

Cellular biophysics

By rectifying the motion of individual swimming bacteria, microfluidic structures can be use to extract mechanical motion from a population of motile bacterial cells. This way, bacteria-powered rotors can be built.

Optics

The merger of microfluidics and optics is typical known as Optofluidics. An example of an optofluidic device is a Tuneable Microlens Array

Acoustic droplet ejection (ADE)

Acoustic droplet ejection uses a pulse of ultrasound to move low volumes of fluids (typically nanoliters or picoliters) without any physical contact. This technology focuses acoustic energy into a fluid sample in order to eject droplets as small as a millionth of a millionth of a liter (picoliter = 10^{-12} liter). ADE technology is a very gentle process, and it can be used to transfer proteins, high molecular weight DNA and live cells without damage or loss of viability. This feature makes the technology suitable for a wide variety of applications including proteomics and cell-based assays.

Fuel cells

Microfluidic fuel cells can use laminar flow to separate the fuel and its oxidant to control the interaction of the two fluids without a physical barrier as would be required in conventional fuel cells.

Ceramic Pot Water Filters

In recent times, developing nations are adopting clay based ceramic water filters for cost effective water filtration. These water filters made from molds manufactured by mixing clay and waste plant materials such as rice husk, sawdust,dried plant biomass etc. in some volumetric or weight proportions. These vessels are in use from very early ages in Africa and Asia. The initial water percolation through there clay pots are alkaline. Recently it was found that this alkaline nature can be easily predicted by use of simple models incorporating micro/nano fluid transport processes such as Capillary Osmosis, Thermo Osmosis, Electro-osmosis and Discharge from these clay ceramic devices.

A tool for cell biological research

Microfluidic technology is creating powerful tools for cell biologists to control the complete cellular environement, leading to new questions and new discoveries . We can list here diverse advantages for microbiology of these tools:

- Microenvironemental control
- Precise spatiotemporal concentration gradients
- Mechanical deformation
- Force measurements of adherent cells
- Confining cells
- Exerting a controlled force
- Fast and precise temperature control
- Electric field integration
- Cell culture

Chapter 11

Accelerator Physics and Quantum Optics

Accelerator physics

Accelerator physics deals with the problems of building and operating particle accelerators.

The experiments conducted with particle accelerators are not regarded as part of **accelerator physics**. These belong (according to the objectives of the experiments) to particle physics, nuclear physics, condensed matter physics, materials physics, etc. as well as to other sciences and technical fields. The types of experiments done at a particular accelerator and/or its other uses are largely constrained by the characteristics of the accelerator itself, such as energy (per particle), types of particles, beam intensity, beam quality, etc.

Accelerator physics itself is the study of the motion of the particle beam through the machine, control and manipulation of the beam, interaction with the machine itself, and measurements of the various parameters associated with particle beams.

Equations of motion

The motion of charged particles through an accelerator is controlled using applied electro-magnetic fields, and the equations of motion may be derived from (or, since in many cases a general solution is not possible, approximated from) relativistic Hamiltonian mechanics. Typically, a separate Hamiltonian is written down for each element (e.g. for a single quadrupole magnet, or accelerating structure) to allow the equations of motion to be solved for this one element. Once this has been done for each element encountered in the machine, the full trajectory of each particle may be calculated for the entire machine.

In many cases a general solution of the full Hamiltonian is not possible, so it is necessary to make approximations. This may take the form of the Paraxial approximation (a Taylor series in the dynamical variables, truncated to low order), however, even in the cases of strongly non-linear magnetic fields, a Lie transform may be used to construct an integrator with a high degree of accuracy, and the paraxial approximation is not necessary.

Diagnostics

A vital component of any accelerator are the diagnostic devices that allow various properties of the particle bunches to be measured.

A typical machine may use many different types of measurement device in order to measure different properties. These include (but are not limited to) Beam Position Monitors (BPMs) to measure the position of the bunch, screens (fluorescent screens, Optical Transition Radiation (OTR) devices) to image the profile of the bunch, wire-scanners to measure its cross-section, and toroids or ICTs to measure the bunch charge (i.e. the number of particles per bunch).

While many of these devices rely on well understood technology, designing a device capable of measuring a beam for a particular machine is a complex task requiring much expertise. Not only is a full understanding of the physics of the operation of the device necessary, but it is also necessary to ensure that the device is capable of measuring the expected parameters of the machine under consideration.

Success of the full range of beam diagnostics often underpins the success of the machine as a whole.

Machine tolerances

Errors in the alignment of components, field strength, etc., are inevitable in machines of this scale, so it is important to consider the tolerances under which a machine may operate.

Engineers will provide the physicists with expected tolerances for the alignment and manufacture of each component to allow full physics simulations of the expected behaviour of the machine under these conditions. In many cases it will be found that the performance is degraded to an unacceptable level, requiring either re-engineering of the components, or the invention of algorithms that allow the machine performance to be 'tuned' back to the design level.

This may require many simulations of different error conditions in order to determine the relative success of each tuning algorithm, and to allow recommendations for the collection of algorithms to be deployed on the real machine.

Interactions between the beam and the machine

Due to the strong electro-magnetic fields that follow the beam, it is possible for it to interact with any electrical impedance in the walls of the beam pipe. This may be in the form of a resistive impedance (i.e. the finite resistivity of the beam pipe material) or an inductive/capacitive impedance (due to the geometric changes in the beam pipe's cross section).

These impedances will induce so called 'wake-fields' (a strong warping of the electromagnetic field of the beam) that can interact with later particles. Since this interaction may have a negative effect, it must be studied to determine its magnitude, and to determine any actions that may be taken to mitigate it.

Quantum optics

Quantum optics is a field of research in physics, dealing with the application of quantum mechanics to phenomena involving light and its interactions with matter.

History of quantum optics

Light is made up of particles called photons and hence inherently is "grainy" (quantized). Quantum optics is the study of the nature and effects of light as quantized photons. The first indication that light might be quantized came from Max Planck in 1899 when he correctly modeled blackbody radiation. By assuming blackbody radiation is quantized, Bohr showed that the atoms were also quantized, in the sense that they could only emit discrete amounts of energy. The understanding of the interaction between light and matter following these developments not only formed the basis of quantum optics but were also crucial for the development of quantum mechanics as a whole. However, the subfields of quantum mechanics dealing with matter-light interaction were principally regarded as research into matter rather than into light; hence one rather spoke of atom physics and quantum electronics in 1960. Laser science—i.e., research into principles, design and application of these devices—became an important field, and the quantum mechanics underlying the laser's principles was studied now with more emphasis on the properties of light, and the name *quantum optics* became customary.

As laser science needed good theoretical foundations, and also because research into these soon proved very fruitful, interest in quantum optics rose. Following the work of 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. This led to the introduction of the coherent state as a quantum description of laser light and the realization that some states of light could not be described with classical waves. In 1977, Kimble et al. demonstrated the first source of light which required a quantum description: a single atom that emitted one photon at a time. This was the first conclusive evidence that light was made up of photons. Another quantum state of light with certain advantages over any classical state, squeezed light, was soon proposed. At the same time, development of short and ultrashort laser pulses—created by Q switching and modelocking techniques—opened the way to the study of unimaginably fast ("ultrafast") processes. Applications for solid state research (e.g. Raman spectroscopy) were found, and mechanical forces of light on matter were studied. The latter led to levitating and positioning clouds of atoms or even small biological samples in an optical trap or optical tweezers by laser beam. This, along with Doppler cooling was the crucial technology needed to achieve the celebrated Bose-Einstein condensation.

Other remarkable results are the demonstration of quantum entanglement, quantum teleportation, and (recently, in 1995) quantum logic gates. The latter are of much interest in quantum information theory, a subject which partly emerged from quantum optics, partly from theoretical computer science.

Today's fields of interest among quantum optics researchers include parametric down-conversion, parametric oscillation, even shorter (attosecond) light pulses, use of quantum optics for quantum information, manipulation of single atoms, Bose-Einstein condensates, their application, and how to manipulate them (a sub-field often called atom optics), coherent perfect absorbers, and much more.

Research into quantum optics that aims to bring photons into use for information transfer and computation is now often called photonics to emphasize the claim that photons and photonics will take the role that electrons and electronics now have.

Concepts of quantum optics

According to quantum theory, light may be considered not only as an electro-magnetic wave but also as a "stream" of particles called photons which travel with *c*, the vacuum speed of light. These particles should not be considered to be classical billiard balls, but as quantum mechanical particles described by a wavefunction spread over a finite region.

Each particle carries one quantum of energy equal to *hf*, where h is Planck's constant and f is the frequency of the light. The postulation of the quantization of light by Max Planck in 1899 and the discovery of the general validity of this idea in Albert Einstein's 1905 explanation of the photoelectric effect soon led physicists to realize the possibility of population inversion and the possibility of the laser.

This kind of use of statistical mechanics is the fundament of most concepts of quantum optics: Light is described in terms of field operators for creation and annihilation of photons—i.e. in the language of quantum electrodynamics.

A frequently encountered state of the light field is the coherent state as introduced by George Sudarshan in 1963. This state, which can be used to approximately describe the output of a single-frequency laser well above the laser threshold, exhibits Poissonian photon number statistics. Via certain nonlinear interactions, a coherent state can be transformed into a squeezed coherent state, which can exhibit super- or sub- Poissonean photon statistics. Such light is called squeezed light. Other important quantum aspects are related to correlations of photon statistics between different beams. For example, parametric nonlinear processes can generate so-called twin beams, where ideally each photon of one beam is associated with a photon in the other beam.

Atoms are considered as quantum mechanical oscillators with a discrete energy spectrum with the transitions between the energy eigenstates being driven by the absorption or emission of light according to Einstein's theory with the oscillator strength depending on the quantum numbers of the states.

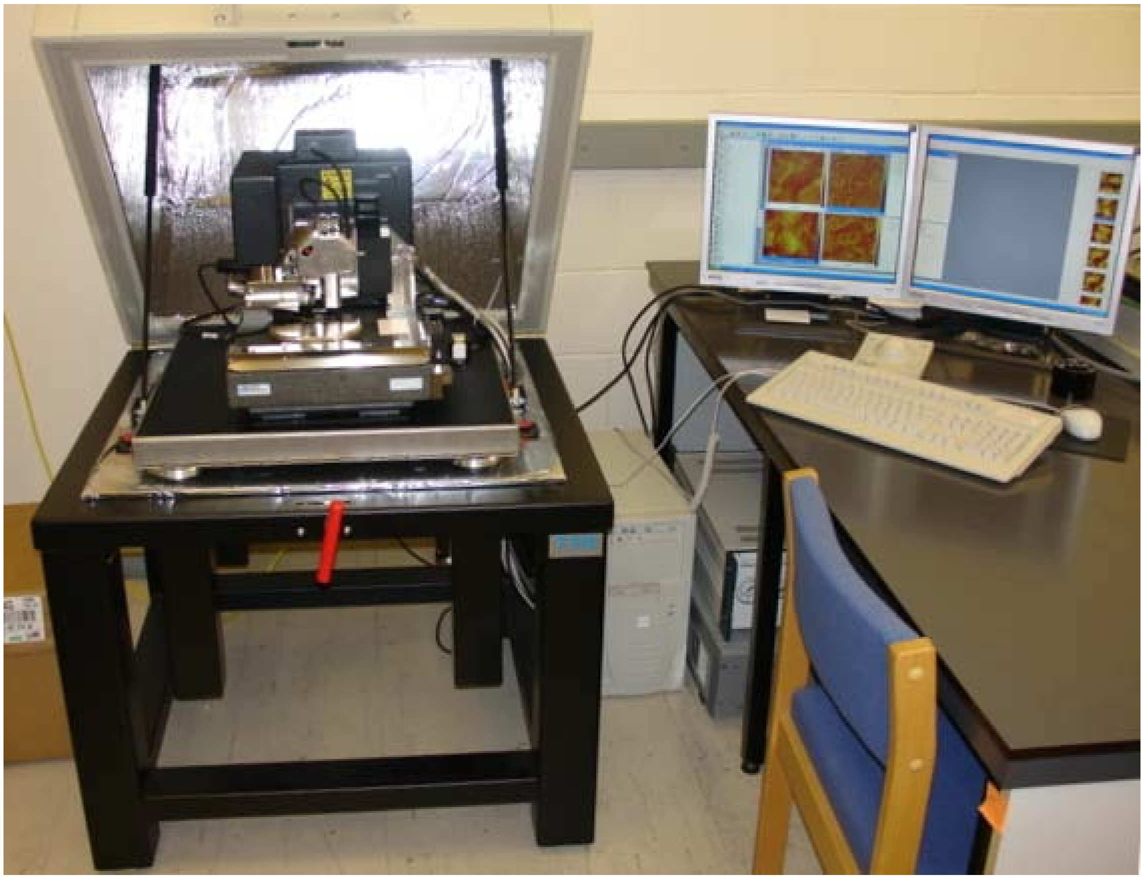
For solid state matter one uses the energy band models of solid state physics. This is important as understanding how light is detected (typically by a solid-state device that absorbs it) is crucial for understanding experiments.

Quantum electronics

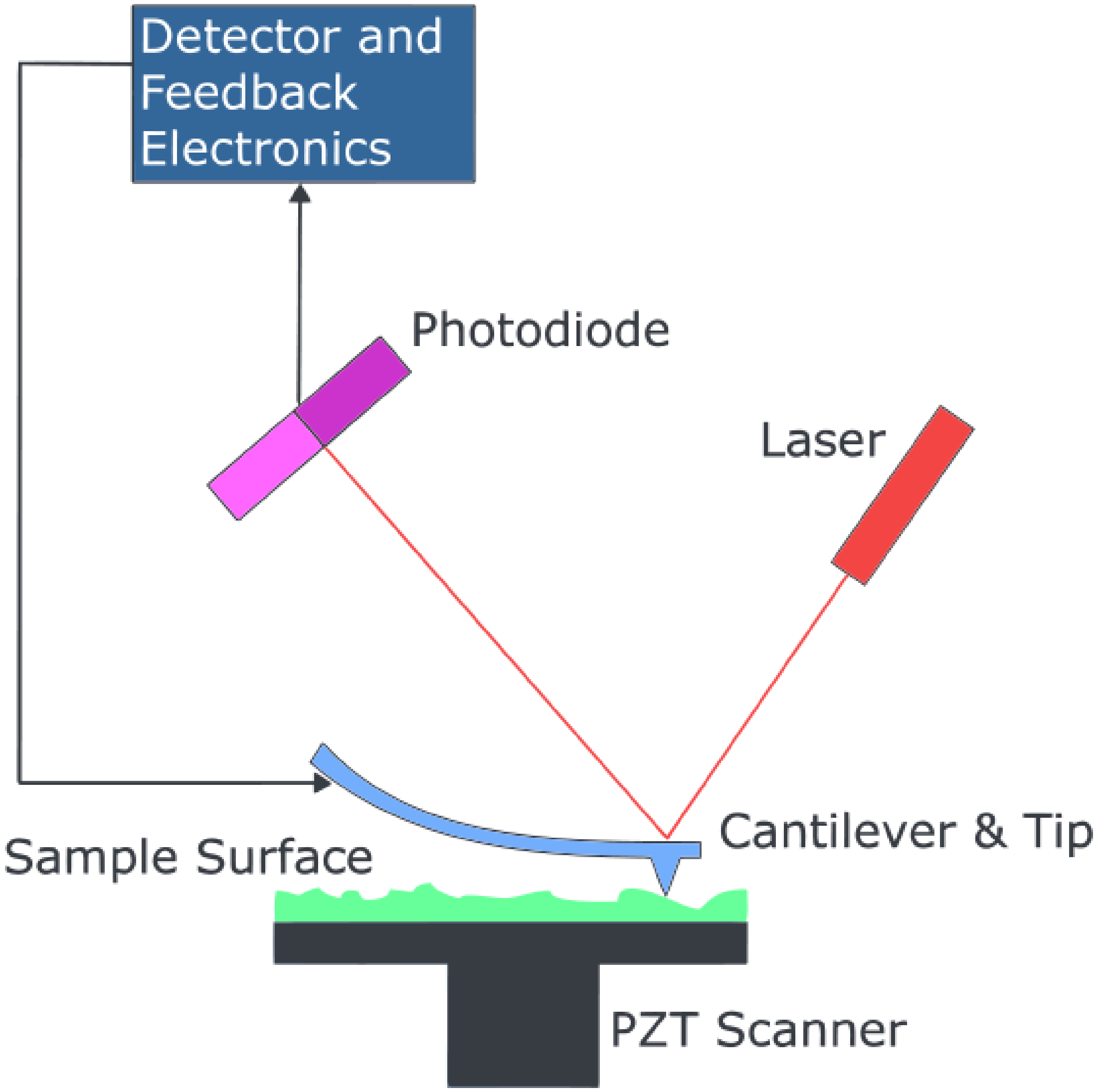
This term was used for the area of physics dealing with the effects of quantum mechanics on the behavior of electrons in matter, and their interactions with photons. It is today rarely considered a sub-field in its own right, as it has been absorbed by other fields. Solid state physics regularly takes quantum mechanics into account, and is usually concerned with electrons. Specific application to electronics is researched within semiconductor physics. The term also encompassed the basic processes of laser operation where photons are interacting with electrons: absorption, spontaneous emission, and stimulated emission. The term was mainly used between the 1950s and the 1970s. Today, the research output of this field is mainly used in quantum optics, especially for the part of it that draws not from atomic physics but from solid-state physics. Its usage overlapped Quantum Hall effect and Quantum cellular automata.

Chapter 12

Atomic Force Microscopy



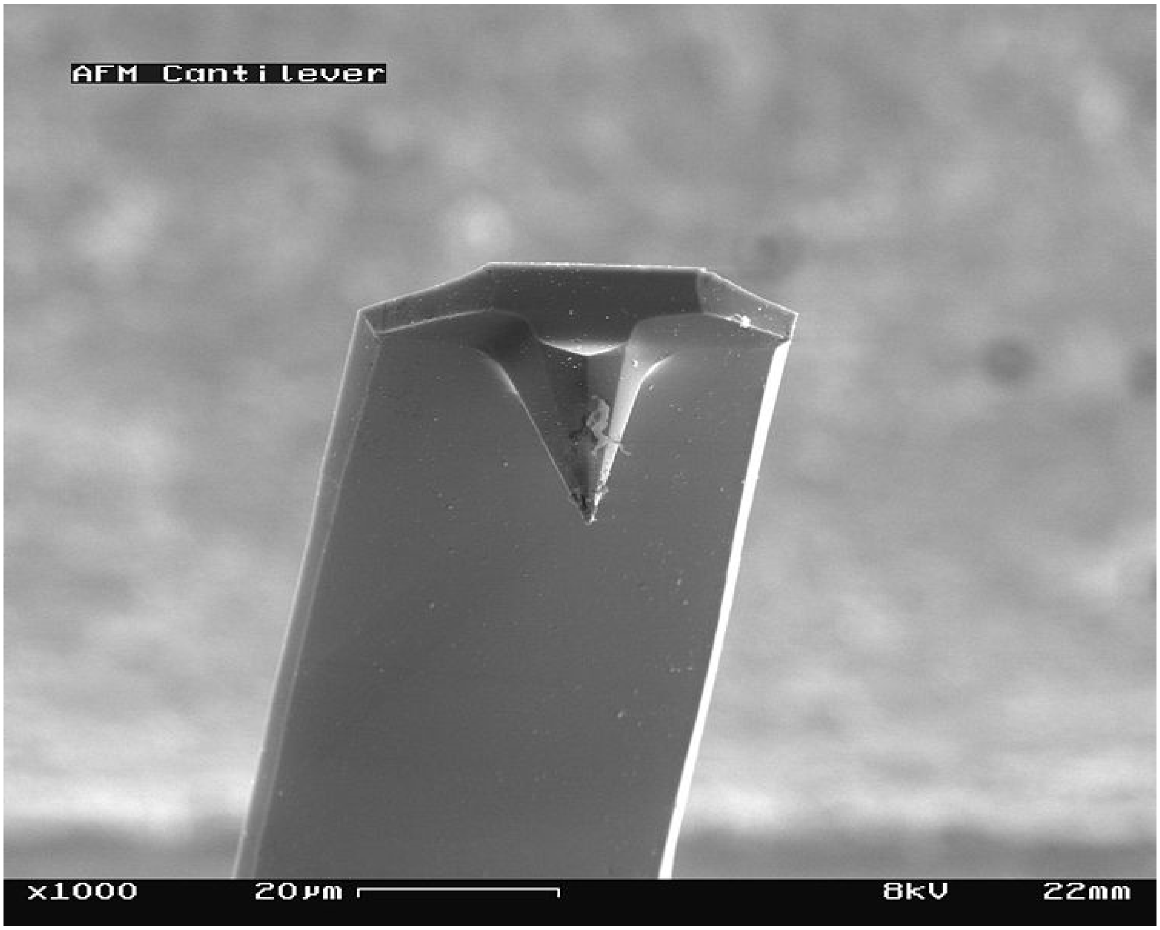
A commercial AFM setup



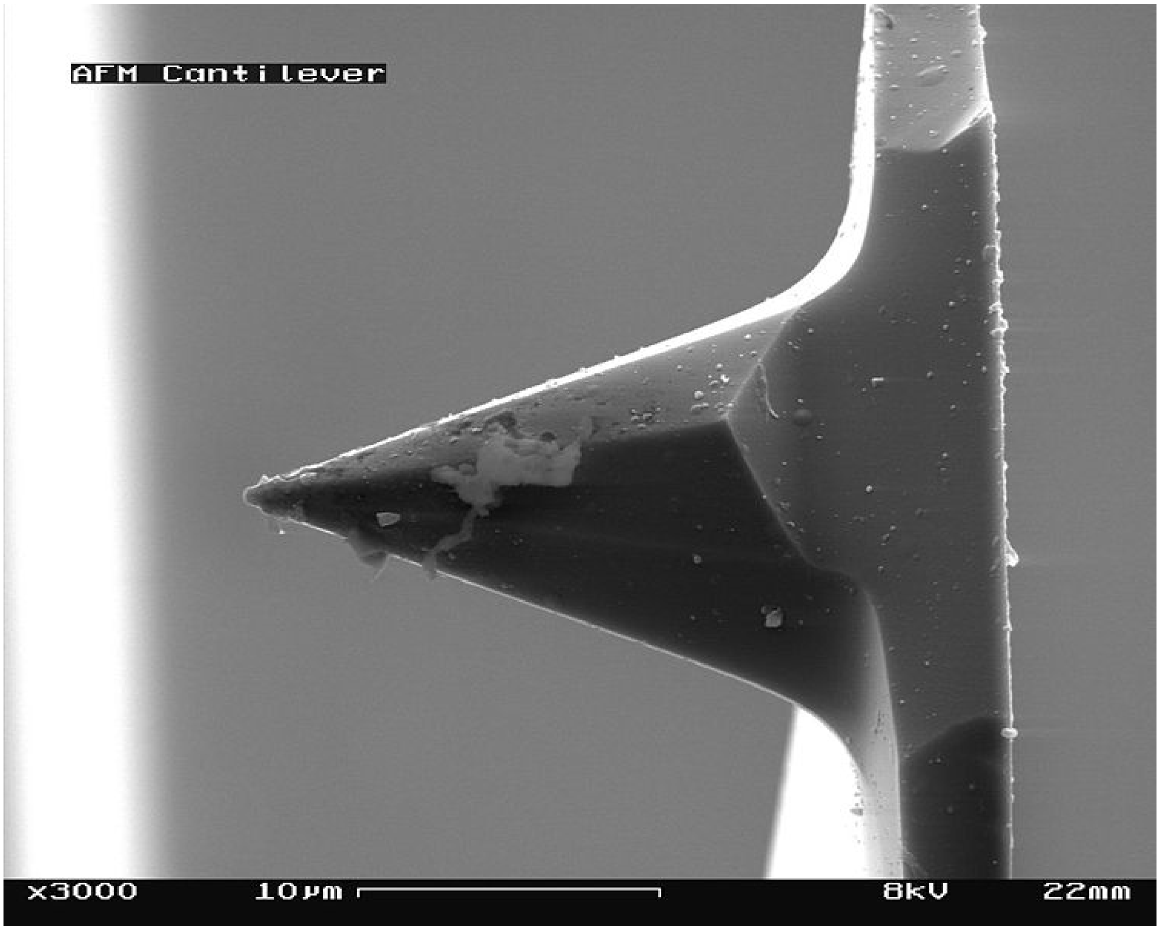
Block diagram of atomic force microscope

Atomic force microscopy (AFM) or scanning force microscopy (SFM) is a very high-resolution type of scanning probe microscopy, with demonstrated resolution on the order of fractions of a nanometer, more than 1000 times better than the optical diffraction limit. The precursor to the AFM, the scanning tunneling microscope, was developed by Gerd Binnig and Heinrich Rohrer in the early 1980s at IBM Research - Zurich, a development that earned them the Nobel Prize for Physics in 1986. Binnig, Quate and Gerber invented the first atomic force microscope (also abbreviated as AFM) in 1986. The first commercially available atomic force microscope was introduced in 1989. The AFM is one of the foremost tools for imaging, measuring, and manipulating matter at the nanoscale. The information is gathered by "feeling" the surface with a mechanical probe. Piezoelectric elements that facilitate tiny but accurate and precise movements on (electronic) command enable the very precise scanning. In some variations, electric potentials can also be scanned using conducting cantilevers. In newer more advanced versions, currents can even be passed through the tip to probe the electrical conductivity or transport of the underlying surface, but this is much more challenging with very few research groups reporting reliable data.

Basic principles

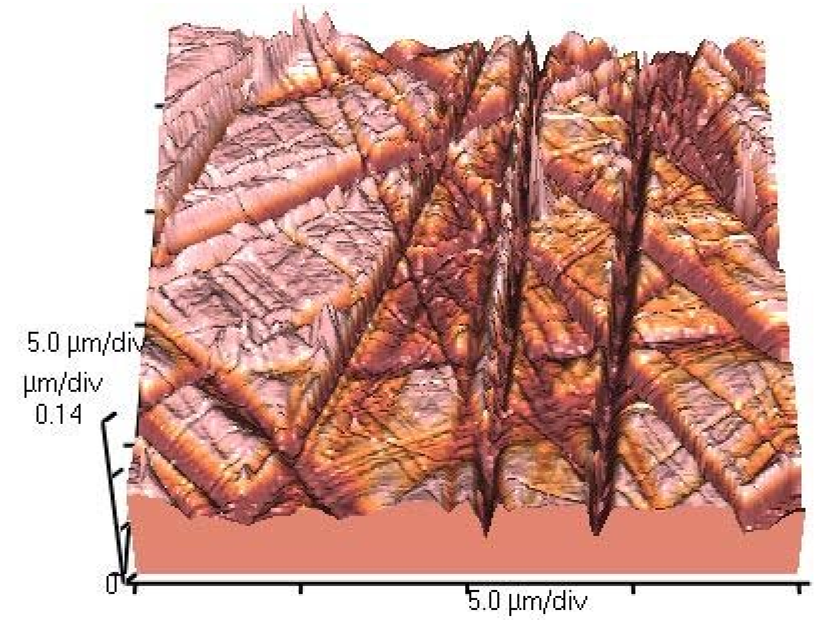


Electron micrograph of a used AFM cantilever image width ~100 micrometers



and ~30 micrometers

The AFM consists of a cantilever with a sharp tip (probe) at its end that is used to scan the specimen surface. The cantilever is typically silicon or silicon nitride with a tip radius of curvature on the order of nanometers. When the tip is brought into proximity of a sample surface, forces between the tip and the sample lead to a deflection of the cantilever according to Hooke's law. Depending on the situation, forces that are measured in AFM include mechanical contact force, van der Waals forces, capillary forces, chemical bonding, electrostatic forces, magnetic forces, Casimir forces, solvation forces, etc. Along with force, additional quantities may simultaneously be measured through the use of specialized types of probe. Typically, the deflection is measured using a laser spot reflected from the top surface of the cantilever into an array of photodiodes. Other methods that are used include optical interferometry, capacitive sensing or piezoresistive AFM cantilevers. These cantilevers are fabricated with piezoresistive elements that act as a strain gauge. Using a Wheatstone bridge, strain in the AFM cantilever due to deflection can be measured, but this method is not as sensitive as laser deflection or interferometry.



Atomic force microscope topographical scan of a glass surface. The micro and nano-scale features of the glass can be observed, portraying the roughness of the material. The image space is (x,y,z) = (20um x 20um x 420nm).

If the tip was scanned at a constant height, a risk would exist that the tip collides with the surface, causing damage. Hence, in most cases a feedback mechanism is employed to adjust the tip-to-sample distance to maintain a constant force between the tip and the sample. Traditionally, the sample is mounted on a piezoelectric tube, that can move the sample in the *z* direction for maintaining a constant force, and the *x* and *y* directions for scanning the sample. Alternatively a 'tripod' configuration of three piezo crystals may be employed, with each responsible for scanning in the x,y and z directions. This eliminates some of the distortion effects seen with a tube scanner. In newer designs, the tip is mounted on a vertical piezo scanner while the sample is being scanned in X and Y using another piezo block. The resulting map of the area $z = f(x,y)$ represents the topography of the sample.

The AFM can be operated in a number of modes, depending on the application. In general, possible imaging modes are divided into static (also called *contact*) modes and a variety of dynamic (or non-contact) modes where the cantilever is vibrated.

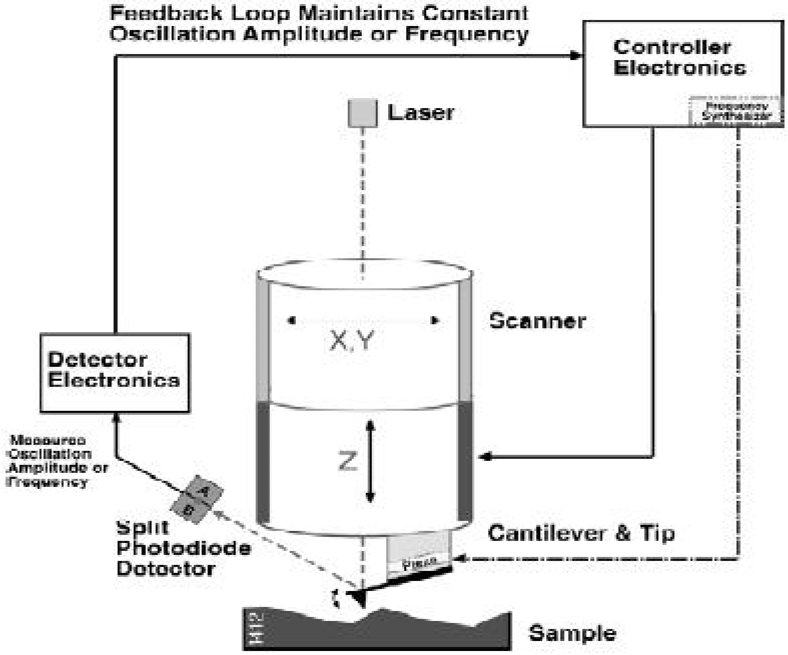
Imaging modes

The primary modes of operation for an AFM are static mode and dynamic mode. In static mode, the cantilever is "dragged" across the surface of the sample and the contours of the surface are measured directly using the deflection of the cantilever. In the dynamic mode, the cantilever is externally oscillated at or close to its fundamental resonance frequency or a harmonic. The oscillation amplitude, phase and resonance frequency are modified by tip-sample interaction forces. These changes in oscillation with respect to the external reference oscillation provide information about the sample's characteristics.

Contact mode

In the static mode operation, the static tip deflection is used as a feedback signal. Because the measurement of a static signal is prone to noise and drift, low stiffness cantilevers are used to boost the deflection signal. However, close to the surface of the sample, attractive forces can be quite strong, causing the tip to "snap-in" to the surface. Thus static mode AFM is almost always done in contact where the overall force is repulsive. Consequently, this technique is typically called "contact mode". In contact mode, the force between the tip and the surface is kept constant during scanning by maintaining a constant deflection.

Non-contact mode



AFM - non-contact mode

In this mode, the tip of the cantilever does not contact the sample surface. The cantilever is instead oscillated at a frequency slightly above its resonant frequency where the amplitude of oscillation is typically a few nanometers (<10 nm). The van der Waals forces, which are strongest from 1 nm to 10 nm above the surface, or any other long range force which extends above the surface acts to decrease the resonance frequency of the cantilever. This decrease in resonant frequency combined with the feedback loop system maintains a constant oscillation amplitude or frequency by adjusting the average tip-to-sample distance. Measuring the tip-to-sample distance at each (x,y) data point allows the scanning software to construct a topographic image of the sample surface.

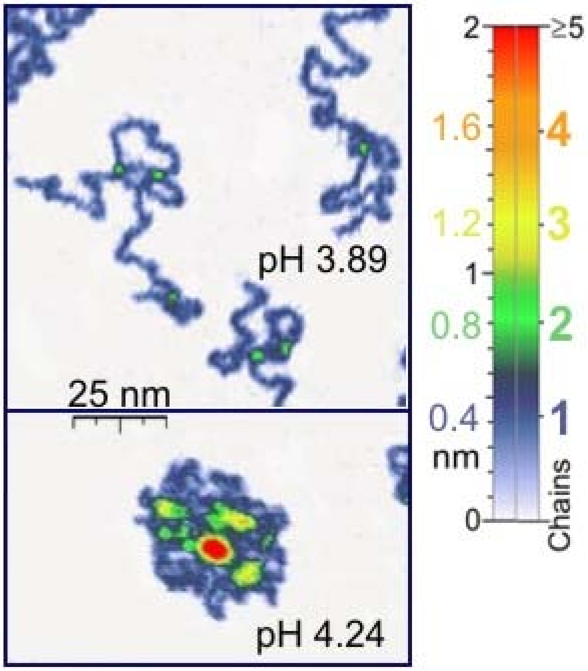
Non-contact mode AFM does not suffer from tip or sample degradation effects that are sometimes observed after taking numerous scans with contact AFM. This makes non-contact AFM preferable to contact AFM for measuring soft samples. In the case of rigid samples, contact and non-contact images may look the same. However, if a few monolayers of adsorbed fluid are lying on the surface of a rigid sample, the images may look quite different. An AFM operating in contact mode will penetrate the liquid layer to image the underlying surface, whereas in non-contact mode an AFM will oscillate above the adsorbed fluid layer to image both the liquid and surface.

Schemes for dynamic mode operation include frequency modulation and the more common amplitude modulation. In frequency modulation, changes in the oscillation frequency provide information about tip-sample interactions. Frequency can be measured with very high sensitivity and thus the frequency modulation mode allows for the use of very stiff cantilevers. Stiff cantilevers provide stability very close to the surface and, as a result, this technique was the first AFM technique to provide true atomic resolution in ultra-high vacuum conditions.

In amplitude modulation, changes in the oscillation amplitude or phase provide the feedback signal for imaging. In amplitude modulation, changes in the phase of oscillation can be used to discriminate between different types of materials on the surface. Amplitude modulation can be operated either in the non-contact or in the intermittent contact regime. In dynamic contact mode, the cantilever is oscillated such that the separation distance between the cantilever tip and the sample surface is modulated.

Amplitude modulation has also been used in the non-contact regime to image with atomic resolution by using very stiff cantilevers and small amplitudes in an ultra-high vacuum environment.

Tapping mode



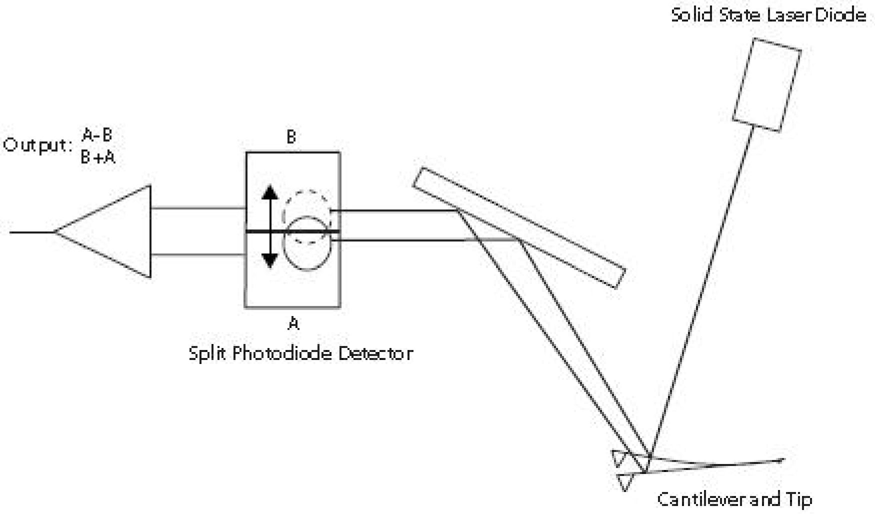
Single polymer chains (0.4 nm thick) recorded in a tapping mode under aqueous media with different pH.

In ambient conditions, most samples develop a liquid meniscus layer. Because of this, keeping the probe tip close enough to the sample for short-range forces to become detectable while preventing the tip from sticking to the surface presents a major problem for non-contact dynamic mode in ambient conditions. Dynamic contact mode (also called intermittent contact or tapping mode) was developed to bypass this problem.

In *tapping mode*, the cantilever is driven to oscillate up and down at near its resonance frequency by a small piezoelectric element mounted in the AFM tip holder similar to non-contact mode. However, the amplitude of this oscillation is greater than 10 nm, typically 100 to 200 nm. Due to the interaction of forces acting on the cantilever when the tip comes close to the surface, Van der Waals force, dipole-dipole interaction, electrostatic forces, etc. cause the amplitude of this oscillation to decrease as the tip gets closer to the sample. An electronic servo uses the piezoelectric actuator to control the height of the cantilever above the sample. The servo adjusts the height to maintain a set cantilever oscillation amplitude as the cantilever is scanned over the sample. A *tapping AFM* image is therefore produced by imaging the force of the intermittent contacts of the tip with the sample surface.

This method of "tapping" lessens the damage done to the surface and the tip compared to the amount done in contact mode. Tapping mode is gentle enough even for the visualization of supported lipid bilayers or adsorbed single polymer molecules (for instance, 0.4 nm thick chains of synthetic polyelectrolytes) under liquid medium. With proper scanning parameters, the conformation of single molecules can remain unchanged for hours.

AFM cantilever deflection measurement



AFM beam deflection detection

Laser light from a solid state diode is reflected off the back of the cantilever and collected by a position sensitive detector (PSD) consisting of two closely spaced photodiodes whose output signal is collected by a differential amplifier. Angular displacement of the cantilever results in one photodiode collecting more light than the other photodiode, producing an output signal (the difference between the photodiode signals normalized by their sum) which is proportional to the deflection of the cantilever. It detects cantilever deflections <10 nm (thermal noise limited). A long beam path (several centimeters) amplifies changes in beam angle.

Force spectroscopy

Another major application of AFM (besides imaging) is force spectroscopy, the direct measurement of tip-sample interaction forces as a function of the gap between the tip and sample (the result of this measurement is called a force-distance curve). For this method, the AFM tip is extended towards and retracted from the surface as the deflection of the cantilever is monitored as a function of piezoelectric displacement. These measurements have been used to measure nanoscale contacts, atomic bonding, Van der Waals forces, and Casimir forces, dissolution forces in liquids and single molecule stretching and rupture forces. Furthermore, AFM was used to measure, in an aqueous environment, the dispersion force due to polymer adsorbed on the substrate. Forces of the order of a few piconewtons can now be routinely measured with a vertical distance resolution of better than 0.1 nanometers. Force spectroscopy can be performed with either static or dynamic modes. In dynamic modes, information about the cantilever vibration is monitored in addition to the static deflection.

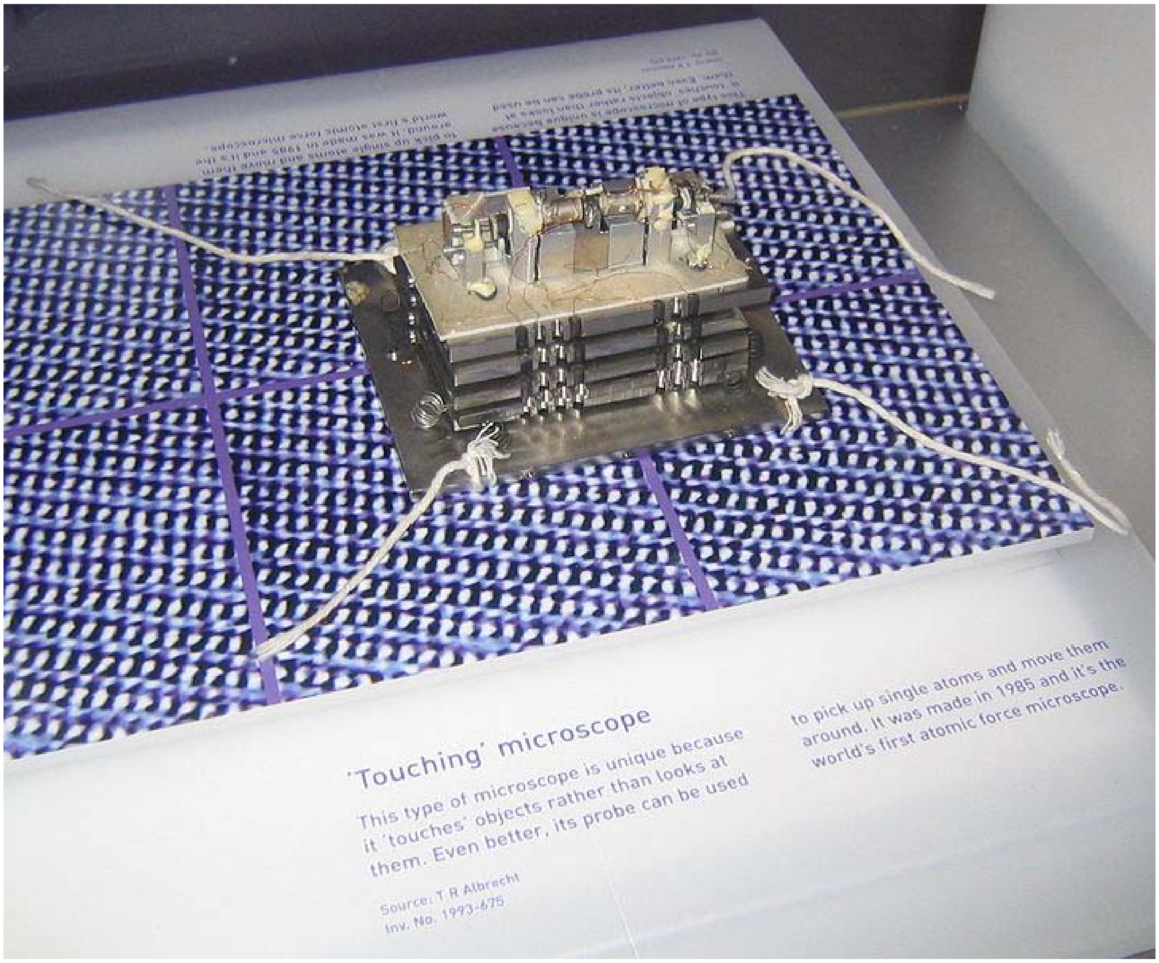
Problems with the technique include no direct measurement of the tip-sample separation and the common need for low stiffness cantilevers which tend to 'snap' to the surface. The snap-in can be reduced by measuring in liquids or by using stiffer cantilevers, but in the latter case a more sensitive deflection sensor is needed. By applying a small dither to the tip, the stiffness (force gradient) of the bond can be measured as well.

Identification of individual surface atoms

The AFM can be used to image and manipulate atoms and structures on a variety of surfaces. The atom at the apex of the tip "senses" individual atoms on the underlying surface when it forms incipient chemical bonds with each atom. Because these chemical interactions subtly alter the tip's vibration frequency, they can be detected and mapped. This principle was used to distinguish between atoms of silicon, tin and lead on an alloy surface, by comparing these 'atomic fingerprints' to values obtained from large-scale density functional theory (DFT) simulations.

The trick is to first measure these forces precisely for each type of atom expected in the sample, and then to compare with forces given by DFT simulations. The team found that the tip interacted most strongly with silicon atoms, and interacted 23% and 41% less strongly with tin and lead atoms, respectively. Thus, each different type of atom can be identified in the matrix as the tip is moved across the surface.

Advantages and disadvantages


'Touching' microscope
This type of microscope is unique because it 'touches' objects rather than looks at them. Even better, its probe can be used
to pick up single atoms and move them around. It was made in 1985 and it's the world's first atomic force microscope.
Source: T R Albrecht
Inv. No. 1993-675

The first atomic force microscope

Just like any other tool, an AFM's usefulness has limitations. When determining whether or not analyzing a sample with an AFM is appropriate, there are various advantages and disadvantages that must be considered.

Advantages

AFM has several advantages over the scanning electron microscope (SEM). Unlike the electron microscope which provides a two-dimensional projection or a two-dimensional image of a sample, the AFM provides a three-dimensional surface profile. Additionally, samples viewed by AFM do not require any special treatments (such as metal/carbon coatings) that would irreversibly change or damage the sample. While an electron microscope needs an expensive vacuum environment for proper operation, most AFM modes can work perfectly well in ambient air or even a liquid environment. This makes it possible to study biological macromolecules and even living organisms. In principle, AFM can provide higher resolution than SEM. It has been shown to give true atomic resolution in ultra-high vacuum (UHV) and, more recently, in liquid environments. High resolution AFM is comparable in resolution to scanning tunneling microscopy and transmission electron microscopy.

Disadvantages

A disadvantage of AFM compared with the scanning electron microscope (SEM) is the single scan image size. In one pass, the SEM can image an area on the order of square millimeters with a depth of field on the order of millimeters. Whereas the AFM can only image a maximum height on the order of 10-20 micrometers and a maximum scanning area of about 150×150 micrometers. One method of improving the scanned area size for AFM is by using parallel probes in a fashion similar to that of millipede data storage.

The scanning speed of an AFM is also a limitation. Traditionally, an AFM cannot scan images as fast as a SEM, requiring several minutes for a typical scan, while a SEM is capable of scanning at near real-time, although at relatively low quality. The relatively slow rate of scanning during AFM imaging often leads to thermal drift in the image making the AFM microscope less suited for measuring accurate distances between topographical features on the image. However, several fast-acting designs were suggested to increase microscope scanning productivity including what is being termed videoAFM (reasonable quality images are being obtained with videoAFM at video rate: faster than the average SEM). To eliminate image distortions induced by thermal drift, several methods have been introduced.

AFM images can also be affected by hysteresis of the piezoelectric material and cross-talk between the *x*, *y*, *z* axes that may require software enhancement and filtering. Such filtering could "flatten" out real topographical features. However, newer AFMs utilize closed-loop scanners which practically eliminate these problems. Some AFMs also use separated orthogonal scanners (as opposed to a single tube) which also serve to eliminate part of the cross-talk problems.

As with any other imaging technique, there is the possibility of image artifacts, which could be induced by an unsuitable tip, a poor operating environment, or even by the sample itself. These image artifacts are unavoidable however, their occurrence and effect on results can be reduced through various methods.

Due to the nature of AFM probes, they cannot normally measure steep walls or overhangs. Specially made cantilevers and AFMs can be used to modulate the probe sideways as well as up and down (as with dynamic contact and non-contact modes) to measure sidewalls, at the cost of more expensive cantilevers, lower lateral resolution and additional artifacts.

Piezoelectric scanners

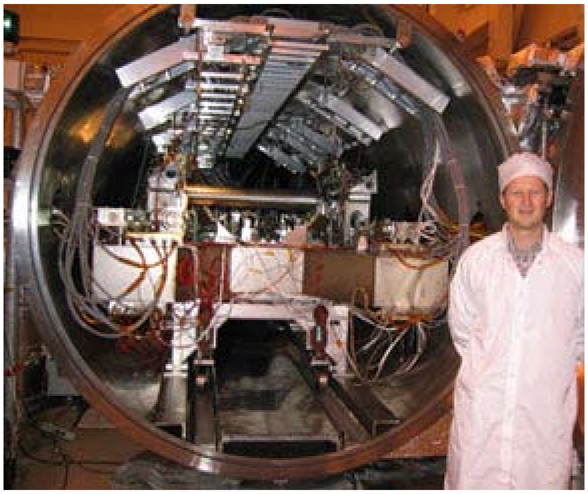
AFM scanners are made from piezoelectric material, which expands and contracts proportionally to an applied voltage. Whether they elongate or contract depends upon the polarity of the voltage applied. The scanner is constructed by combining independently operated piezo electrodes for X, Y, and Z into a single tube, forming a scanner which can manipulate samples and probes with extreme precision in 3 dimensions.

Scanners are characterized by their sensitivity which is the ratio of piezo movement to piezo voltage, i.e., by how much the piezo material extends or contracts per applied volt. Because of differences in material or size, the sensitivity varies from scanner to scanner. Sensitivity varies non-linearly with respect to scan size. Piezo scanners exhibit more sensitivity at the end than at the beginning of a scan. This causes the forward and reverse scans to behave differently and display hysteresis between the two scan directions. This can be corrected by applying a non-linear voltage to the piezo electrodes to cause linear scanner movement and calibrating the scanner accordingly.

The sensitivity of piezoelectric materials decreases exponentially with time. This causes most of the change in sensitivity to occur in the initial stages of the scanner's life. Piezoelectric scanners are run for approximately 48 hours before they are shipped from the factory so that they are past the point where they may have large changes in sensitivity. As the scanner ages, the sensitivity will change less with time and the scanner would seldom require recalibration.

Chapter 13

Metrology



A scientist stands in front of a microarcsecond (1 millionth of 1 arcsecond or 1 millionth of 1/3600 degree) testbed.

Metrology is the science of measurement. Metrology includes all theoretical and practical aspects of measurement. The word comes from Greek μέτρον (*metron*), "measure" + "λόγος" (*logos*), amongst others meaning "speech, oration, discourse, quote, study, calculation, reason". In Ancient Greek the term μετρολογία (*metrologia*) meant "theory of ratios".

Introduction

Metrology is defined by the *International Bureau of Weights and Measures* (BIPM) as "the science of measurement, embracing both experimental and theoretical determinations at any level of uncertainty in any field of science and technology." The *ontology* and international vocabulary of metrology (VIM) is maintained by the International Organisation for Standardisation.

Metrology is a very broad field and may be divided into three subfields:

Subfield	Definition
Scientific or fundamental metrology	concerns the establishment of *quantity systems*, unit systems, *units of measurement*, the development of new measurement methods, realisation of measurement standards and the transfer of traceability from these standards to users in society.
Applied or industrial metrology	concerns the application of measurement science to manufacturing and other processes and their use in society, ensuring the suitability of measurement instruments, their calibration and quality control of measurements.
Legal metrology	concerns regulatory requirements of measurements and measuring instruments for the protection of health, public safety, the environment, enabling taxation, protection of consumers and fair trade.

A core concept in metrology is (metrological) traceability, defined as "the property of the result of a measurement or the value of a standard whereby it can be related to stated references, usually national or international standards, through an unbroken chain of comparisons, all having stated uncertainties." The level of traceability establishes the level of comparability of the measurement: whether the result of a measurement can be compared to the previous one, a measurement result a year ago, or to the result of a measurement performed anywhere else in the world.

Traceability is most often obtained by calibration, establishing the relation between the indication of a measuring instrument and the value of a measurement standard. These standards are usually coordinated by national metrological institutes: National Institute of Standards and Technology, National Physical Laboratory, UK, Physikalisch-Technische Bundesanstalt, etc.

Tracebility, accuracy, precision, systematic bias, evaluation of measurement uncertainty are critical parts of a quality management system.

Basics

Mistakes can make measurements and counts incorrect. Even if there are no mistakes, nearly all measurements are still inexact. The term 'error' is reserved for that inexactness, also called measurement uncertainty. Among the few exact measurements are:

- The absence of the quantity being measured, such as a voltmeter with its leads shorted together: the meter should read zero exactly.
- Measurement of an accepted constant under qualifying conditions, such as the triple point of pure water: the thermometer should read 273.16 kelvin (0.01 degrees Celsius, 32.018 degrees Fahrenheit) when qualified equipment is used correctly.

- Self-checking ratio metric measurements, such as a potentiometer: the ratio in between steps is independently adjusted and verified to be beyond influential inexactness.

All other measurements either have to be checked to be sufficiently correct or left to chance. Metrology is the science that establishes the correctness of specific measurement situations. This is done by anticipating and allowing for both mistakes and error. The precise distinction between measurement error and mistakes is not settled and varies by country. Repeatability and reproducibility studies help quantify the precision: one common method is an ANOVA Gauge R&R study.

Calibration is the process where metrology is applied to measurement equipment and processes to ensure conformity with a known standard of measurement, usually traceable to a national standards board.

Society

Sufficiently correct measurements are essential to commerce. About nine out of every ten people working in metrology specialize in commercial measurement, most at the technician level. Correct measurements are beneficial to manufacturing, but other methods are available and sometimes are more appropriate.

Metrology has thrived at the interface between science and manufacturing. Aerospace, commercial nuclear power, medicine, medical devices and semiconductors rely on metrology to translate theoretical science into mass produced reality.

The basic concepts of metrology appear simple on the surface, and metrology is rarely taught in a systematic manner above the technician level. Within most businesses, metrology core beliefs such as recording all setups and observations for possible future reference are opposed to the general business practice of minimizing recordkeeping to limit litigation effects.

Applied metrology

Metrology laboratories are places where both metrology and calibration work are performed. Calibration laboratories generally specialize in calibration work only.

Both metrology and calibration laboratories must isolate the work performed from influences that might affect the work. Temperature, humidity, vibration, electrical power supply, radiated energy and other influences are often controlled. Generally, it is the rate of change or instability that is more detrimental than whatever value prevails.

Calibration technicians execute calibration work. In large organizations, the work is further divided into three groups:

Group	Definition
Set-up people	arrange the equipment needed for calibration and verify that it works correctly.
Operators	execute the calibration procedures and collect data.
Tear-down people	dismantle set-ups, check the components for damage and then put the components into a stored state. This is the entry-level position for people who didn't start in the equipment warehouse or transportation functions

Alternately, the technicians can be divided by major discipline areas: physical, dimensional, electrical, RF, microwave and so on. But the principles are the same regardless of the equipment.

Metrology technicians perform investigation work in addition to calibrations. They also apply proven principles to known situations and evaluate unexpected or contradictory results.

Specific education in metrology was formerly limited to sub-professional work. Most of the branches of the US Military train 'enlisted-grade' technicians to meet their specific needs.

Large industrial organizations also develop people who demonstrate aptitude in testing functions. When this is combined with an engineering degree, it qualifies the person as a metrology engineer. Over the last 15 years, Universities such as the University of North Carolina at Charlotte created a specific curriculum in metrology engineering. In England, metrology was part of the fifth year of some undergraduate engineering programmes.

Metrologists are people who perform metrology work at and above the technician levels, generally without the benefit or acknowledgement of a college degree.

The metrology and calibration work described above is always accompanied by documentation. The documentation can be divided into two types; one related to the task and the other related the administrative program. Task documentation includes calibration procedures and the data collected. Administrative program documentation includes equipment identification data, 'calibration certificates', calibration time interval information and 'as-found' or 'out-of-tolerance' notifications.

Administrative programs provide standardization of the metrology and calibration work and make it possible to independently verify that the work was performed. Generally, the administrative program is specific to the organization performing the work and addresses customer requirements. General administrative program specifications created by industry groups, such as the ANS (ANSI) Z540 series may also be covered in the administrative program. Other specifications created by the US Food and Drug Administration, US Federal Aviation Administration or other agencies would supplement or replace ANS Z540 for work performed in their domains. Often administrative programs can be as complicated and detailed as the measurement work itself.

An administrative program that has insufficient actual metrology or calibration capability is derisively referred to as a "lick and stick" program.

Standards

Standards are objects or ideas that are designated as being authoritative for some accepted reason. Whatever value they possess is useful for comparison to unknowns for the purpose of establishing or confirming an assigned value based on the standard. The design of this comparison process for measurements is metrology. The execution of measurement comparisons for the purpose of establishing the relationship between a standard and some other measuring device is calibration.

The ideal standard is independently reproducible without uncertainty. This is what the creators of the “meter” length standard were attempting to do in the 19th century when they defined a meter as one ten-millionth of the distance from the equator to one of the Earth’s poles. Later, it was learned that the Earth’s surface is an unreliable basis for a standard. The Earth is not spherical and it is constantly changing in shape. But the special alloy meter bars that were created and accepted in that time period standardized international length measurement until the 1950s. Careful calibrations allowed tolerances as small as 10 parts per million to be distributed and reproduced in metrology laboratories worldwide, regardless of whether the rest of the metric system was implemented and in spite of the shortfalls of the meter’s original basis.

Historical International Prototype Meter bars

Modern standards

Currently, only five independent units of measure are internationally recognized: temperature interval, linear distance, electrical current, frequency and mass. All measurements of all types are based on one or more of these independent units. Two supplemental independent units are also recognized internationally, both dealing with angle measurement.

For example, Ohm's law is a widely known concept in electrical study. Of the three units of measure involved, only current (ampere) is an independent unit. Voltage and resistance units are dependent on current units, as defined by Ohm's law.

In the United States, ASTM Standard Practice E 380,replaced by IEEE/ASTM SI10 , adapts independent unit of measure theory to practical measurement activity.

It is believed that each of independent units of measure will be defined in terms of the other four independent units eventually. Length (meter) and time (second) are already connected this way. If an accurate time base is available, then a length standard can be reproduced without a meter bar artifact, using the known constant speed of light. Lesser known is the relationship between the luminance (candela) and current (ampere). The candela is defined in terms of the watt, which in turn is derived from the ampere. This difficult to recreate standard is supplemented by an incandescent bulb design that is used as a secondary and transfer standard. These bulbs recreate the candela when a specific amount of current is applied.

The development of standards follows the needs of technology. As a result, some units of measure have much more resolution than others. The second is reproducible to 1 part in 10^{14}. As it became possible to measure time more precisely, solar time, believed to be a constant, proved to be very slightly irregular. This resulted in leap second adjustments to keep UTC synchronised with solar time.

Luminance (candela) can only be reproduced to 5% of reading despite having sensors that have accuracies of +/- 50 parts per million (0.005%) precision. This is due to the standard not being accurately reproducible.

Temperature (kelvin) is defined by agreed fixed points. These points are defined by the state changes of nearly pure materials, generally as they move from liquid to solid. Between these fixed points, Standard Platinum Resistance Thermometers (SPRTs), constructed a specified manner, are used to interpolate temperature values. This mosaic of approaches produces measurement uncertainty which is not uniform over the entire range of temperature measurement. Temperature measurement is coordinated by the International Practical Temperature Scale, maintained by the BIPM.

These non-commercial measurement details used to be academic curiosities. However, engineering, manufacturing and ordinary living now routinely challenge the limits of measurement.

Industry-specific standards

In addition to standards created by national and international standards organizations, many large and small industrial companies also define metrology standards and procedures to meet their particular needs for technically and economically competitive manufacturing. These standards and procedures, while drawing in part upon the national and international standards, also address the issues of what specific instrument technology will be used to measure each quantity, how often each quantity will be measured, and which definition of each quantity will be used as the basis for accomplishing the process control that their manufacturing and product specifications require. Industrial metrology standards include dynamic control plans, also known as “dimensional control plans”, or “DCPs”, for their products.

In industrial metrology, several issues beyond accuracy constrain the usability of metrology methods. These include

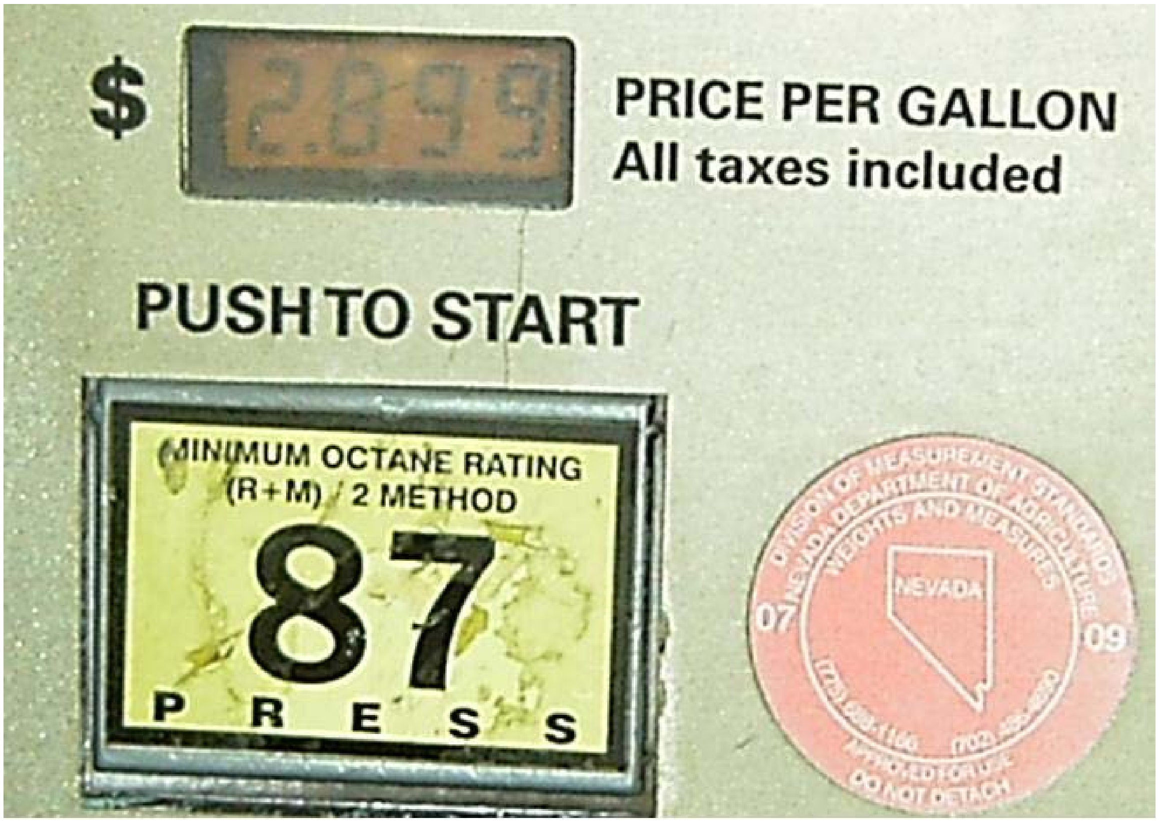
- The speed with which measurements can be accomplished on parts or surfaces in the process of manufacturing, which must match the TAKT Time of the production line.
- The completeness with which the manufactured part can be measured such as described in high-definition metrology,
- The ability of the measurement mechanism to operate reliably in a manufacturing plant environment considering temperature, vibration, dust, and a host of other potential hostile factors,
- The ability of the measurement results, as they are presented, to be assimilated by the manufacturing operators or automation in time to effectively control the manufacturing process variables, and
- The total financial cost of measuring each part.

National standards

Every country maintains its own metrology system. In the United States, the National Institute of Standards and Technology (NIST) plays the dual role of maintaining and furthering both commercial and scientific metrology. NIST does not enforce measurement accuracy directly.

The accuracy and traceability of commercial measurements is enforced per the laws of the individual states. Commercial measurement generally involves any material sold by any unit of measure. Some intuitive or obvious measurement is generally exempted, such as selling cloth on a cutting table that has a yardstick fastened to it. All counting-based transactions are generally exempt also. But each state has its own rules, responding to the accumulated concerns of the state residents.

Commercial metrology is also known as "weights and measures" and is essential to commerce of any kind above the pure barter level. Every state maintains its own weights and measures functionality with traceability to the national standards maintained by NIST. Large states further divide this effort by county, where a "Sealer" or other appointee is responsible for the validity of most common commercial measurements such as mass balances (scales) in grocery stores and gasoline pump measurements of volume. The sealer's staff and agents make periodic inspections to catch merchant cheaters, maintaining the integrity of commercial measurements.



Typical State Seal application.

Depending on the specific state, other state government agencies can be involved. For example, electricity watt-hour meters and water delivery flow meters are commonly monitored by the state's "public utilities commission" who enforces the measurement tolerances and traceability to NIST through the utility providers. Highway State Police and the State Highway Department generally run the commercial truck weight measurement programs for safety purposes and to minimize the damage to road surfaces that overloaded trucks cause. Nearly all states license weighmasters, weighmistresses, scale calibrators and other specialists involved in commercial measuring equipment maintenance.

The term "commercial metrology" is also used to describe calibration laboratories that are not owned by the companies they serve.

Scientific metrology addresses measurement phenomena not quantified in ordinary commerce, such as the test bed pictured at the beginning. Calibration laboratories that serve scientific metrology are regulated as businesses only. They may choose to have their work accredited by voluntary certification organizations based on customer desires, but there is no requirement to do so. Irresolvable disputes involving scientific metrology are generally settled in the civil court systems. Some federal government entities like the Federal Communications Commission and the Environmental Protection Administration are considered to be the final authority in their domains rather than the NIST. Disputes involving only metrology issues with those organizations probably would not be heard in any courts.

Historical development

Metrology has existed in some form or another since antiquity. The earliest forms of metrology were simply arbitrary standards set up by regional or local authorities, often based on practical measures such as the length of an arm. The earliest examples of these standardized measures are length, time, and weight. These standards were established in order to facilitate commerce and record human activity.

Little progress was made with regard to proto-metrology until various scientists, chemists, and physicists started making headway during the scientific revolution. With the advances in the sciences, the comparison of experiment to theory required a rational system of units, and something more closely resembling modern metrology began to come into being. The discovery of atoms, electricity, thermodynamics, and other fundamental scientific principles could be applied to standards of measurement, and many inventions made it easier to quantitatively or qualitatively assess physical properties, using the defined units of measurement established by science.

Metrology was thus one of the precursors to the Industrial Revolution, and was necessary for the implementation of mass production, equipment commonality, and assembly lines.

Modern metrology has its roots in the French Revolution, with the political motivation to harmonize units all over France and the concept of establishing units of measurement based on constants of nature, and thus making measurement units available "for all people, for all time". In this case deriving a unit of length from the dimensions of the Earth, and a unit of mass from a cube of water. The result was platinum standards for the meter and the kilogram established as the basis of the metric system on June 22, 1799. This further led to the creation of the Système International d'Unités, or the International System of Units. This system has gained unprecedented worldwide acceptance as definitions and standards of modern measurement units. Though not the official system of units of all nations, the definitions and specifications of SI are globally accepted and recognized. The SI is maintained under the auspices of the Metre Convention and its institutions, the General Conference on Weights and Measures, or CGPM, its executive branch the International Committee for Weights and Measures, or CIPM, and its technical institution the International Bureau of Weights and Measures, or BIPM.

As the authorities on SI, these organizations establish and promulgate the SI, with the ambition to be able to service all. This includes introducing new units, such as the relatively new unit, the mole, to encompass metrology in chemistry. These units are then established and maintained through various agencies in each country, and establish a hierarchy of measurement standards that can be traced back to the established standard unit, a concept known as metrological traceability. The U.S. agencies holding this responsibility are the National Institute of Standards and Technology (NIST) and the American National Standards Institute (ANSI).

The development of standards also does involve individual and small group achievements. In 1893, Edward Weston (chemist) and his company perfected his Saturated Standard Cell design, which allowed the volt to be reproduced to 1 part in ten to the fourth power directly. This advance made a huge practical difference at a critical moment in the development of modern electrical devices. Groupings of saturated cells, called banks, can still be found in some metrology and calibration laboratories today. Edward Weston did not pursue patents for his cell design. By doing this, his superior design quickly replaced similar but inferior patented devices worldwide without much discussion.

Mechanisms

At the base of metrology is the definition, realisation and dissemination of units of measurement. Physical or chemical properties are quantised by assigning a property value in some multiple of a measurement unit.

The basic 'lineage' of measurement standards are:

- The definition of a unit, based on some physical constant, such as absolute zero, the freezing point of water, etc.; or an agreed-upon arbitrary standard.
- The realisation of the unit by experimental methods and the scaling into multiples and submultiples, by establishment of primary standards. In some cases an approximation is used, when the realisation of the units is less precise than other methods of generating a scale of the quantity in question. This is presently the situation for the electrical units in the SI, where voltage and resistance are defined in terms of the ampere, but are used in practice from realisations based on the Josephson effect and the quantised Hall effect.
- the transfer of traceability from the primary standards to secondary and working standards. This is achieved by calibration.

Theoretically, metrology, as the science of measurement, attempts to validate the data obtained from test equipment. Though metrology is the science of measurement, in practical applications, it is the enforcement, verification and validation of predefined standards for:

Criterion	Definition
Accuracy	is the degree of exactness which the final product corresponds to the measurement standard.
Precision	refers to the ability of a measurement to be consistently reproduced.
Reliability	refers to the consistency of accurate results over consecutive measurements over time.
Traceability	refers to the ongoing validations that the measurement of the final product conforms to the original standard of measurement.

These standards can vary widely, but are often mandated by governments, agencies, and treaties such as the International Organization for Standardization, the Metre Convention, or the FDA. These agencies promulgate policies and regulations that standardize industries, countries, and streamline international trade, products, and measurements. Metrology is, at its core, an analysis of the uncertainty of individual measurements, and attempts to validate each measurement made with a given instrument, and the data obtained from it. The dissemination of traceability to consumers in society is often performed by a dedicated calibration laboratory with a recognized quality system in compliance with such standards. National laboratory accreditation schemes have been established to offer third-party assessment of such quality systems. A central requirement of these accreditations is documented traceability to national or international standards.

Some common standards include:

- ISO 17025:2005—General Requirements for Calibration Laboratories
- ISO 9000—Quality Systems Management
- ISO 14000—Environmental Management
- 21 CFR Part 210/211—FDA Regulations concerning GMP (Good Manufacturing Practices) Quality Systems
- 21 CFR Part 110—FDA Regulations concerning Food Industry GMP's.

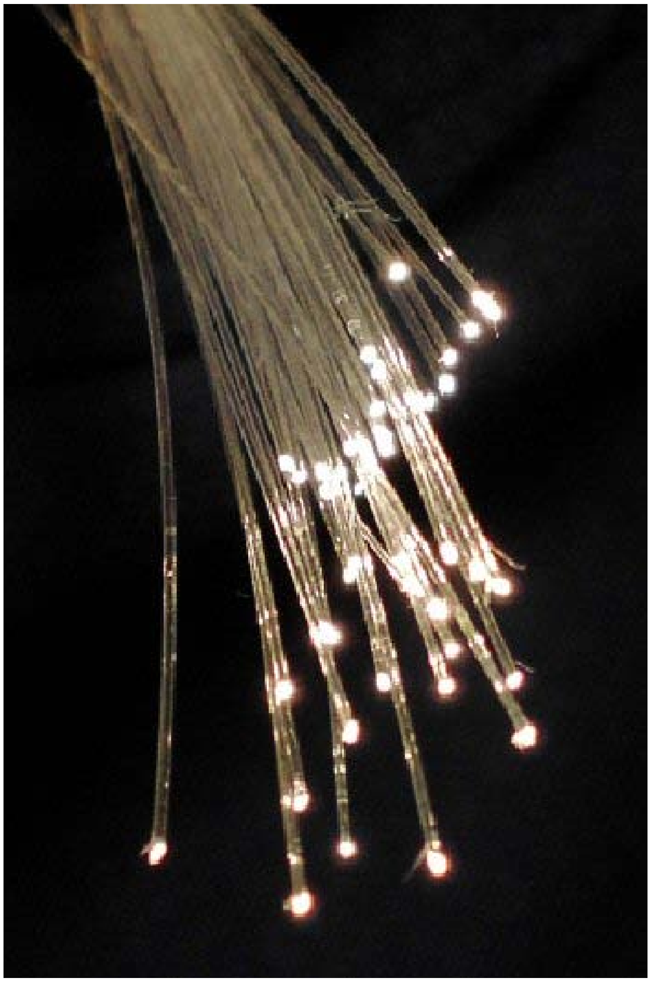
Time and frequency metrology

This area of metrology studies components and their characteristics, especially

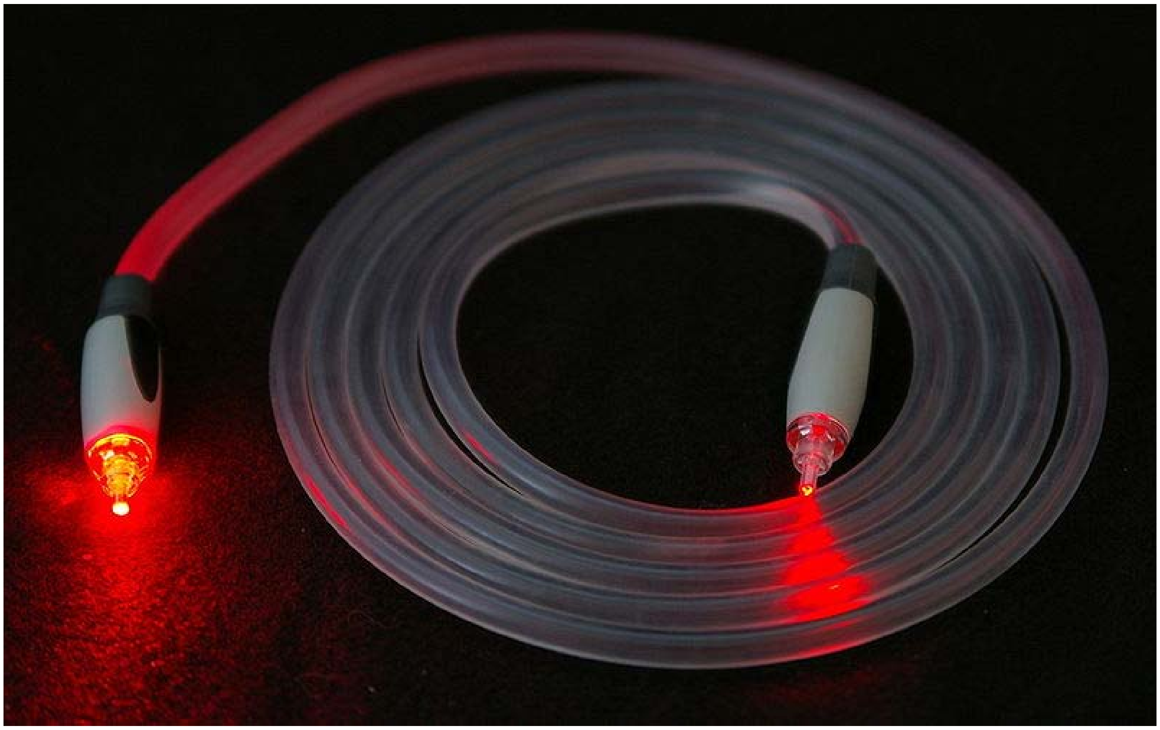
- frequency standards
- synthesizers
- oscillators
- digital clocks

Chapter 14

Optical Fiber



A bundle of optical fibers



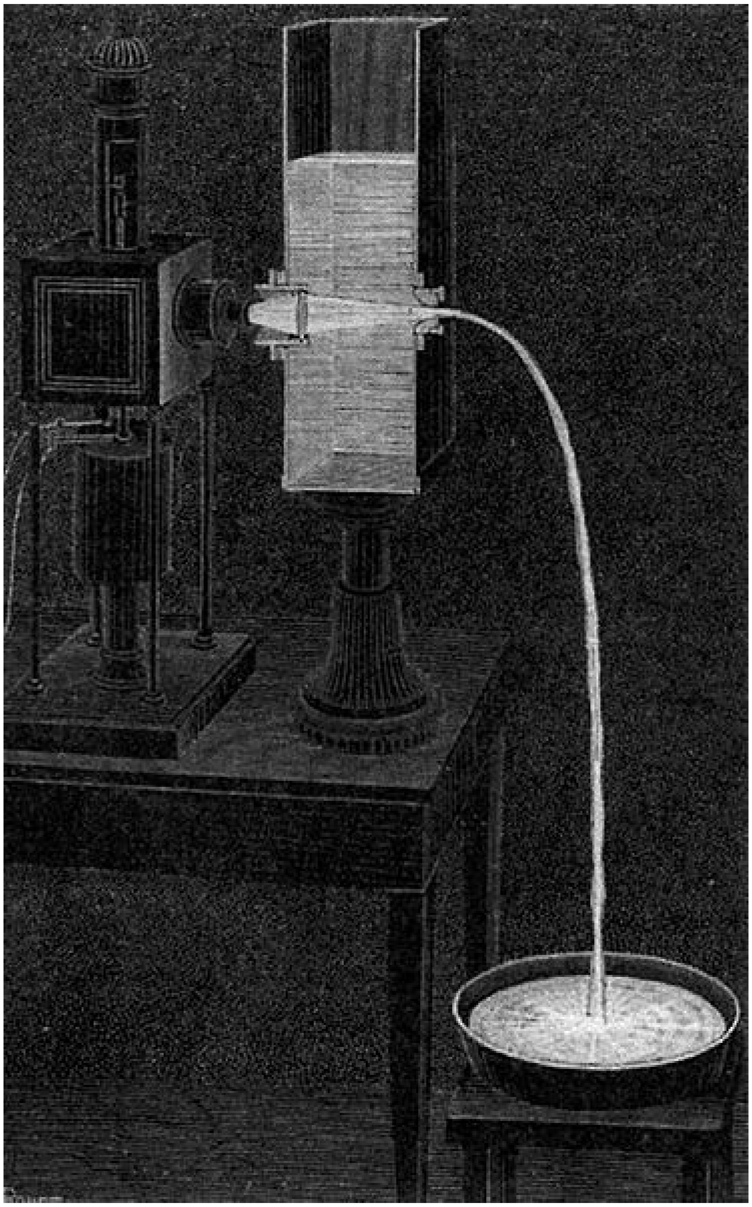
A TOSLINK fiber optic audio cable being illuminated at one end

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

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

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

History



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

Fiber optics, though used extensively in the modern world, is a fairly simple and old technology. Guiding of light by refraction, the principle that makes fiber optics possible, was first demonstrated by Daniel Colladon and Jacques Babinet in Paris in the early 1840s. John Tyndall included a demonstration of it in his public lectures in London a dozen years later. Tyndall also wrote about the property of total internal reflection in an introductory book about the nature of light in 1870: "When the light passes from air into water, the refracted ray is bent *towards* the perpendicular... When the ray passes from water to air it is bent *from* the perpendicular... If the angle which the ray in water encloses with the perpendicular to the surface be greater than 48 degrees, the ray will not quit the water at all: it will be *totally reflected* at the surface.... The angle which marks the limit where total reflection begins is called the limiting angle of the medium. For water this angle is 48°27', for flint glass it is 38°41', while for diamond it is 23°42'." Unpigmented human hairs have also been to shown to act as an optical fibre.

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

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

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

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

The crucial attenuation limit of 20 dB/km was first achieved in 1970, by researchers Robert D. Maurer, Donald Keck, Peter C. Schultz, and Frank Zimar working for

American glass maker Corning Glass Works, now Corning Incorporated. They demonstrated a fiber with 17 dB/km attenuation by doping silica glass with titanium. A few years later they produced a fiber with only 4 dB/km attenuation using germanium dioxide as the core dopant. Such low attenuation ushered in optical fiber telecommunication. In 1981, General Electric produced fused quartz ingots that could be drawn into fiber optic strands 25 miles (40 km) long.

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

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

Applications

Optical fiber communication

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

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

Fiber optic sensors

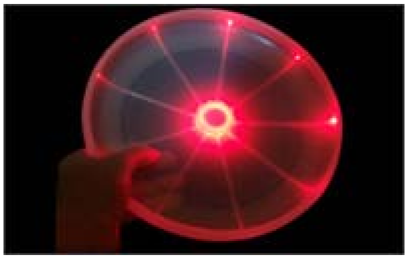
Fibers have many uses in remote sensing. In some applications, the sensor is itself an optical fiber. In other cases, fiber is used to connect a non-fiberoptic sensor to a measurement system. Depending on the application, fiber may be used because of its small size, or the fact that no electrical power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

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

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

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

Other uses of optical fibers



A frisbee illuminated by fiber optics

Light reflected from optical fiber illuminates exhibited model

Fiber optic front sight on a hand gun

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

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

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

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

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

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

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

Principle of operation

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

Index of refraction

The index of refraction is a way of measuring the speed of light in a material. Light travels fastest in a vacuum, such as outer space. The speed of light in a vacuum is about 300,000 kilometres (186 thousand miles) per second. Index of refraction is calculated by dividing the speed of light in a vacuum by the speed of light in some other medium. The index of refraction of a vacuum is therefore 1, by definition. The typical value for the cladding of an optical fiber is 1.46. The core value is typically 1.48. The larger the index of refraction, the slower light travels in that medium. From this information, a good rule of thumb is that signal using optical fiber for communication will travel at around 200 million meters per second. Or to put it another way, to travel 1000 kilometers in fiber, the signal will take 5 milliseconds to propagate. Thus a phone call carried by fiber between

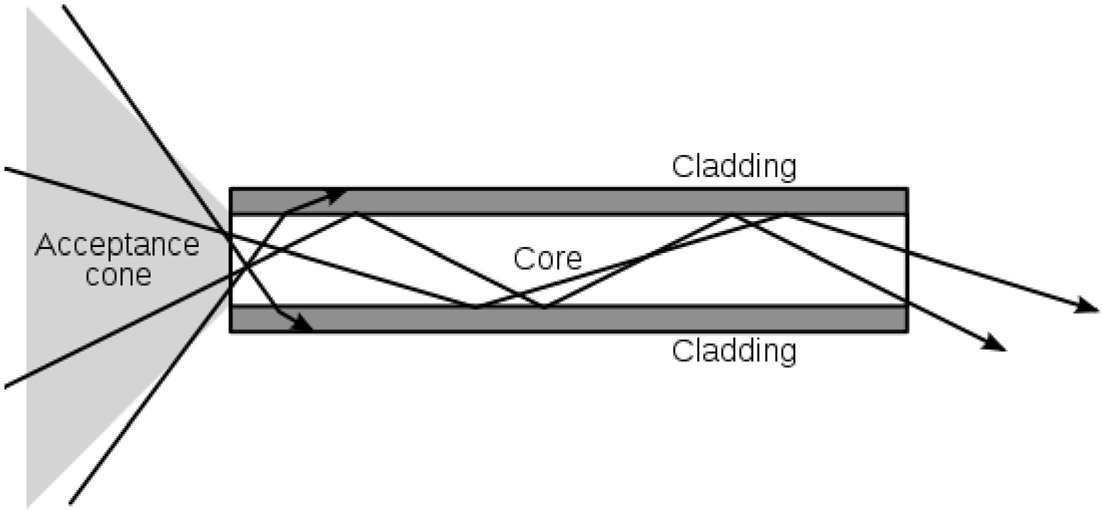
Sydney and New York, a 12000 kilometer distance, means that there is an absolute minimum delay of 60 milliseconds (or around 1/16 of a second) between when one caller speaks to when the other hears. (Of course the fiber in this case will probably travel a longer route, and there will be additional delays due to communication equipment switching and the process of encoding and decoding the voice onto the fiber).

Total internal reflection

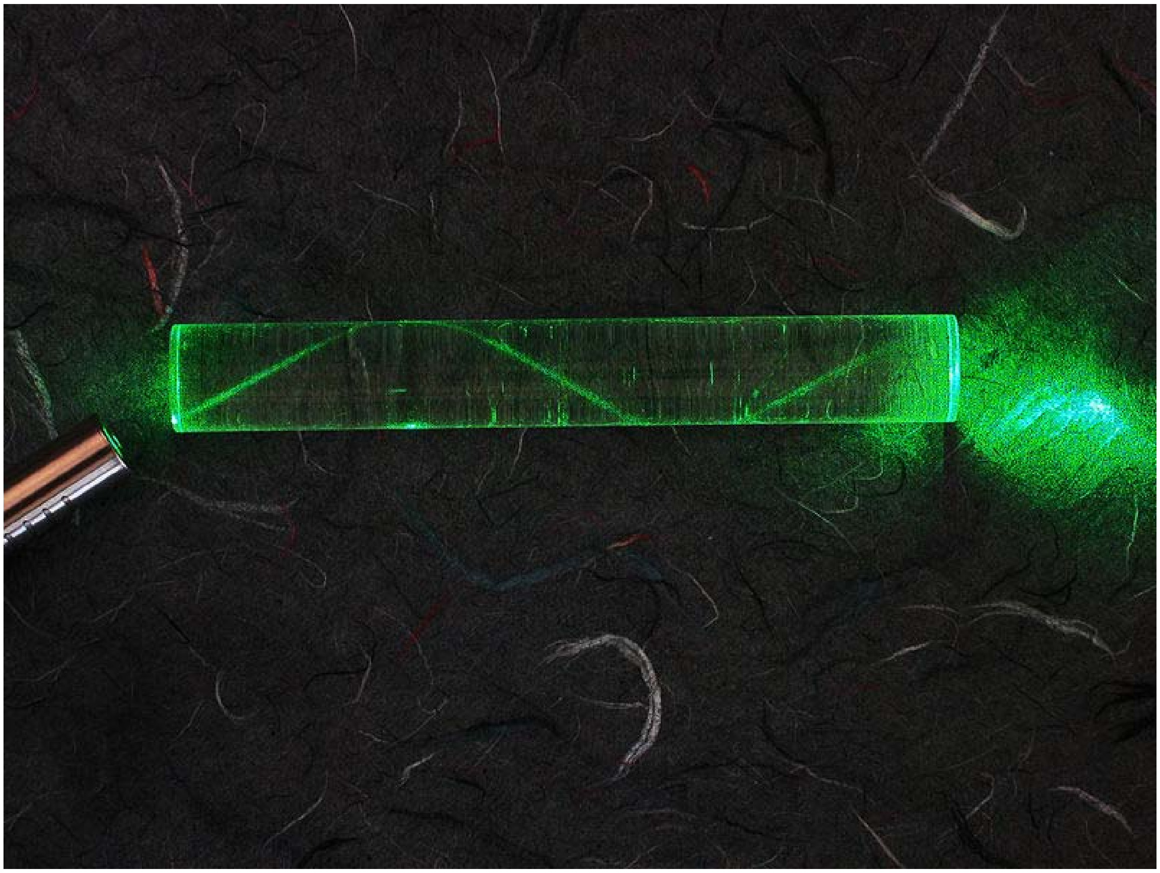
When light traveling in a dense medium hits a boundary at a steep angle (larger than the "critical angle" for the boundary), the light will be completely reflected. This effect is used in optical fibers to confine light in the core. Light travels along the fiber bouncing back and forth off of the boundary.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

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

Multi-mode fiber

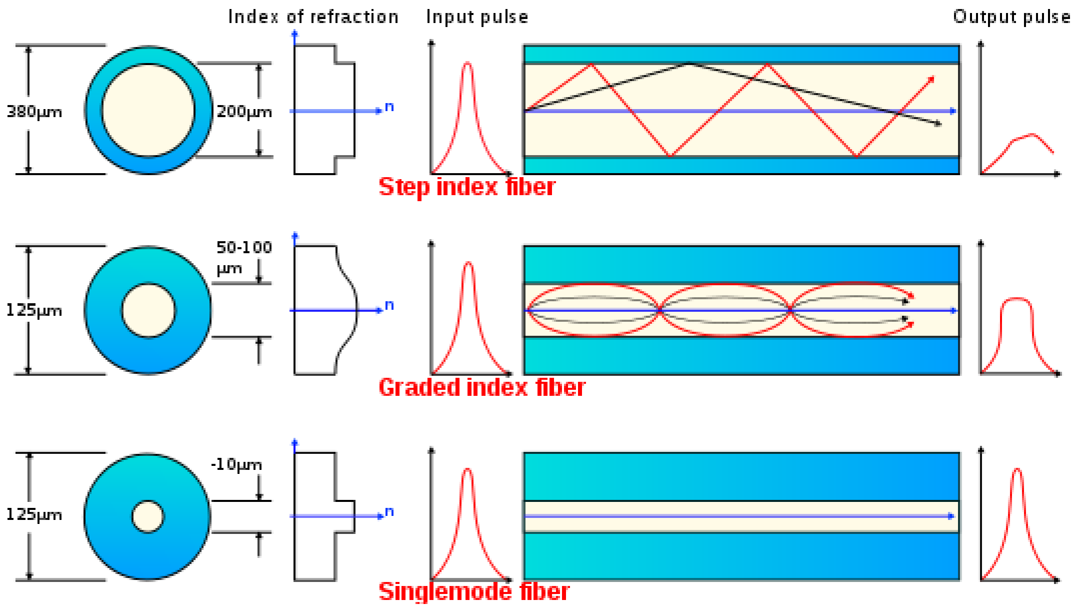


The propagation of light through a multi-mode optical fiber.



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

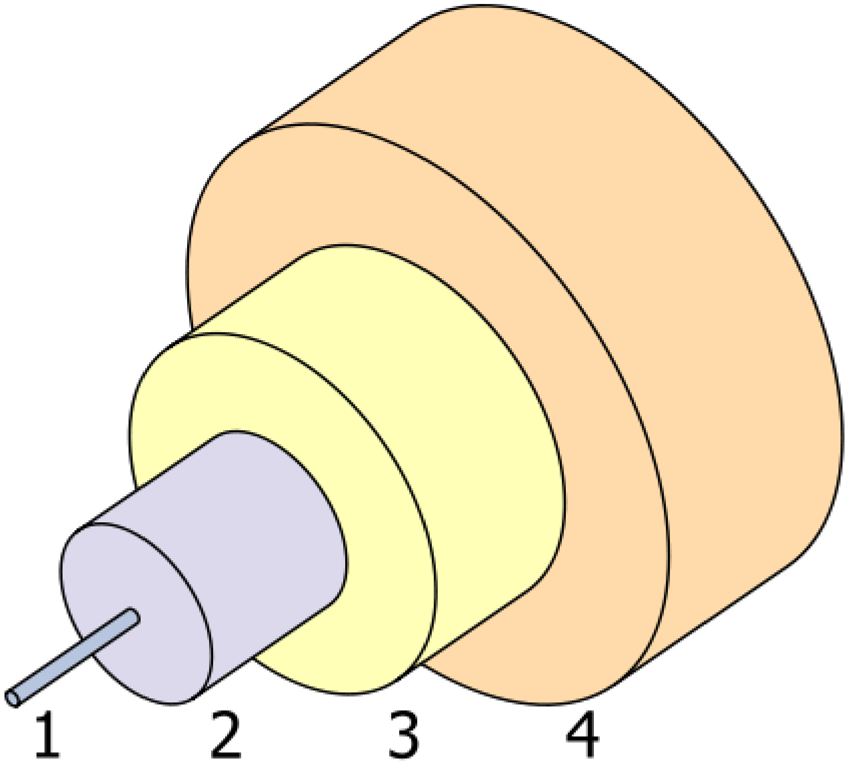
Fiber with large core diameter (greater than 10 micrometers) may be analyzed by geometrical optics. Such fiber is called *multi-mode fiber*, from the electromagnetic analysis (see below). In a step-index multi-mode fiber, rays of light are guided along the fiber core by total internal reflection. Rays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



Optical fiber types.

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

Single-mode fiber



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

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

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

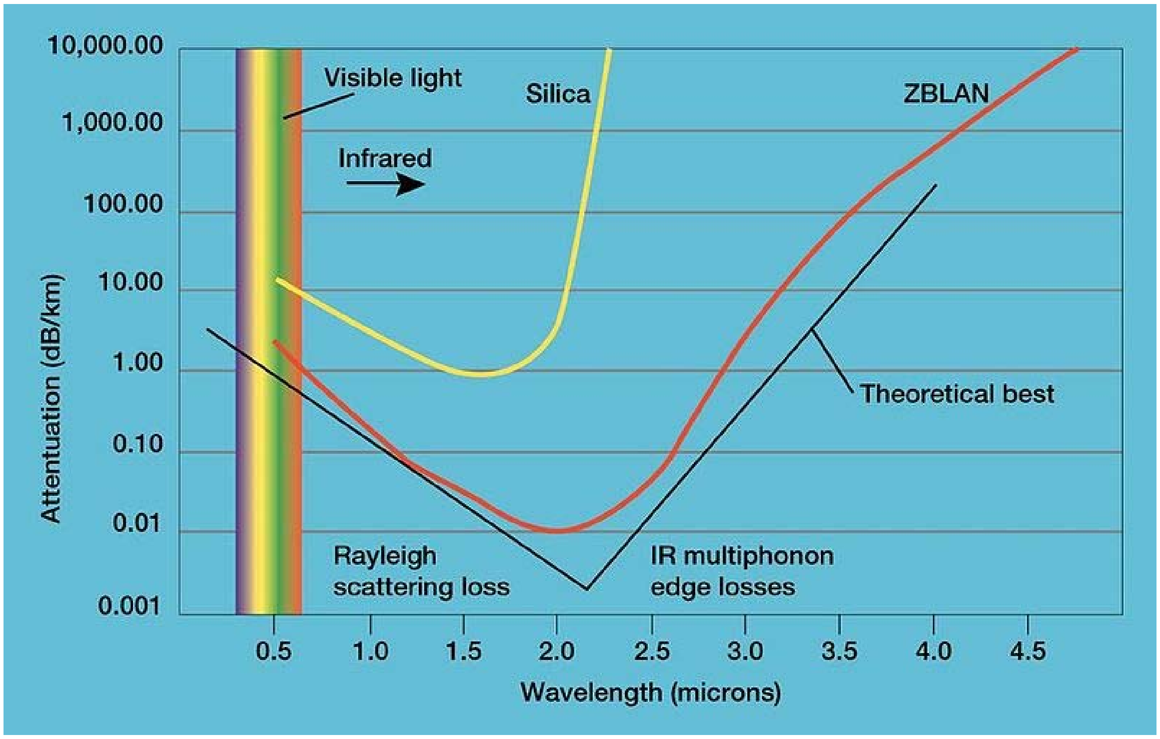
The most common type of single-mode fiber has a core diameter of 8–10 micrometers and is designed for use in the near infrared. The mode structure depends on the wavelength of the light used, so that this fiber actually supports a small number of additional modes at visible wavelengths. Multi-mode fiber,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

Special-purpose fiber

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

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

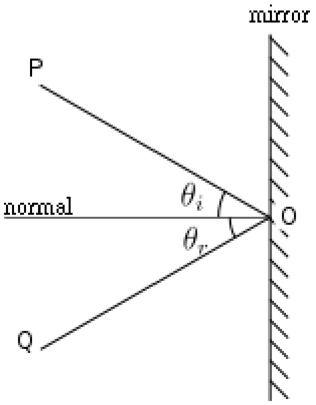
Mechanisms of attenuation



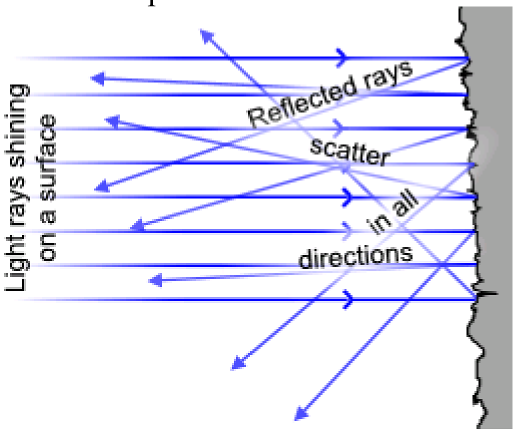
Light attenuation by ZBLAN and silica fibers

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

Light scattering



Specular reflection



Diffuse reflection

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

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

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

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

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

UV-Vis-IR absorption

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

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

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

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

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

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

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

Manufacturing

Materials

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

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

Silica

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

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

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

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

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

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

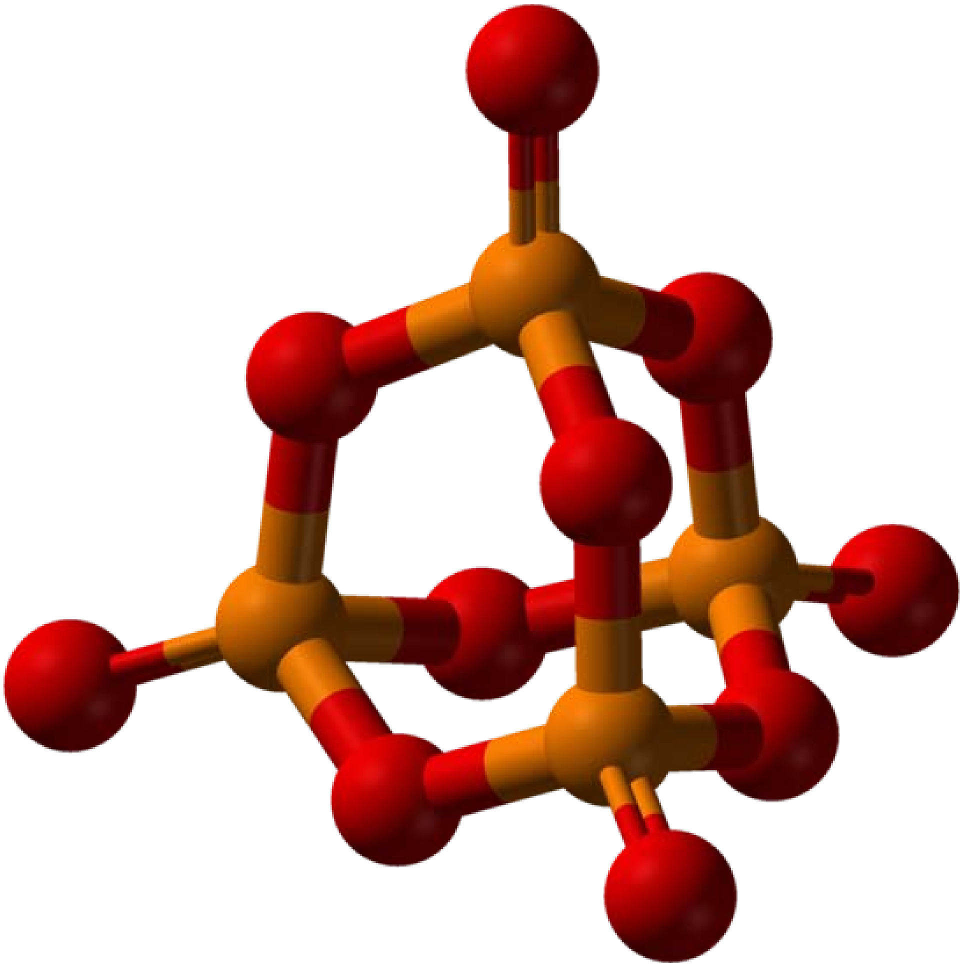
Fluorides

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

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

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

Phosphates



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

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

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

Chalcogenides

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

Process

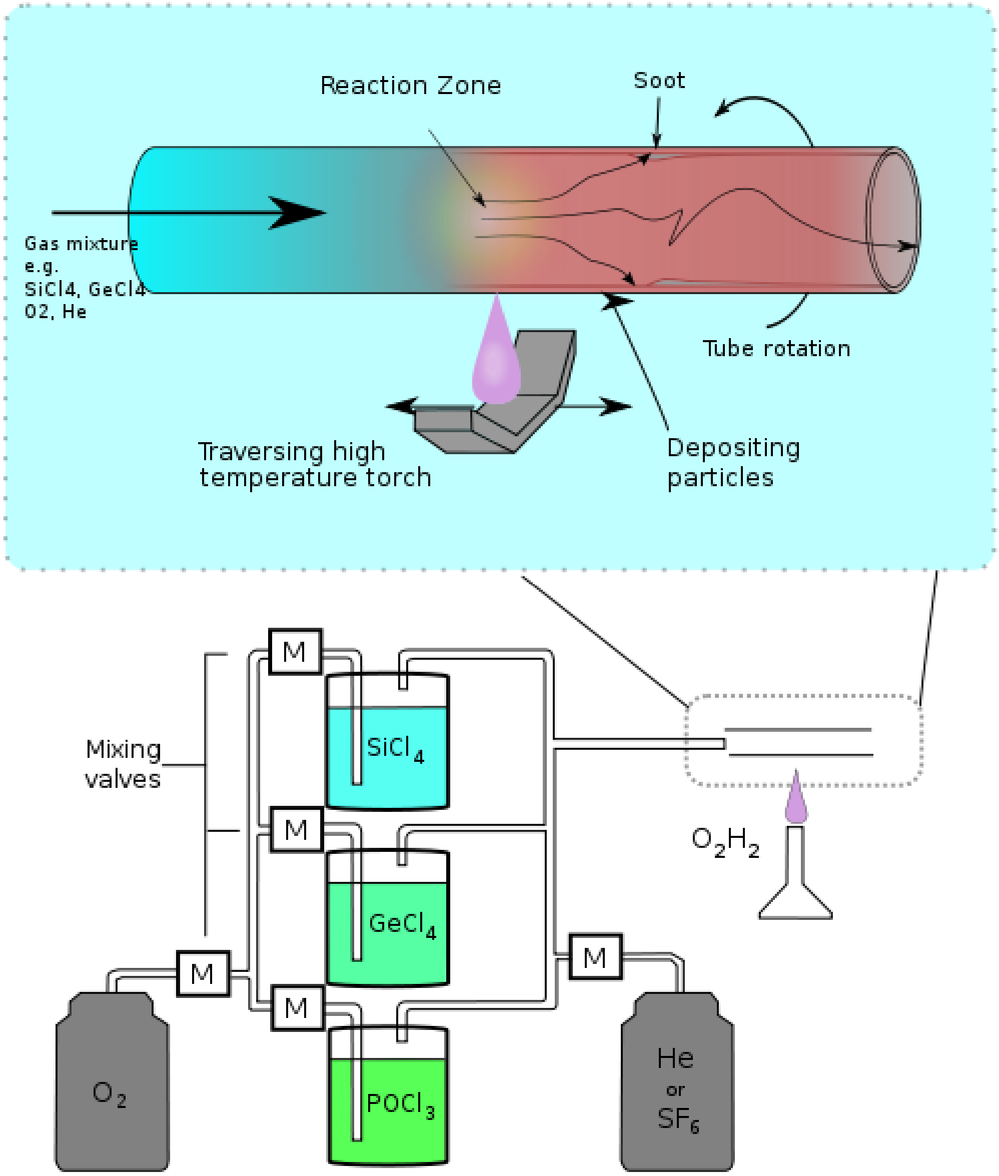


Illustration of the modified chemical vapor deposition (inside) process

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

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

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

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

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

Coatings

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

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

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

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

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

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

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

Practical issues

Optical fiber cables

An optical fiber cable

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

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

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

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

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

Termination and splicing

ST connectors on multi-mode fiber.

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

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

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

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

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

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

Free-space coupling

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

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

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

Fiber fuse

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

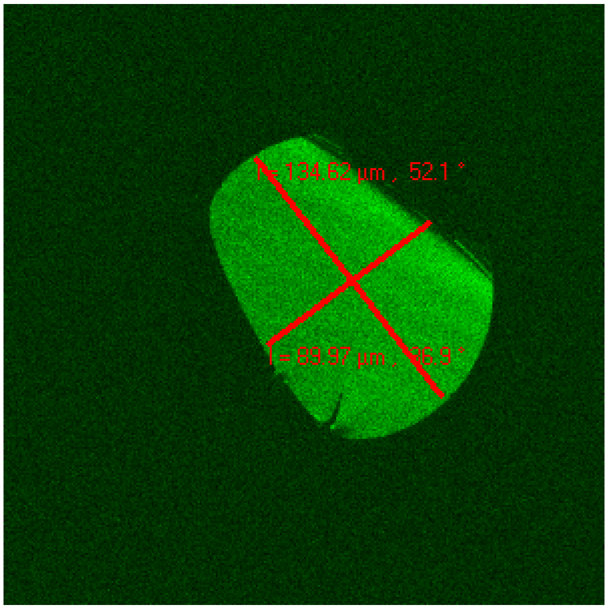
Example

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

Electric power transmission

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

Preform



Cross-section of a fiber drawn from a D-shaped **preform**

A preform is a piece of glass used to draw an optical fiber. The preform may consist of several pieces of a glass with different refractive index, to provide the core and cladding of the fiber. The shape of the preform may be circular, although for some applications such as double-clad fibers another form is preferred. In fiber lasers based on double-clad fiber, an asymmetric shape improves the filling factor for laser pumping.

Due to the surface tension, the shape is smoothed during the drawing process, and the shape of the resulting fiber does not reproduce the sharp edges of the preform. Nevertheless, the careful polishing of the **preform** is important, any defects of the **preform** surface affect the optical and mechanical properties of the resulting fiber. In particular, the preform for the test-fiber shown in the figure was not polished well, and the cracks are seen with confocal optical microscope.

Chapter 15

Semiconductor

A **semiconductor** is a material with electrical conductivity due to electron flow (as opposed to ionic conductivity) intermediate in magnitude between that of a conductor and an insulator. This means a conductivity roughly in the range of 10^3 to 10^{-8} siemens per centimeter. Semiconductor materials are the foundation of modern electronics, including radio, computers, telephones, and many other devices. Such devices include transistors, solar cells, many kinds of diodes including the light-emitting diode, the silicon controlled rectifier, and digital and analog integrated circuits. Similarly, semiconductor solar photovoltaic panels directly convert light energy into electrical energy. In a metallic conductor, current is carried by the flow of electrons. In semiconductors, current is often schematized as being carried either by the flow of electrons or by the flow of positively charged "holes" in the electron structure of the material. Actually, however, in both cases only electron movements are involved.

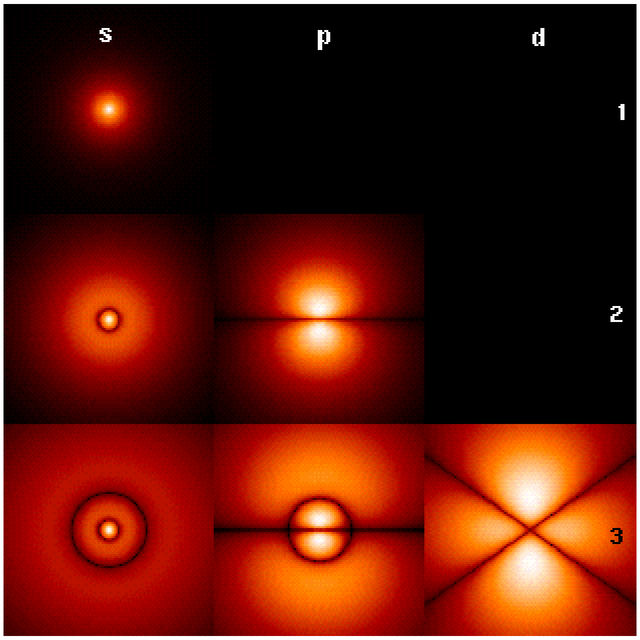
Common semiconducting materials are crystalline solids, but amorphous and liquid semiconductors are known. These include hydrogenated amorphous silicon and mixtures of arsenic, selenium and tellurium in a variety of proportions. Such compounds share with better known semiconductors intermediate conductivity and a rapid variation of conductivity with temperature, as well as occasional negative resistance. Such disordered materials lack the rigid crystalline structure of conventional semiconductors such as silicon and are generally used in thin film structures, which are less demanding for as concerns the electronic quality of the material and thus are relatively insensitive to impurities and radiation damage. Organic semiconductors, that is, organic materials with properties resembling conventional semiconductors, are also known.

Silicon is used to create most semiconductors commercially. Dozens of other materials are used, including germanium, gallium arsenide, and silicon carbide. A pure semiconductor is often called an “intrinsic” semiconductor. The electronic properties and the conductivity of a semiconductor can be changed in a controlled manner by adding very small quantities of other elements, called “dopants”, to the intrinsic material. In crystalline silicon typically this is achieved by adding impurities of boron or phosphorus to the melt and then allowing the melt to solidify into the crystal. This process is called "doping".

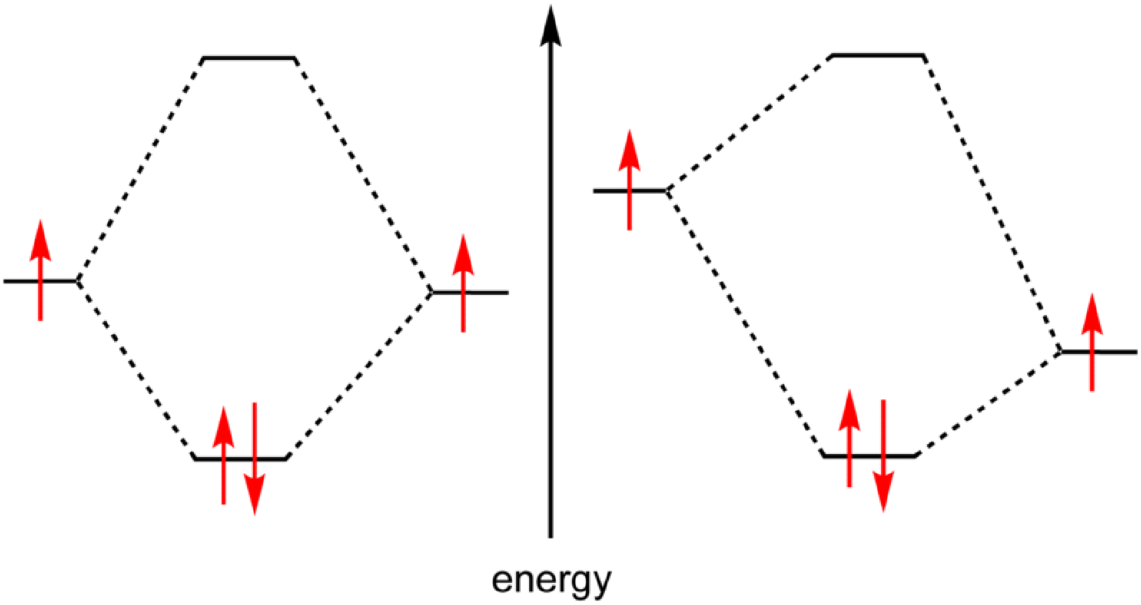
Explaining semiconductor energy bands

There are three popular ways to classify the electronic structure of a crystal.

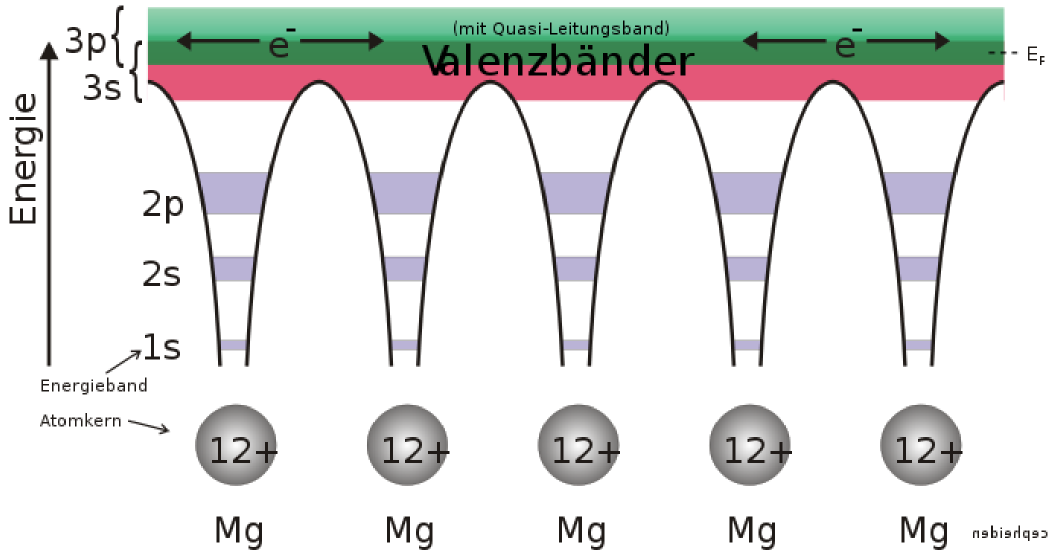
- Band structure
- atoms – crystal – vacuum



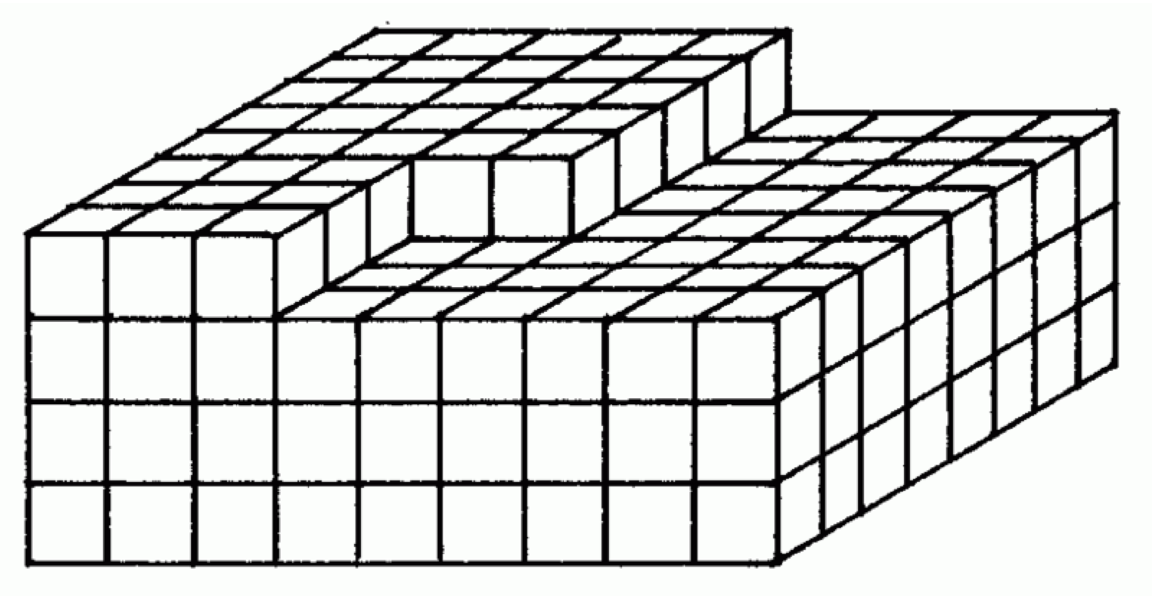
In a single H-atom an electron resides in well known orbitals. Note that the orbitals are called s,p,d in order of increasing circular current.



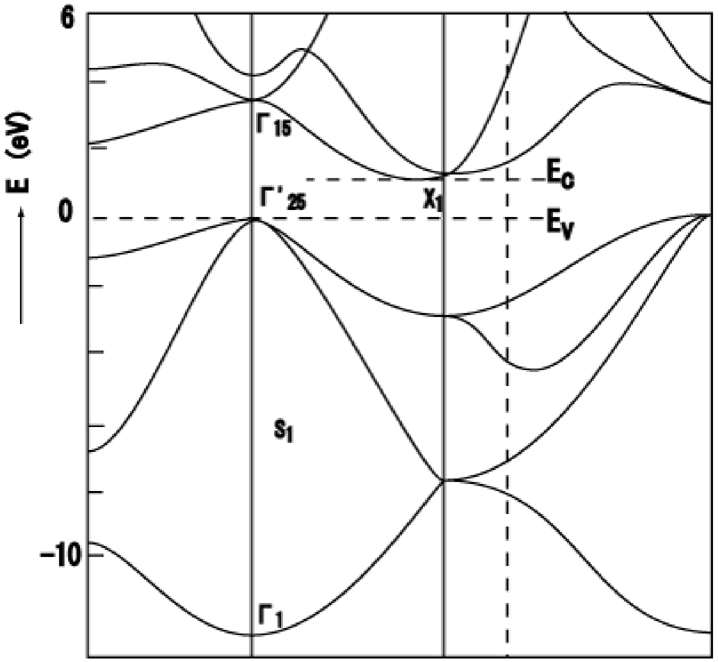
Putting two atoms together leads to delocalized orbitals across two atoms, yielding a covalent bond. Due to the Pauli exclusion principle, every state can contain only one electron.



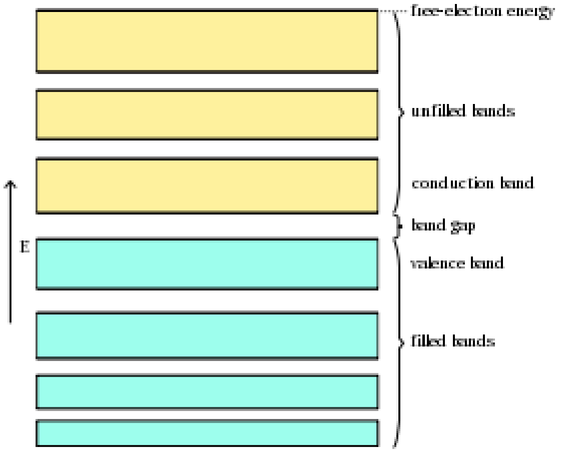
This can be continued with more atoms. Note: This picture shows a metal, not an actual semiconductor.



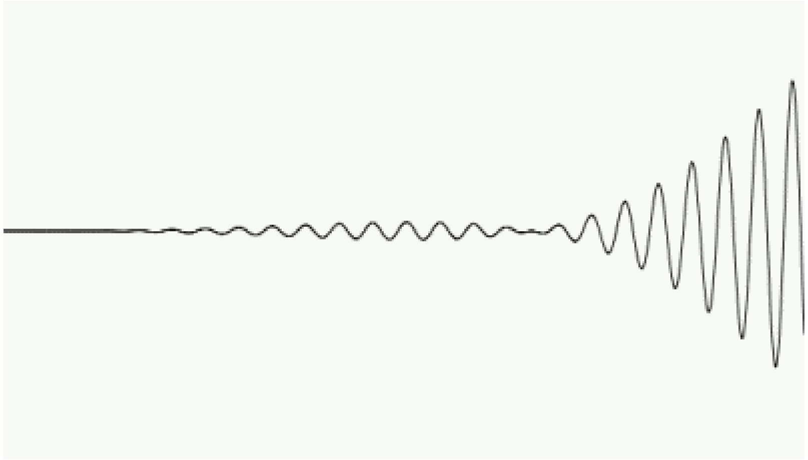
Continuing to add creates a crystal, which may then be cut into a tape and fused together at the ends to allow circular currents.



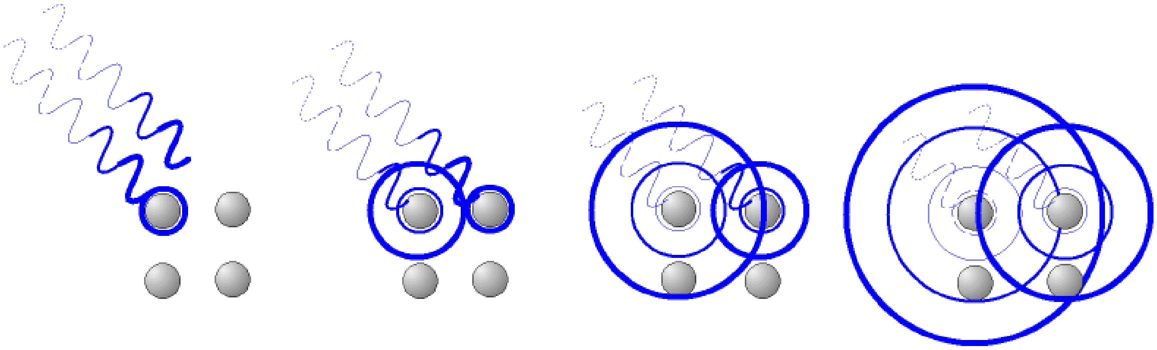
For this regular solid the band structure can be calculated or measured.



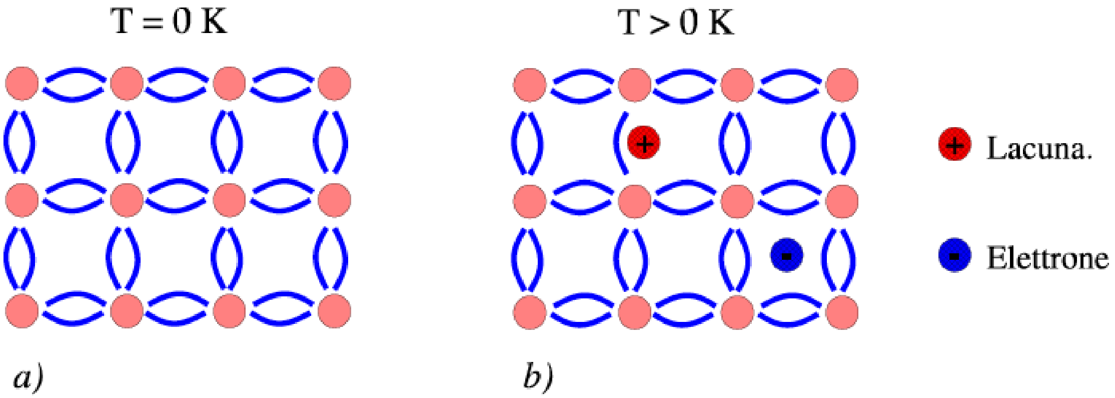
Integrating over the k axis gives the bands of a semiconductor showing a full valence band and an empty conduction band. Generally stopping at the vacuum level is undesirable, because some people want to calculate: photoemission, inverse photoemission



After the band structure is determined states can be combined to generate wave packets. As this is analogous to wave packages in free space, the results are similar.



An alternative description, which does not really appreciate the strong Coulomb interaction, shoots free electrons into the crystal and looks at the scattering.



A third alternative description uses strongly localized unpaired electrons in chemical bonds, which looks almost like a Mott insulator.

Energy bands and electrical conduction

In classic crystalline semiconductors, the electrons can have energies only within certain bands (i.e. ranges of levels of energy). Energetically, these bands are located between the energy of the ground state, corresponding to electrons tightly bound to the atomic nuclei of the material, and the free electron energy. The latter is the energy required for an electron to escape entirely from the material. The energy bands each correspond to a large number of discrete quantum states of the electrons, and most of the states with low energy (closer to the nucleus) are full, up to a particular band called the *valence band*. Semiconductors and insulators are distinguished from metals because the valence band in them is nearly filled with electrons under usual operating conditions, while very few (semiconductor) or virtually none (insulator) of them are available in the *conduction band*, the band immediately above the valence band.

The ease with which electrons in a semiconductor can be excited from the valence band to the conduction band depends on the band gap between the bands. The size of this energy bandgap serves as an arbitrary dividing line (roughly 4 eV) between semiconductors and insulators.

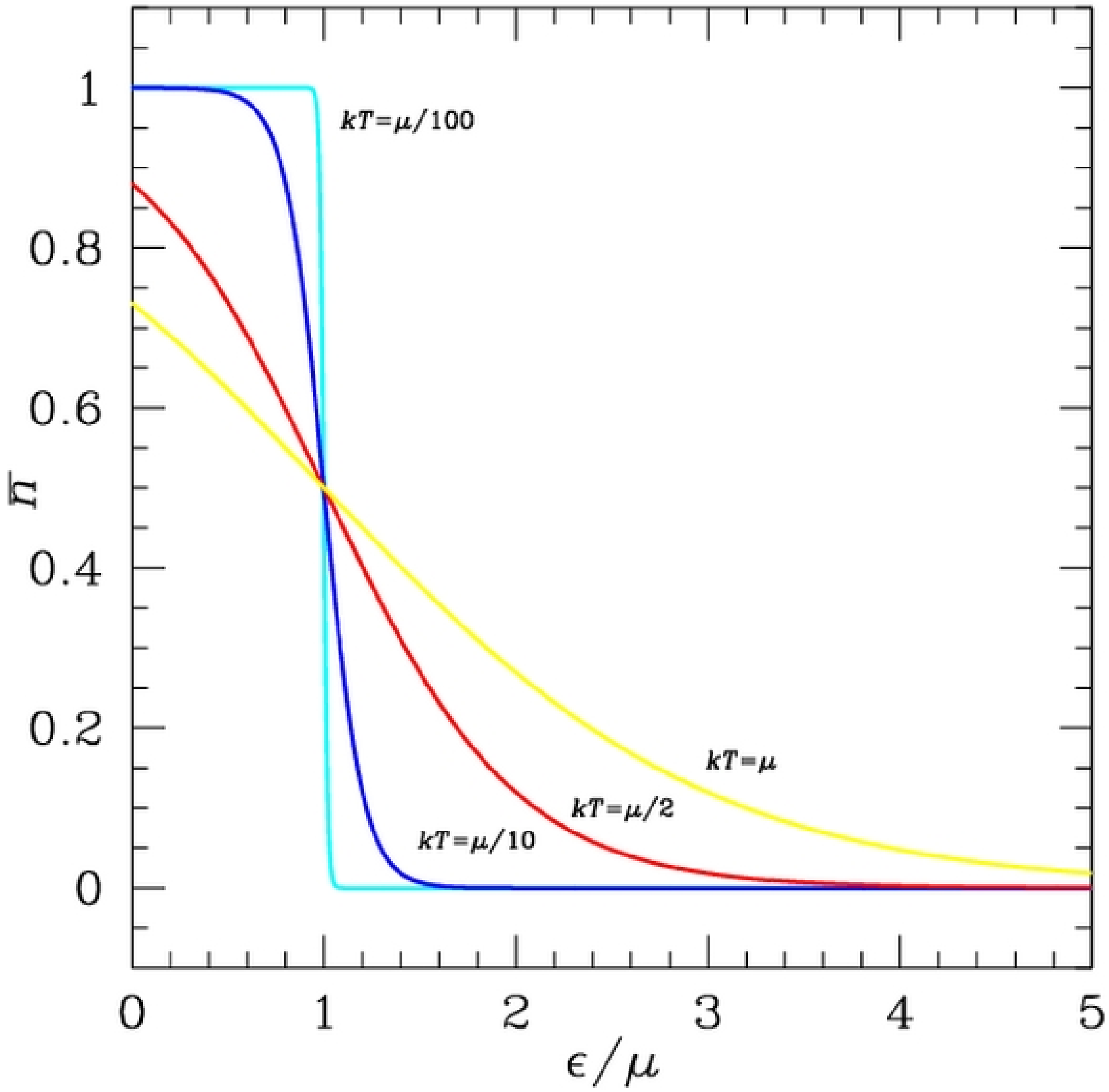
With covalent bonds, an electron moves by hopping to a neighboring bond. The Pauli exclusion principle requires the electron to be lifted into the higher anti-bonding state of that bond. For delocalized states, for example in one dimension – that is in a nanowire, for every energy there is a state with electrons flowing in one direction and another state with the electrons flowing in the other. For a net current to flow, more states for one direction than for the other direction must be occupied. For this to occur, energy is required, as in the semiconductor the next higher states lie above the band gap. Often this is stated as: full bands do not contribute to the electrical conductivity. However, as the temperature of a semiconductor rises above absolute zero, there is more energy in the semiconductor to spend on lattice vibration and — more importantly for us — on lifting some electrons into an energy states of the conduction band. The current-carrying electrons in the conduction band are known as "free electrons", although they are often simply called "electrons" if context allows this usage to be clear.

Electrons excited to the conduction band also leave behind electron holes, or unoccupied states in the valence band. Both the conduction band electrons and the valence band holes contribute to electrical conductivity. The holes themselves don't actually move, but a neighboring electron can move to fill the hole, leaving a hole at the place it has just come from, and in this way the holes appear to move, and the holes behave as if they were actual positively charged particles.

One covalent bond between neighboring atoms in the solid is ten times stronger than the binding of the single electron to the atom, so freeing the electron does not imply destruction of the crystal structure.

Holes: electron absence as a charge carrier

The concept of holes can also be applied to metals, where the Fermi level lies *within* the conduction band. With most metals the Hall effect indicates electrons are the charge carriers. However, some metals have a mostly filled conduction band. In these, the Hall effect reveals positive charge carriers, which are not the ion-cores, but holes. In the case of a metal, only a small amount of energy is needed for the electrons to find other unoccupied states to move into, and hence for current to flow. Sometimes even in this case it may be said that a hole was left behind, to explain why the electron does not fall back to lower energies: It cannot find a hole. In the end in both materials electron-phonon scattering and defects are the dominant causes for resistance.



Fermi-Dirac distribution. States with energy ε below the Fermi energy, here μ, have higher probability *n* to be occupied, and those above are less likely to be occupied. Smearing of the distribution increases with temperature.

The energy distribution of the electrons determines which of the states are filled and which are empty. This distribution is described by Fermi-Dirac statistics. The distribution is characterized by the temperature of the electrons, and the *Fermi energy* or *Fermi level*. Under absolute zero conditions the Fermi energy can be thought of as the energy up to which available electron states are occupied. At higher temperatures, the Fermi energy is the energy at which the probability of a state being occupied has fallen to 0.5.

The dependence of the electron energy distribution on temperature also explains why the conductivity of a semiconductor has a strong temperature dependency, as a semiconductor operating at lower temperatures will have fewer available free electrons and holes able to do the work.

Energy–momentum dispersion

In the preceding description an important fact is ignored for the sake of simplicity: the *dispersion* of the energy. The reason that the energies of the states are broadened into a band is that the energy depends on the value of the wave vector, or *k-vector*, of the electron. The k-vector, in quantum mechanics, is the representation of the momentum of a particle.

The dispersion relationship determines the effective mass, m^*, of electrons or holes in the semiconductor, according to the formula:

$$m^* = \hbar^2 \cdot \left[\frac{d^2E(k)}{dk^2}\right]^{-1} .$$

The effective mass is important as it affects many of the electrical properties of the semiconductor, such as the electron or hole mobility, which in turn influences the *diffusivity* of the charge carriers and the electrical conductivity of the semiconductor.

Typically the effective mass of electrons and holes are different. This affects the relative performance of *p-channel* and *n-channel* IGFETs.

The top of the valence band and the bottom of the conduction band might not occur at that same value of *k*. Materials with this situation, such as silicon and germanium, are known as *indirect bandgap* materials. Materials in which the band extrema are aligned in *k*, for example gallium arsenide, are called *direct bandgap* semiconductors. Direct gap semiconductors are particularly important in optoelectronics because they are much more efficient as light emitters than indirect gap materials.

Carrier generation and recombination

When ionizing radiation strikes a semiconductor, it may excite an electron out of its energy level and consequently leave a hole. This process is known as *electron–hole pair generation*. Electron-hole pairs are constantly generated from thermal energy as well, in the absence of any external energy source.

Electron-hole pairs are also apt to recombine. Conservation of energy demands that these recombination events, in which an electron loses an amount of energy larger than the band gap, be accompanied by the emission of thermal energy (in the form of phonons) or radiation (in the form of photons).

In some states, the generation and recombination of electron–hole pairs are in equipoise. The number of electron-hole pairs in the steady state at a given temperature is determined by quantum statistical mechanics. The precise quantum mechanical mechanisms of generation and recombination are governed by conservation of energy and conservation of momentum.

As the probability that electrons and holes meet together is proportional to the product of their amounts, the product is in steady state nearly constant at a given temperature, providing that there is no significant electric field (which might "flush" carriers of both types, or move them from neighbour regions containing more of them to meet together) or externally driven pair generation. The product is a function of the temperature, as the probability of getting enough thermal energy to produce a pair increases with temperature, being approximately $\exp(-E_G/kT)$, where k is Boltzmann's constant, T is absolute temperature and E_G is band gap.

The probability of meeting is increased by carrier traps—impurities or dislocations which can trap an electron or hole and hold it until a pair is completed. Such carrier traps are sometimes purposely added to reduce the time needed to reach the steady state.

Semi-insulators

Some materials are classified as **semi-insulators**. These have electrical conductivity nearer to that of electrical insulators. Semi-insulators find niche applications in micro-electronics, such as substrates for HEMT. An example of a common semi-insulator is gallium arsenide.

Doping

The property of semiconductors that makes them most useful for constructing electronic devices is that their conductivity may easily be modified by introducing impurities into their crystal lattice. The process of adding controlled impurities to a semiconductor is known as *doping*. The amount of impurity, or dopant, added to an *intrinsic* (pure) semiconductor varies its level of conductivity. Doped semiconductors are often referred to as *extrinsic*. By adding impurity to pure semiconductors, the electrical conductivity may be varied not only by the number of impurity atoms but also, by the type of impurity atom and the changes may be thousand folds and million folds. For example, 1 cm^3 of a metal or semiconductor specimen has a number of atoms on the order of 10^{22}. Since every atom in metal donates at least one free electron for conduction in metal, 1 cm^3 of metal contains free electrons on the order of 10^{22}. At the temperature close to 20 °C , 1 cm^3 of pure germanium contains about 4.2×10^{22} atoms and 2.5×10^{13} free electrons and 2.5×10^{13} holes (empty spaces in crystal lattice having positive charge) The addition of 0.001% of arsenic (an impurity) donates an extra 10^{17} free electrons in the same volume and the electrical conductivity increases about 10,000 times."

Dopants

The materials chosen as suitable dopants depend on the atomic properties of both the dopant and the material to be doped. In general, dopants that produce the desired controlled changes are classified as either electron acceptors or donors. A donor atom that activates (that is, becomes incorporated into the crystal lattice) donates weakly bound valence electrons to the material, creating excess negative charge carriers. These weakly bound electrons can move about in the crystal lattice relatively freely and can facilitate conduction in the presence of an electric field. (The donor atoms introduce some states under, but very close to the conduction band edge. Electrons at these states can be easily excited to the conduction band, becoming free electrons, at room temperature.) Conversely, an activated acceptor produces a hole. Semiconductors doped with *donor* impurities are called *n-type*, while those doped with *acceptor* impurities are known as *p-type*. The n and p type designations indicate which charge carrier acts as the material's majority carrier. The opposite carrier is called the minority carrier, which exists due to thermal excitation at a much lower concentration compared to the majority carrier.

For example, the pure semiconductor silicon has four valence electrons. In silicon, the most common dopants are IUPAC group 13 (commonly known as *group III*) and group 15 (commonly known as *group V*) elements. Group 13 elements all contain three valence electrons, causing them to function as acceptors when used to dope silicon. Group 15 elements have five valence electrons, which allows them to act as a donor. Therefore, a silicon crystal doped with boron creates a p-type semiconductor whereas one doped with phosphorus results in an n-type material.

Carrier concentration

The concentration of dopant introduced to an intrinsic semiconductor determines its concentration and indirectly affects many of its electrical properties. The most important factor that doping directly affects is the material's carrier concentration. In an intrinsic semiconductor under thermal equilibrium, the concentration of electrons and holes is equivalent. That is,

$$n = p = n_i.$$

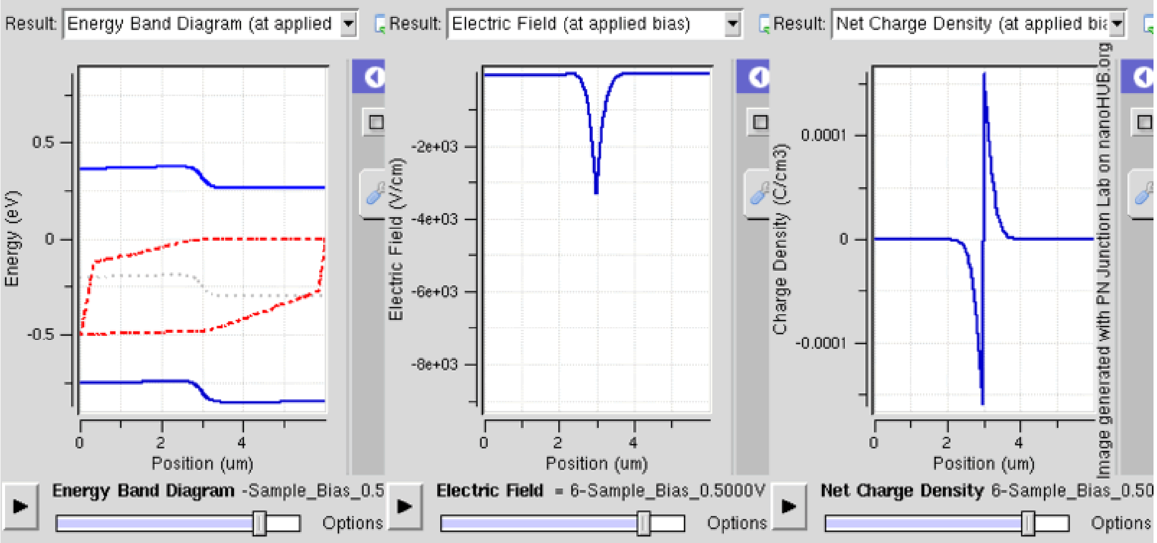
If we have a non-intrinsic semiconductor in thermal equilibrium the relation becomes:

$$n_0 \cdot p_0 = n_i^2$$

where n_0 is the concentration of conducting electrons, p_0 is the electron hole concentration, and n_i is the material's intrinsic carrier concentration. Intrinsic carrier concentration varies between materials and is dependent on temperature. Silicon's n_i, for example, is roughly 1.08×10^{10} cm^{-3} at 300 kelvins (room temperature).

In general, an increase in doping concentration affords an increase in conductivity due to the higher concentration of carriers available for conduction. Degenerately (very highly) doped semiconductors have conductivity levels comparable to metals and are often used in modern integrated circuits as a replacement for metal. Often superscript plus and minus symbols are used to denote relative doping concentration in semiconductors. For example, n^+ denotes an n-type semiconductor with a high, often degenerate, doping concentration. Similarly, p^- would indicate a very lightly doped p-type material. It is useful to note that even degenerate levels of doping imply low concentrations of impurities with respect to the base semiconductor. In crystalline intrinsic silicon, there are approximately 5×10^{22} atoms/cm³. Doping concentration for silicon semiconductors may range anywhere from 10^{13} cm^{-3} to 10^{18} cm^{-3}. Doping concentration above about 10^{18} cm^{-3} is considered degenerate at room temperature. Degenerately doped silicon contains a proportion of impurity to silicon on the order of parts per thousand. This proportion may be reduced to parts per billion in very lightly doped silicon. Typical concentration values fall somewhere in this range and are tailored to produce the desired properties in the device that the semiconductor is intended for.

Effect on band structure



Result: Energy Band Diagram (at applied

Result: Electric Field (at applied bias)

Result: Net Charge Density (at applied bia

Energy Band Diagram -Sample_Bias_0.5

Electric Field = 6-Sample_Bias_0.5000V

Net Charge Density 6-Sample_Bias_0.50

Image generated with PN Junction Lab on nanoHUB.org

Band diagram of PN junction operation in forward bias mode showing reducing depletion width. Both p and n junctions are doped at a 1e15/cm3 doping level, leading to built-in potential of ~0.59V. Reducing depletion width can be inferred from the shrinking charge profile, as fewer dopants are exposed with increasing forward bias.

Doping a semiconductor crystal introduces allowed energy states within the band gap but very close to the energy band that corresponds to the dopant type. In other words, donor impurities create states near the conduction band while acceptors create states near the valence band. The gap between these energy states and the nearest energy band is usually referred to as dopant-site bonding energy or E_B and is relatively small. For example, the E_B for boron in silicon bulk is 0.045 eV, compared with silicon's band gap of about 1.12 eV. Because E_B is so small, it takes little energy to ionize the dopant atoms and create free carriers in the conduction or valence bands. Usually the thermal energy available at room temperature is sufficient to ionize most of the dopant.

Dopants also have the important effect of shifting the material's Fermi level towards the energy band that corresponds with the dopant with the greatest concentration. Since the Fermi level must remain constant in a system in thermodynamic equilibrium, stacking layers of materials with different properties leads to many useful electrical properties. For example, the p-n junction's properties are due to the energy band bending that happens as a result of lining up the Fermi levels in contacting regions of p-type and n-type material.

This effect is shown in a *band diagram*. The band diagram typically indicates the variation in the valence band and conduction band edges versus some spatial dimension, often denoted *x*. The Fermi energy is also usually indicated in the diagram. Sometimes the *intrinsic Fermi energy*, E_i, which is the Fermi level in the absence of doping, is shown. These diagrams are useful in explaining the operation of many kinds of semiconductor devices.

Preparation of semiconductor materials

Semiconductors with predictable, reliable electronic properties are necessary for mass production. The level of chemical purity needed is extremely high because the presence of impurities even in very small proportions can have large effects on the properties of the material. A high degree of crystalline perfection is also required, since faults in crystal structure (such as dislocations, twins, and stacking faults) interfere with the semiconducting properties of the material. Crystalline faults are a major cause of defective semiconductor devices. The larger the crystal, the more difficult it is to achieve the necessary perfection. Current mass production processes use crystal ingots between 100 mm and 300 mm (4–12 inches) in diameter which are grown as cylinders and sliced into wafers.

Because of the required level of chemical purity and the perfection of the crystal structure which are needed to make semiconductor devices, special methods have been developed to produce the initial semiconductor material. A technique for achieving high purity includes growing the crystal using the Czochralski process. An additional step that can be used to further increase purity is known as zone refining. In zone refining, part of a solid crystal is melted. The impurities tend to concentrate in the melted region, while the desired material recrystalizes leaving the solid material more pure and with fewer crystalline faults.

In manufacturing semiconductor devices involving heterojunctions between different semiconductor materials, the lattice constant, which is the length of the repeating element of the crystal structure, is important for determining the compatibility of materials.