Technological Advancements and Innovations

Sima Butts

First Edition, 2012

ISBN 978-81-323-4034-8

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

Table of Contents

Chapter-1

Cryonics

Technicians prepare a patient for cryopreservation.

Cryonics (from Greek *kryos-* meaning *icy cold*) is the low-temperature preservation of humans and animals who can no longer be sustained by contemporary medicine, with the hope that healing and resuscitation may be possible in the future. Cryopreservation of people or large animals is not reversible with current technology. The stated rationale for cryonics is that people who are considered dead by current legal or medical definitions may not necessarily be dead according to the more stringent information-theoretic definition of death. It is proposed that cryopreserved people might someday be recovered by using highly advanced future technology.

The future repair technologies assumed by cryonics are still hypothetical and not widely known or recognized. Responding to skepticism from scientists such as Steve Jones, an open letter supporting cryonics was written and signed by currently 62 scientists. As of 2010, only around 200 people have undergone the procedure since it was first proposed in 1962. In the United States, cryonics can only be legally performed on humans after they have been pronounced legally dead as otherwise it would count as murder or assisted suicide.

Cryonics procedures ideally begin within minutes of cardiac arrest, and use cryoprotectants to prevent ice formation during cryopreservation. However, the idea of cryonics also includes preservation of people after longer post-mortem delays because of the possibility that brain structures encoding memory and personality may still persist or be inferable. Whether sufficient brain information still exists for cryonics to work under some preservation conditions may be intrinsically unprovable by present knowledge. Therefore, most proponents of cryonics see it as an intervention with prospects for success that vary widely depending on circumstances.

Premises of cryonics

A central premise of cryonics is that long-term memory, personality, and identity are stored in durable cell structures and patterns within the brain that do not require continuous brain activity to survive. This premise is generally accepted in medicine; it is known that under certain conditions the brain can stop functioning and still later recover with retention of long-term memory. Additional scientific premises of cryonics are that (1) brain structures encoding personality and long-term memory persist for some time after clinical death, (2) these structures are preserved by cryopreservation, and (3) future technologies that could restore encoded memories to functional expression in a healed person are theoretically possible.

Cryonics is controversial because the technologies of premise (3) are so advanced that premises (1) and (2) are considered irrelevant by most scientists. Whether biological traces of memory or personhood might persist after clinical death is obviously a question of interest. Similarly, outside of cryonics there is no interest in the question of whether memory encoding might survive cryopreservation because the question is regarded as meaningless until cryopreservation can be reversed. At present only cells, tissues, and some small organs can be reversibly cryopreserved. Medical science is primarily concerned with what is demonstrably achievable, not what is theoretically possible. There are therefore no established scientific specialties or journals directly concerned with the scientific questions posed by cryonics.

Cryonics advocates claim that it is possible to preserve the fine cell structures of the brain in which memory and identity reside with present technology. They say that demonstrably reversible preservation is not necessary to achieve the present-day goal of cryonics, which is preservation of brain information that encodes memory and personal identity. They believe that current cryonics procedures can preserve the anatomical basis of mind, and that this should be sufficient to prevent information-theoretic death until future repairs might be possible.

A moral premise of cryonics is that cryopreserving people when there is no other hope is the right thing to do, sometimes even under poor conditions that make the scientific premises of cryonics highly uncertain. Some cryonicists believe as a matter of principle that anyone who would ordinarily be regarded as dead should instead be made a "permanent patient" subject to whatever future advances might bring.

Obstacles to success

Preservation injury

Long-term cryopreservation can be achieved by cooling to near 77.15 Kelvin, the boiling point of liquid nitrogen. It is a common mistaken belief that cells will lyse (burst) due to the formation of ice crystals within the cell, since this only occurs if the freezing rate exceeds the osmotic loss of water to the extracellular space. However, damage from freezing can still be serious; ice may still form between cells, causing mechanical and chemical damage. Cryonics organizations use cryoprotectants to reduce this damage. Cryoprotectant solutions are circulated through blood vessels to remove and replace water inside cells with chemicals that prevent freezing. This can reduce damage greatly, but freezing of whole people still causes injuries that are not reversible with present technology.

When used at high concentrations, cryoprotectants stop ice formation completely. Cooling and solidification without freezing is called vitrification. The first cryoprotectant solutions able to vitrify at very slow cooling rates while still being compatible with tissue survival were developed in the late 1990s by cryobiologists Gregory Fahy and Brian Wowk for the purpose of banking transplantable organs. These solutions were adopted for use in cryonics by the Alcor Life Extension Foundation, for which they are believed to permit vitrification of some parts of the human body, especially the brain. This has allowed animal brains to be vitrified, warmed back up, and examined for ice damage using light and electron microscopy. No ice crystal damage was found. The Cryonics Institute also uses a vitrification solution developed by their staff cryobiologist, Dr. Yuri Pichugin, applying it principally to the brain.

Vitrification in cryonics is different than vitrification in mainstream cryobiology because vitrification in cryonics is not reversible with current technology. It is only structural vitrification. When successful it can prevent freezing injury in some body parts, but at the price of toxicity caused by cryoprotectant chemicals. The nature of this toxicity is still poorly understood. Cryonicists assume that toxicity is more subtle and repairable than obvious structural damage that would otherwise be caused by freezing. If, for example, toxicity is due to denatured proteins, those proteins could be repaired or replaced.

Ischemic injury

Ischemia means inadequate or absent blood circulation that deprives tissue of oxygen and nutrients. At least several minutes of ischemia is a typical part of cryonics because of the common legal requirement that cryonics procedures do not begin until after blood circulation stops. The heart must stop beating so that legal death can be declared. When there is advance notice of impending clinical death, it is sometimes possible to deploy a team of technicians to perform a “standby”. The team artificially restores blood circulation and breathing using techniques similar to CPR as soon as possible after the heart stops. The aim is to keep tissues alive after legal death by analogy to conventional medical procedures in which viable organs and tissues are obtained for transplant from legally deceased donors. Legal death does not mean that all the cells of the body have died.

Often in cryonics the brain is without oxygen for many minutes at warm temperatures, or even hours if the heart stops unexpectedly. This causes ischemic injury to the brain and other tissues that makes resuscitation impossible by present medical technology. Cryonicists justify preservation under such conditions by noting recent advances that allow brain resuscitation after longer periods of ischemia than the traditional 4-to-6-minute limit, and persistence of brain structure and even some brain cell function after long periods of clinical death. They argue that definitions of death change as technology advances, and the early stages of what is called "death" today is actually a form of ischemic injury that will be reversible in the future. They claim that personal survival during long periods of clinical death is determined by information-theoretic criteria.

Revival

Those who believe that revival may someday be possible generally look toward advanced bioengineering, molecular nanotechnology, nanomedicine, or mind uploading as key technologies. Revival requires repairing damage from lack of oxygen, cryoprotectant toxicity, thermal stress (fracturing), freezing in tissues that do not successfully vitrify, and reversing the effects that caused the patient's death. In many cases extensive tissue regeneration will be necessary. Hypothetical revival scenarios generally envision repairs being performed by vast numbers of microscopic organisms or devices. These devices would restore healthy cell structure and chemistry at the molecular level, ideally before warming. More radically, mind transfer has also been suggested as a possible revival approach if and when technology is ever developed to scan the memory contents of a preserved brain.

It has sometimes been written that cryonics revival will be a last-in-first-out process. In this view, preservation methods will get progressively better until eventually they are demonstrably reversible, after which medicine will begin to reach back and revive people cryopreserved by more primitive methods. Revival of people cryopreserved by early cryonics technology may require centuries, if it is possible at all. People cryopreserved in the future, with better technology, may require less advanced technology to be revived because they will have been cryopreserved with better technology that caused less damage to tissue. The "last in, first out" view of cryonics has been criticized because the quality of cryopreservation depends on many factors other than the era in which cryopreservation takes place.

It has been claimed that if technologies for general molecular analysis and repair are ever developed, then theoretically any damaged body could be "revived". Survival would then depend on whether preserved brain information was sufficient to permit restoration of all or part of the personal identity of the original person, with amnesia being the final dividing line between success and failure.

Neuropreservation

Neuropreservation is cryopreservation of the brain, often within the head, with surgical removal and disposal (usually cremation) of the rest of the body. Neuropreservation, sometimes called “neuro,” is one of two distinct preservation options in cryonics, the other being "whole body" preservation.

Neuropreservation is motivated by the brain's role as the primary repository of memory and personal identity. (For instance, spinal cord injury victims, organ transplant patients, and amputees retain their personal identity.) It is also motivated by the belief that reversing any type of cryonic preservation is so difficult and complex that any future technology capable of it must by its nature be capable of generalized tissue regeneration, including growth of a new body around a repaired brain. Some suggested revival scenarios for whole body patients even involve discarding the original body and regenerating a new body because tissues are so badly damaged by the preservation process. These considerations, along with lower costs, easier transportation in emergencies, and the specific focus on brain preservation quality, have motivated many cryonicists to choose neuropreservation.

The advantages and disadvantages of neuropreservation are often debated among cryonics advocates. Critics of neuropreservation note that the body is a record of much life experience, including learned motor skills (muscle memory). While few cryonicists doubt that a revived neuro patient would be the same person, there are wider questions about how a regenerated body might feel different from the original. Partly for these reasons (as well as for better public relations), the Cryonics Institute preserves only whole bodies. Some proponents of neuropreservation agree with these concerns, but still feel that lower costs and better brain preservation justify concentrating preservation efforts on the brain. About two-thirds of the patients stored at Alcor are neuropreservation patients. Although the American Cryonics Society no longer offers the neuropreservation option, about half of their patients are "neuros".

Financial issues

Financing storage of a cryonics patient at a cryonics organization by an on-going payment system was done in the early days of cryonics, but this system proved to be unworkable. Cryonics patients are to be stored for many decades, if not a century or two or longer, and a reliable source of outside funding is highly unlikely. Pay-as-you go funding was part of the reason for the CSC Chatsworth financial failure described in the history section. All modern cryonics organizations require full payment for all future costs associated with storage "in perpetuity" before patient cryostorage will be accepted.

Costs of cryonics vary greatly, ranging from the basic fee of $10,000 for neuro (head or brain only) cryopreservation at the European cryonics company KrioRus, to more than

$200,000 for whole body cryopreservation by Alcor with overseas and last-minute fees. Alcor's neuropreservation (just the head) is priced at $80,000. There is an extra $500 annual membership fee during life by Alcor. After payment of an initiaton fee, ACS full members pay an annual fee of $300 currently. To some extent these cost differences reflect differences in how fees are quoted. The Cryonics Institute fee does not include "standby" (a team waits for death to occur and begins procedures at bedside), transportation costs, or funeral director expenses outside of Michigan, which must be purchased as extras.

While cryonics is sometimes suspected of being greatly profitable, the high expenses of doing cryonics are well documented. The expenses are comparable to major transplant surgeries. The two most expensive things are standby expenses (a team of 5+ people needs to be hired for up to several weeks) and the money that must be set aside to generate interest to pay for storage of the patient in liquid nitrogen in perpetuity (especially for whole body patients).

The most common method of paying for cryonics is life insurance, which spreads the cost over many years. Cryonics advocates are quick to point out that such insurance is especially affordable for young people. Cryonics providers claim that even the most expensive cryonics plans are "affordable for the vast majority" of people in the industrialized world who really want it and plan for it in advance. With the advent of low-cost cryonics provided by companies such as KrioRus (so far in Europe only) cryonics becomes feasible even for last-minute cases.

Philosophical and ethical considerations

Cryonics is based on a view of dying as a process that can be stopped in the minutes, and perhaps hours, following clinical death. If death is not an event that happens suddenly when the heart stops, this raises philosophical questions about what exactly death is. In 2005 an ethics debate in the medical journal, Critical Care, noted "…few if any patients pronounced dead by today's physicians are in fact truly dead by any scientifically rigorous criteria." Cryonics proponent Thomas Donaldson has argued that "death" based on cardiac arrest or resuscitation failure is a purely social construction used to justify terminating care of dying patients. In this view, legal death and its aftermath are a form of euthanasia in which sick people are abandoned. Philosopher Max More suggested a distinction between death associated with circumstances and intention versus death that is absolutely irreversible. Absolutely irreversible death has also been called information-theoretic death, which implies destruction of the brain to such an extent that the original information content can no longer be recovered. Bioethicist James Hughes has written that increasing rights will accrue to cryonics patients as prospects for revival become clearer, noting that recovery of legally dead persons has precedent in the discovery of missing persons.

Ethical and theological opinions of cryonics tend to pivot on the issue of whether cryonics is regarded as interment or medicine. If cryonics is interment, then religious beliefs about death and afterlife may come into consideration. Resuscitation may be deemed impossible by those with religious beliefs because the soul is gone, and according to most religions only God can resurrect the dead. Cryonics advocates complain that theological dismissal of cryonics because it is interment is a circular argument because calling cryonics "interment" presumes a priori that cryonics cannot work. They believe future technical advances will validate their view that cryonics patients are recoverable, and therefore never really dead. If cryonics is regarded as medicine, with legal death as a mere enabling mechanism, then cryonics is a long-term coma with uncertain prognosis. It is continuing to care for sick people when others have given up.

Alcor has published a vigorous Christian defense of cryonics, including excerpts of a sermon by Lutheran Reverend Kay Glaesner. Noted Christian commentator John Warwick Montgomery has defended cryonics. In 1969, a Roman Catholic priest consecrated the cryonics capsule of Ann DeBlasio, one of the first cryonics patients. Many followers of Nikolai Fyodorovich Fyodorov see cryonics as an important step in the Common Cause project (reference: Fedorov seminar in Moscow, Russia on 25.11.2006).

At the request of the American Cryonics Society, in 1995, Philosopher Charles Tandy, Ph.D. authored a paper entitled "Cryonic-Hibernation in Light of the Bioethical Principles of Beauchamp and Childress." Dr. Tandy considered the four bioethical factors or principles articulated by philosophers Beauchamp and Childress as they apply to cryonics. These four principles are 1) respect for autonomy; 2) nonmaleficence; 3) beneficence; and 4) justice. Tandy concluded that in respect to all four principles "biomedical professionals have a strong (not weak) and actual (not prima facie, but binding) obligation to help insure cryonic-hibernation of the cryonics patient."

History

Benjamin Franklin, in a 1773 letter, expressed regret that he lived "in a century too little advanced, and too near the infancy of science" that he could not be preserved and revived to fulfil his "very ardent desire to see and observe the state of America a hundred years hence". In 1922 Alexander Yaroslavsky, member of Russian immortalists-biocosmists movement, wrote "Anabiosys Poem". However, the modern era of cryonics began in 1962 when Michigan college physics teacher Robert Ettinger proposed in a privately published book, *The Prospect of Immortality*, that freezing people may be a way to reach future medical technology. Even though freezing a person is apparently fatal, Ettinger argued that what appears to be fatal today may be reversible in the future. He applied the same argument to the process of dying itself, saying that the early stages of clinical death may be reversible in the future. Combining these two ideas, he suggested that freezing recently deceased people may be a way to save lives.

Slightly before Ettinger's book was complete, Evan Cooper (writing as Nathan Duhring) privately published a book called *Immortality: Physically, Scientifically, Now* that independently suggested the same idea. Cooper founded the Life Extension Society (LES) in 1964 to promote freezing people. Ettinger came to be credited as the originator of cryonics, perhaps because his book was republished by Doubleday in 1964 on recommendation of Isaac Asimov and Fred Pohl, and received more publicity. Ettinger also stayed with the movement longer. Nevertheless, cryonics historian R. Michael Perry has written “Evan Cooper deserves the principal credit for forming an organized cryonics movement.”

Cooper’s Life Extension Society became the seed tree for cryonics societies throughout the country where local cryonics advocates would get together as a result of contact through the LES mailing list. The actual word “cryonics” was invented by Karl Werner, then a student in the studio of William Katavolos at Pratt Institute in Brooklyn, NY, in 1965 in conjunction with the founding of the Cryonics Society of New York (CSNY) by Curtis Henderson and Saul Kent that same year. This was followed by the founding of the Cryonics Society of Michigan (CSM) and Cryonics Society of California (CSC) in 1966, and Bay Area Cryonics Society (BACS) in 1969 (renamed the American Cryonics Society, or ACS, in 1985). Neither CSNY nor CSC are currently in operation. CSM eventually became the Immortalist Society, a non-profit affiliate of the Cryonics Institute (CI), a cryonics service organization founded by Ettinger in 1976. CI now has more current cryonics patients than any other organization.

Although there was at least one earlier aborted case, it is generally accepted that the first person frozen with intent of future resuscitation was Dr. James Bedford, a 73-year-old psychology professor frozen under crude conditions by CSC on January 12, 1967. The case made the cover of a limited print run of *Life Magazine* before the presses were stopped to report the death of three astronauts in the *Apollo 1* fire instead. Bedford is still frozen today at Alcor.

Cryonics suffered a major setback in 1979 when it was discovered that nine bodies stored by the head of the CSC, Robert Nelson, in a cemetery in Chatsworth, California, had thawed due to depletion of funds by relatives, after being maintained for a year and a half at the personal expense of Nelson. Some of the bodies had apparently thawed years earlier without notification. Nelson was sued, and negative publicity slowed cryonics growth for years afterward. Of 17 documented cryonics cases between 1967 and 1973, only James Bedford remains cryopreserved today. Strict financial controls and requirements adopted in response to the Chatsworth scandal have resulted in the successful maintenance of almost all cryonics cases since that era.

The largest cryonics organization today, in terms of membership, was established as a nonprofit organization by Fred and Linda Chamberlain in 1972 as the Alcor Society for Solid State Hypothermia (ALCOR). In 1977, the name was changed to the Alcor Life Extension Foundation. In 1982, the Institute for Advanced Biological Studies (IABS), founded by Mike Darwin and Steve Bridge in Indiana, merged with Alcor. During the 1980s, Darwin worked with UCLA cardiothoracic surgery researcher Jerry Leaf at Alcor to develop a medical model for cryonics procedures. They pioneered the first consistent use of a cryonics procedure now known as a “standby”, in which a team waits to begin life support procedures at the bedside of a cryonics patient as soon as possible after the heart stops.

The oldest incorporated cryonics society still in existence is the American Cryonics Society (ACS). This tax-exempt 501(c)(3) membership organization was incorporated in 1969 as the Bay Area Cryonics Society (BACS) by a group of cryonics advocates that included two prominent Bay Area physicians, Dr. M. Coleman Harris and Dr. Grace Talbot. The first suspensions under BACS auspices were performed in 1974 by Trans Time, Inc., a for-profit company started by BACS members. BACS researcher Dr. Paul Segall, working with Jerry Leaf of CryoVita, developed a medical model to induce hypothermia shortly after pronouncement of death. Dr. Segall later went on to pioneer blood substitutes for use in both cryonic suspension and in mainstream medicine.

Cryonics received new support in the 1980s when MIT engineer Eric Drexler started publishing papers and books foreseeing the new field of molecular nanotechnology. His 1986 book, *Engines of Creation*, included an entire chapter on cryonics applications. Cryonics advocates saw the nascent field of nanotechnology as vindication of their long held view that molecular repair of injured tissue was theoretically possible. In the late 1980s Alcor member Dick Clair (who was dying of AIDS) sued for, and ultimately won for everyone, the right to be cryonically preserved in the State of California. Alcor's membership expanded tenfold within a decade, with a 30% annual growth rate between 1988 and 1992.

On July 24, 1988, a Ph.D. in computer science named Kevin Brown started an electronic mailing list called *CryoNet* that became a powerful tool of communication for the cryonics community. Numerous other mailing lists and web forums for discussing cryonics and the affairs of particular organizations have since appeared, but CryoNet remained a central point of contact for cryonicists until it was shut down on March 17, 2011.

Alcor was disrupted by political turmoil in 1993 when a group of activists left to start the CryoCare Foundation, and associated for-profit companies CryoSpan, Inc. (headed by Paul Wakfer) and BioPreservation, Inc. (headed by Mike Darwin). Darwin and collaborators made many technical advances during this time period, including a landmark study documenting high quality brain preservation by freezing with high concentrations of glycerol. CryoCare ceased operations in 1999 when they were unable to renew their service contract with BioPreservation. CryoCare's two patients stored at CryoSpan were transferred to Alcor. Several ACS patients stored at CryoSpan were transferred to CI.

There have been numerous, often transient, for-profit companies involved in cryonics. For-profit companies were often paired or affiliated with non-profit groups they served. Some of these companies, with non-profits they served in parentheses, were Cryonic Interment, Inc. (CSC), Cryo-Span Corporation (CSNY), Cryo-Care Equipment Corporation (CSC and CSNY), Manrise Corporation (Alcor), CryoVita, Inc. (Alcor), BioTransport, Inc. (Alcor), Trans Time, Inc. (BACS), Soma, Inc. (IABS), CryoSpan, Inc. (CryoCare and ACS), BioPreservation, Inc. (CryoCare and ACS), Kryos, Inc. (ACS), Suspended Animation, Inc. (CI, ACS, and Alcor). Trans Time and Suspended Animation are the only for-profit cryonics organizations that still exist.

The cryonics field seems to have largely consolidated around three non-profit groups, Alcor Life Extension Foundation, Cryonics Institute (CI), and the American Cryonics Society (ACS), all deriving significant income from bequests and donations. A newly formed non-profit group called the Cryonics Society was formally incorporated in 2006 but is devoted solely to promotion and public education of the cryonics concept.

As research in the 1990s revealed in greater detail the damaging effects of freezing, there was a trend to use higher concentrations of glycerol cryoprotectant to prevent freezing injury. In 2001 Alcor began using vitrification, a technology borrowed from mainstream organ preservation research, in an attempt to completely prevent ice formation during cooling. Initially the technology could only be applied to the head when separated from the body. In 2005 Alcor began treating the whole body with their vitrification solution in a procedure called "neurovitrification with whole body cryoprotection". In the same year, the Cryonics Institute began treating the head of their whole body patients with their own vitrification solution.

The Cryonics Institute maintains 103 human patients as of 1 March 2011 (along with about 74 pets) at its Clinton Township, Michigan facility. About a fifth of the cryopreserved humans and a smaller portion of the pets came to the CI facility through contract with the American Cryonics Society (which has no storage facilities of its own). Alcor currently maintains 102 cryonics patients in Scottsdale, Arizona. There are support groups in Europe, Canada, the United Kingdom, and Australia. There is also a smaller cryonics company in Russia maintains 15 human patients and 7 pets called KrioRus, and plans for a facility in Australia. There are also plans being developed by renowned architect Stephen Valentine for a multi-acre futuristic high security facility called Timeship to be built in an undisclosed location in the United States, as well as for an underground facility in Switzerland.

Chapter-2

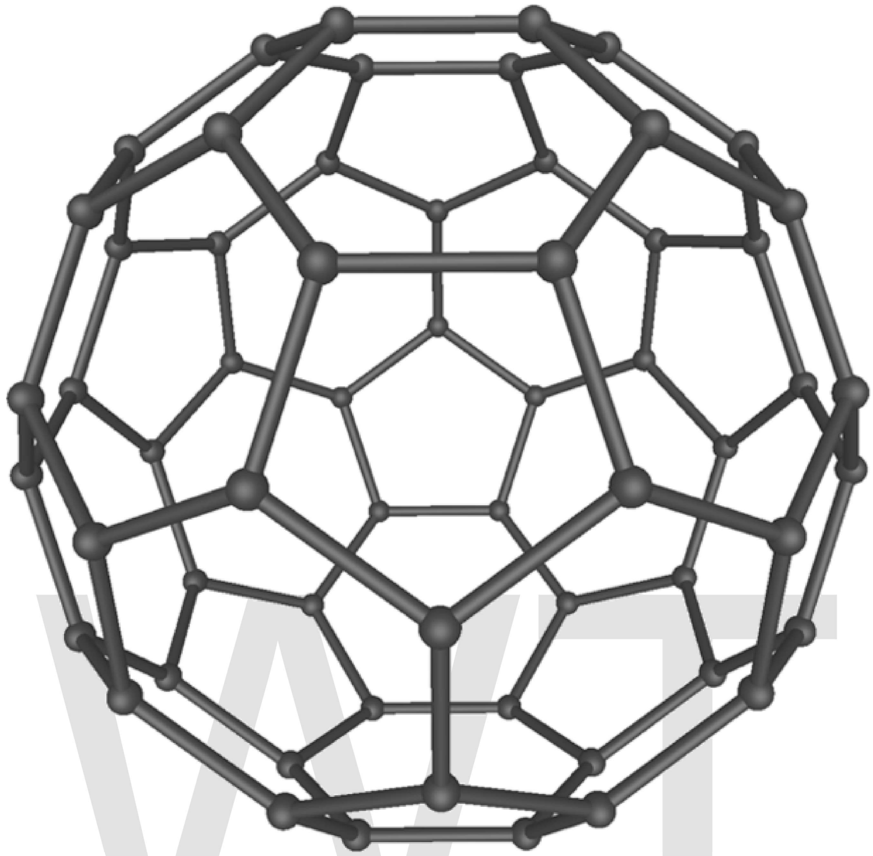
Nanotechnology

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

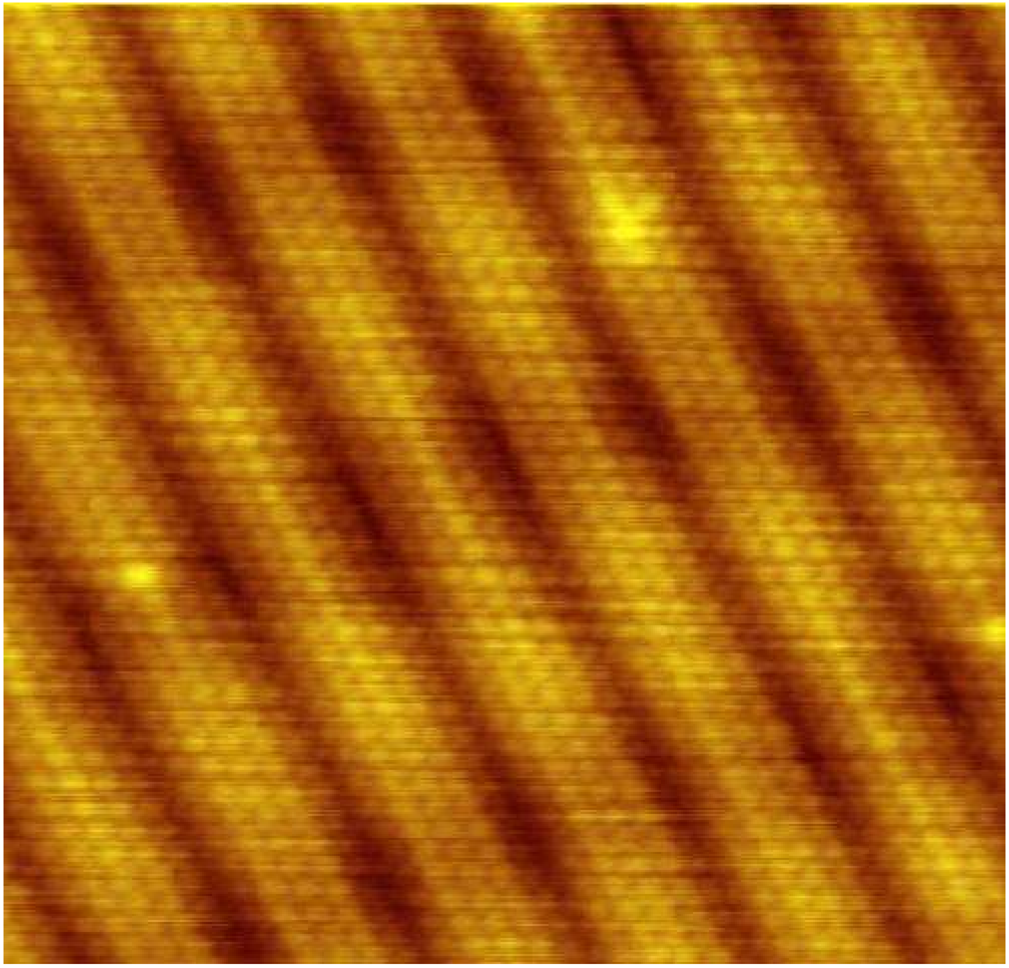


Image of reconstruction on a clean Gold(100) surface, as visualized using scanning tunneling microscopy. The positions of the individual atoms composing the surface are visible.

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.

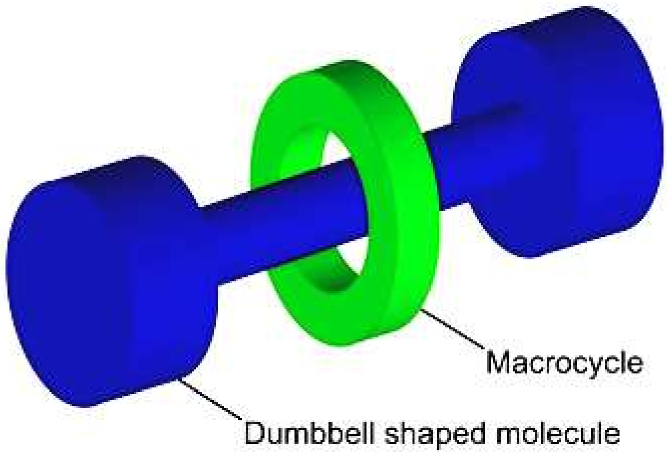
It is hoped that developments in nanotechnology will make possible their construction by some other means, perhaps using biomimetic principles. However, Drexler and other researchers have proposed that advanced nanotechnology, although perhaps initially implemented by biomimetic means, ultimately could 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.

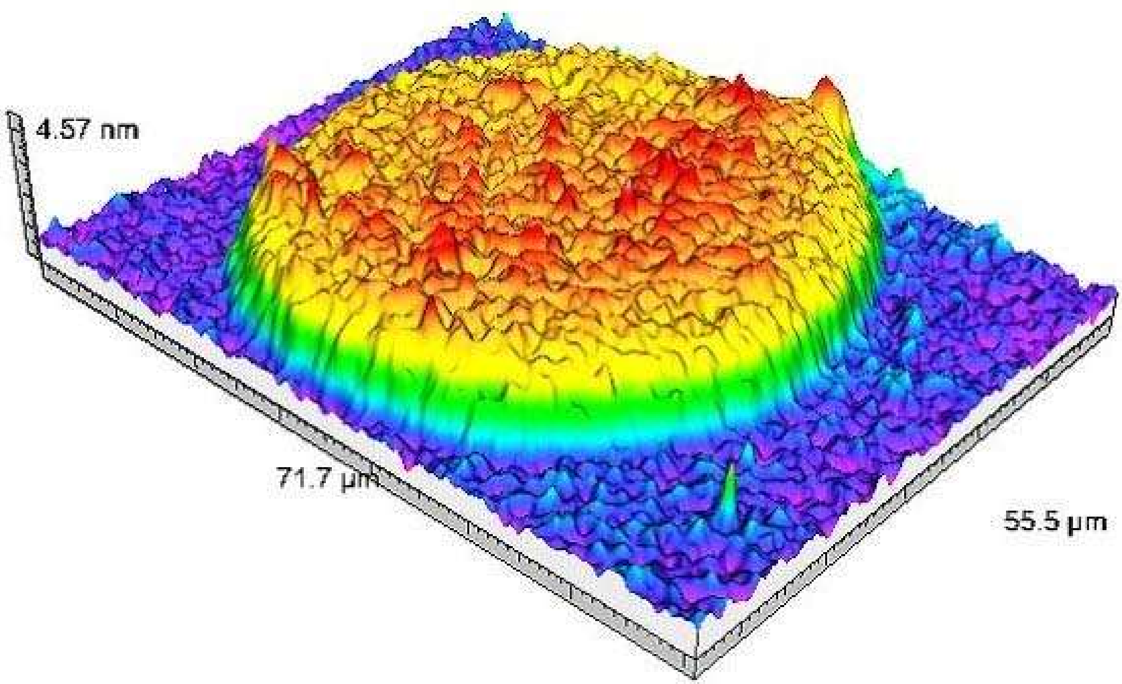
This led to an exchange of letters in the ACS publication Chemical & Engineering News in 2003. Though biology clearly demonstrates that molecular machine systems are possible, non-biological molecular machines are today only in their infancy. Leaders in research on non-biological molecular machines are 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.

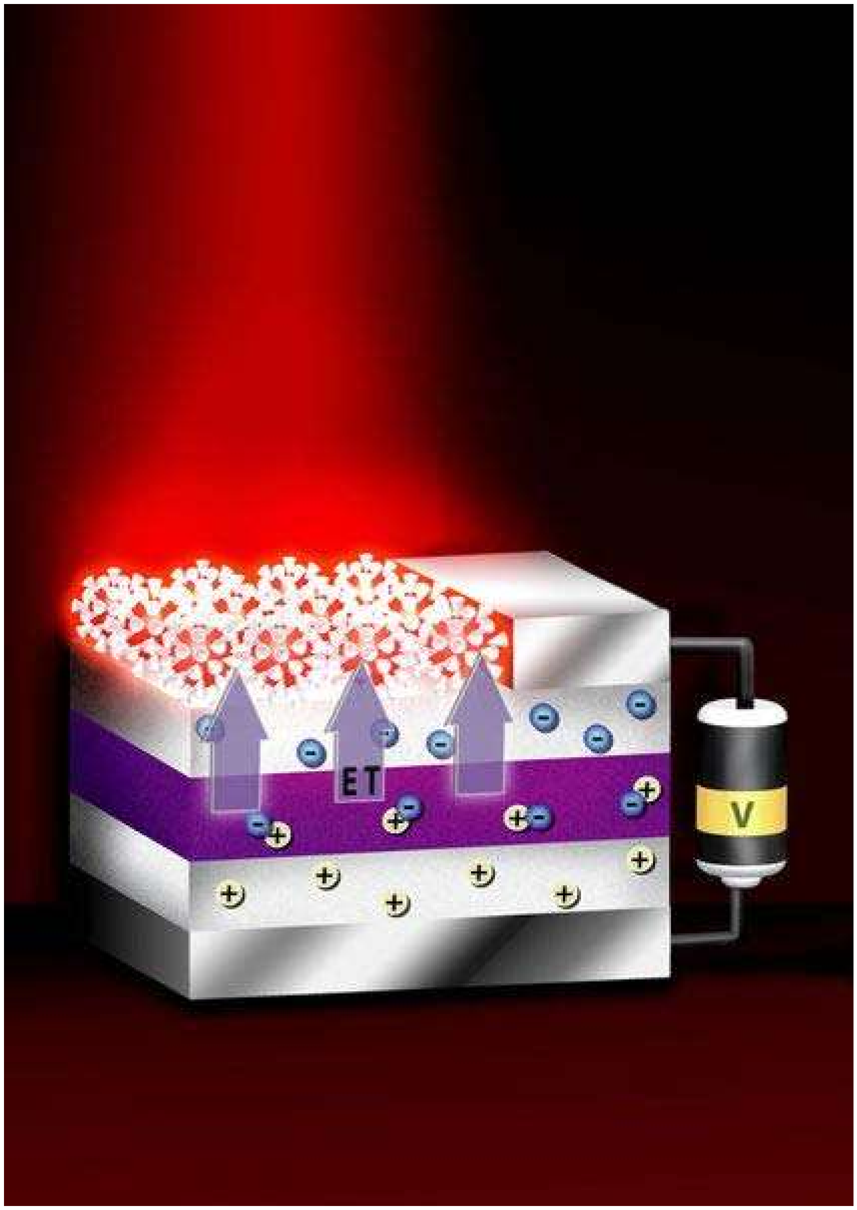
Current research



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

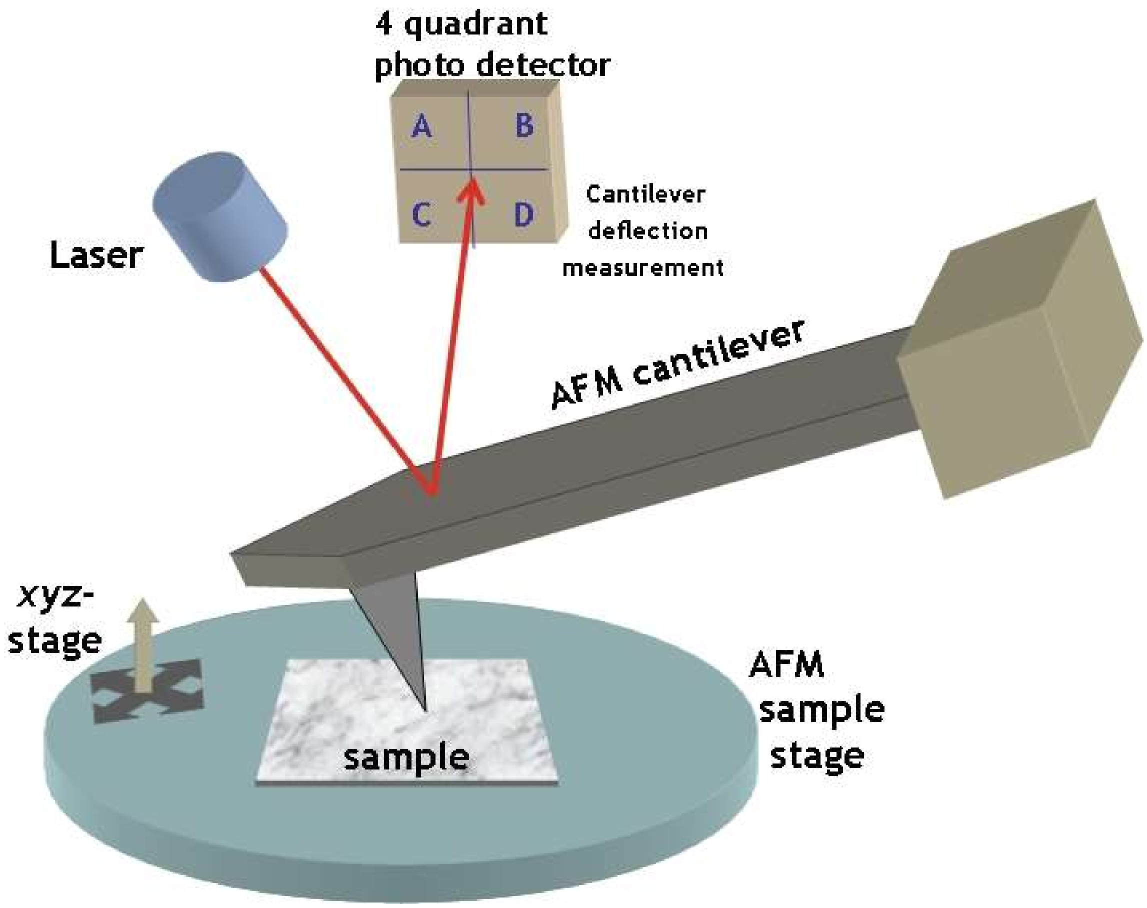
These subfields seek to anticipate what inventions nanotechnology might yield, or attempt to propose an agenda along which inquiry might progress. These often take a big-picture view of nanotechnology, with more emphasis on its societal implications than the details of how such inventions could actually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

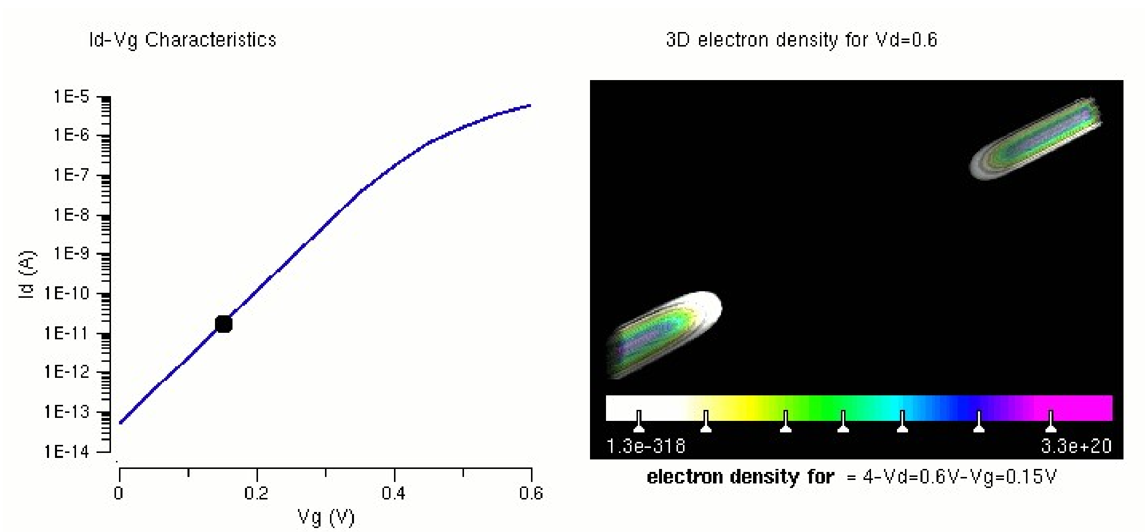
The top-down approach anticipates nanodevices that must be built piece by piece in stages, much as manufactured items are made. Scanning probe microscopy is an important technique both for characterization and synthesis of nanomaterials. Atomic force microscopes and scanning tunneling microscopes can 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-3

Renewable Energy

Burbo Bank Offshore Wind Farm, at the entrance to the River Mersey in North West England.

US President Barack Obama speaks at the DeSoto Next Generation Solar Energy Center, in the USA.

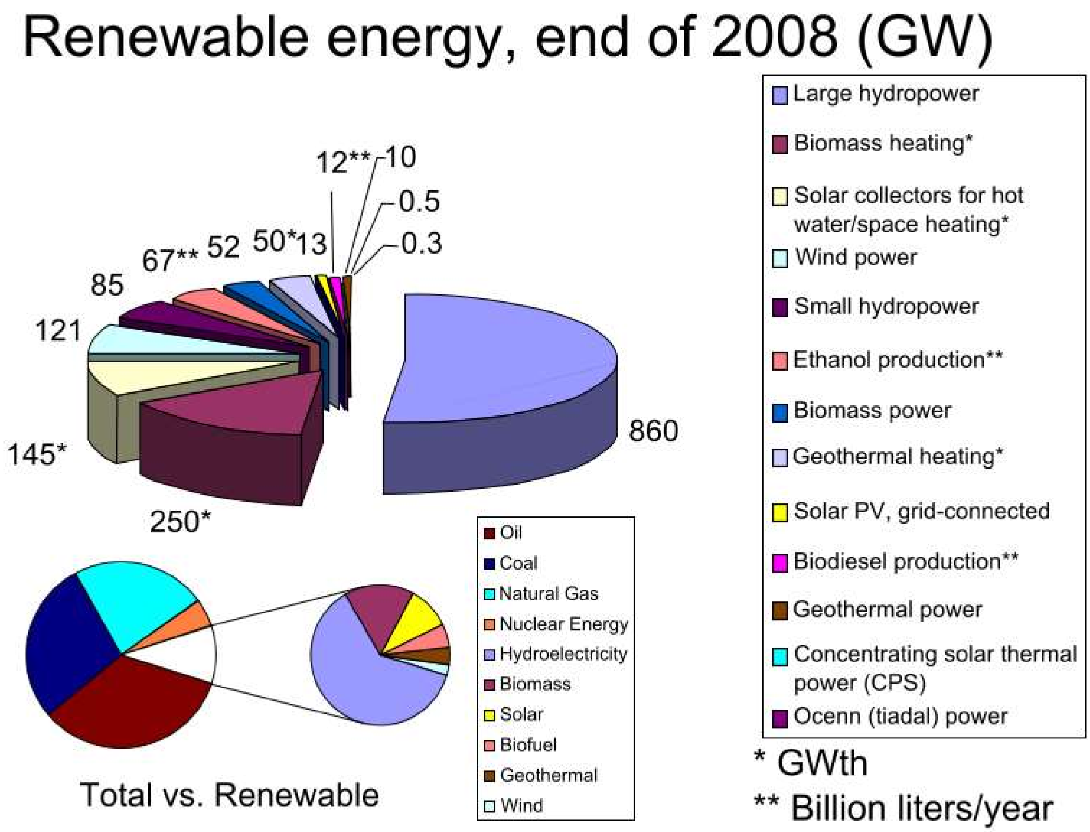
Renewable energy is energy which comes from natural resources such as sunlight, wind, rain, tides, and geothermal heat, which are renewable (naturally replenished). In 2008, about 19% of global final energy consumption came from renewables, with 13% coming from traditional biomass, which is mainly used for heating, and 3.2% from hydroelectricity. New renewables (small hydro, modern biomass, wind, solar, geothermal, and biofuels) accounted for another 2.7% and are growing very rapidly. The share of renewables in electricity generation is around 18%, with 15% of global electricity coming from hydroelectricity and 3% from new renewables.

Wind power is growing at the rate of 30% annually, with a worldwide installed capacity of 158 gigawatts (GW) in 2009, and is widely used in Europe, Asia, and the United States. At the end of 2009, cumulative global photovoltaic (PV) installations surpassed 21 GW and PV power stations are popular in Germany and Spain. Solar thermal power stations operate in the USA and Spain, and the largest of these is the 354 megawatt (MW) SEGS power plant in the Mojave Desert. The world's largest geothermal power installation is The Geysers in California, with a rated capacity of 750 MW. Brazil has one of the largest renewable energy programs in the world, involving production of ethanol fuel from sugar cane, and ethanol now provides 18% of the country's automotive fuel. Ethanol fuel is also widely available in the USA.

While many renewable energy projects are large-scale, renewable technologies are also suited to rural and remote areas, where energy is often crucial in human development. Globally, an estimated 3 million households get power from small solar PV systems. Micro-hydro systems configured into village-scale or county-scale mini-grids serve many areas. More than 30 million rural households get lighting and cooking from biogas made in household-scale digesters. Biomass cookstoves are used by 160 million households.

Climate change concerns, coupled with high oil prices, peak oil, and increasing government support, are driving increasing renewable energy legislation, incentives and commercialization. New government spending, regulation and policies helped the industry weather the global financial crisis better than many other sectors.

Overview



2008 worldwide renewable-energy sources.

Renewable energy flows involve natural phenomena such as sunlight, wind, tides, plant growth, and geothermal heat, as the International Energy Agency explains:

Renewable energy is derived from natural processes that are replenished constantly. In its various forms, it derives directly from the sun, or from heat generated deep within the earth. Included in the definition is electricity and heat generated from solar, wind, ocean, hydropower, biomass, geothermal resources, and biofuels and hydrogen derived from renewable resources.

Renewable energy replaces conventional fuels in four distinct areas: power generation, hot water/ space heating, transport fuels, and rural (off-grid) energy services:

- **Power generation**. Renewable energy provides 18 percent of total electricity generation worldwide. Renewable power generators are spread across many countries, and wind power alone already provides a significant share of electricity in some areas: for example, 14 percent in the U.S. state of Iowa, 40 percent in the northern German state of Schleswig-Holstein, and 20 percent in Denmark. Some countries get most of their power from renewables, including Iceland (100 percent), Brazil (85 percent), Austria (62 percent), New Zealand (65 percent), and Sweden (54 percent).

- **Heating**. Solar hot water makes an important contribution in many countries, most notably in China, which now has 70 percent of the global total (180 GWth). Most of these systems are installed on multi-family apartment buildings and meet a portion of the hot water needs of an estimated 50–60 million households in China. Worldwide, total installed solar water heating systems meet a portion of the water heating needs of over 70 million households. The use of biomass for heating continues to grow as well. In Sweden, national use of biomass energy has surpassed that of oil. Direct geothermal for heating is also growing rapidly.

- **Transport fuels**. Renewable biofuels have contributed to a significant decline in oil consumption in the United States since 2006. The 93 billion liters of biofuels produced worldwide in 2009 displaced the equivalent of an estimated 68 billion liters of gasoline, equal to about 5 percent of world gasoline production.

Mainstream forms of renewable energy

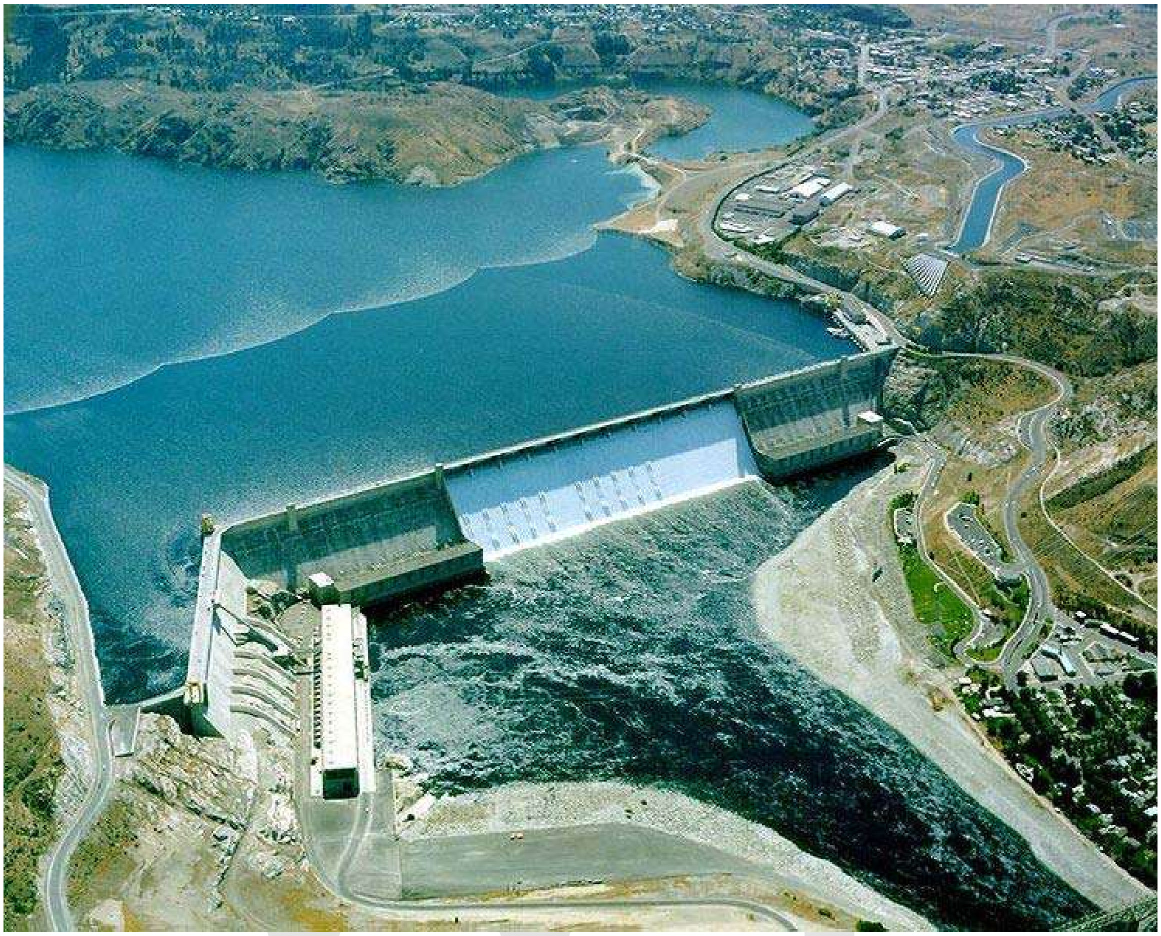
Wind power

Wind Turbines located outside of Palm Springs, California

Airflows can be used to run wind turbines. Modern wind turbines range from around 600 kW to 5 MW of rated power, although turbines with rated output of 1.5–3 MW have become the most common for commercial use; the power output of a turbine is a function of the cube of the wind speed, so as wind speed increases, power output increases dramatically. Areas where winds are stronger and more constant, such as offshore and high altitude sites, are preferred locations for wind farms. Typical capacity factors are 20-40%, with values at the upper end of the range in particularly favourable sites.

Globally, the long-term technical potential of wind energy is believed to be five times total current global energy production, or 40 times current electricity demand. This could require wind turbines to be installed over large areas, particularly in areas of higher wind resources. Offshore resources experience mean wind speeds of ~90% greater than that of land, so offshore resources could contribute substantially more energy.

Hydropower



Grand Coulee Dam is a hydroelectric gravity dam on the Columbia River in the U.S. state of Washington. The dam supplies four power stations with an installed capacity of 6,809 MW and is the largest electric power-producing facility in the United States.

Energy in water can be harnessed and used. Since water is about 800 times denser than air, even a slow flowing stream of water, or moderate sea swell, can yield considerable amounts of energy. There are many forms of water energy:

- Hydroelectric energy is a term usually reserved for large-scale hydroelectric dams. Examples are the Grand Coulee Dam in Washington State and the Akosombo Dam in Ghana.
- Micro hydro systems are hydroelectric power installations that typically produce up to 100 kW of power. They are often used in water rich areas as a remote-area power supply (RAPS). There are many of these installations around the world, including several delivering around 50 kW in the Solomon Islands.
- Run-of-the-river hydroelectricity systems derive kinetic energy from rivers and oceans without using a dam.

- Ocean energy describes all the technologies to harness energy from the ocean and the sea. This includes marine current power, ocean thermal energy conversion, and tidal power.

Solar energy

Monocrystalline solar cell.

Solar energy is the energy derived from the sun through the form of solar radiation. Solar powered electrical generation relies on photovoltaics and heat engines. A partial list of other solar applications includes space heating and cooling through solar architecture, daylighting, solar hot water, solar cooking, and high temperature process heat for industrial purposes.

Solar technologies are broadly characterized as either passive solar or active solar depending on the way they capture, convert and distribute solar energy. Active solar techniques include the use of photovoltaic panels and solar thermal collectors to harness the energy. Passive solar techniques include orienting a building to the Sun, selecting materials with favorable thermal mass or light dispersing properties, and designing spaces that naturally circulate air.

Biomass

Biomass (plant material) is a renewable energy source because the energy it contains comes from the sun. Through the process of photosynthesis, plants capture the sun's energy. When the plants are burned, they release the sun's energy they contain. In this way, biomass functions as a sort of natural battery for storing solar energy. As long as biomass is produced sustainably, with only as much used as is grown, the battery will last indefinitely.

In general there are two main approaches to using plants for energy production: growing plants specifically for energy use, and using the residues from plants that are used for other things. The best approaches vary from region to region according to climate, soils and geography.

Biofuel

Information on pump regarding ethanol fuel blend up to 10%, California.

Liquid biofuel is usually either bioalcohol such as bioethanol or an oil such as biodiesel.

Bioethanol is an alcohol made by fermenting the sugar components of plant materials and it is made mostly from sugar and starch crops. With advanced technology being developed, cellulosic biomass, such as trees and grasses, are also used as feedstocks for ethanol production. Ethanol can be used as a fuel for vehicles in its pure form, but it is usually used as a gasoline additive to increase octane and improve vehicle emissions. Bioethanol is widely used in the USA and in Brazil.

Biodiesel is made from vegetable oils, animal fats or recycled greases. Biodiesel can be used as a fuel for vehicles in its pure form, but it is usually used as a diesel additive to reduce levels of particulates, carbon monoxide, and hydrocarbons from diesel-powered vehicles. Biodiesel is produced from oils or fats using transesterification and is the most common biofuel in Europe.

Biofuels provided 1.8% of the world's transport fuel in 2008.

Geothermal energy

Krafla Geothermal Station in northeast Iceland

Geothermal energy is energy obtained by tapping the heat of the earth itself, both from kilometers deep into the Earth's crust in volcanically active locations of the globe or from shallow depths, as in geothermal heat pumps in most locations of the planet. It is expensive to build a power station but operating costs are low resulting in low energy costs for suitable sites. Ultimately, this energy derives from heat in the Earth's core.

Three types of power plants are used to generate power from geothermal energy: dry steam, flash, and binary. Dry steam plants take steam out of fractures in the ground and use it to directly drive a turbine that spins a generator. Flash plants take hot water, usually at temperatures over 200 °C, out of the ground, and allows it to boil as it rises to the surface then separates the steam phase in steam/water separators and then runs the steam through a turbine. In binary plants, the hot water flows through heat exchangers, boiling an organic fluid that spins the turbine. The condensed steam and remaining geothermal fluid from all three types of plants are injected back into the hot rock to pick up more heat.

The geothermal energy from the core of the Earth is closer to the surface in some areas than in others. Where hot underground steam or water can be tapped and brought to the surface it may be used to generate electricity. Such geothermal power sources exist in certain geologically unstable parts of the world such as Chile, Iceland, New Zealand, United States, the Philippines and Italy. The two most prominent areas for this in the United States are in the Yellowstone basin and in northern California. Iceland produced 170 MW geothermal power and heated 86% of all houses in the year 2000 through geothermal energy. Some 8000 MW of capacity is operational in total.

There is also the potential to generate geothermal energy from hot dry rocks. Holes at least 3 km deep are drilled into the earth. Some of these holes pump water into the earth, while other holes pump hot water out. The heat resource consists of hot underground radiogenic granite rocks, which heat up when there is enough sediment between the rock and the earths surface. Several companies in Australia are exploring this technology.

Renewable energy commercialization

Growth of renewables

During the five-years from the end of 2004 through 2009, worldwide renewable energy capacity grew at rates of 10–60 percent annually for many technologies. For wind power and many other renewable technologies, growth accelerated in 2009 relative to the previous four years. More wind power capacity was added during 2009 than any other renewable technology. However, grid-connected PV increased the fastest of all renewables technologies, with a 60-percent annual average growth rate for the five-year period.

Selected renewable energy indicators

Selected global indicators	2007	2008	2009
Investment in new renewable capacity (annual)	104	130	150 billion USD
Existing renewables power capacity,	1,070	1,140	1,230 GWe

including large-scale hydro			
Existing renewables power capacity, excluding large hydro	240	280	305 GWe
Wind power capacity (existing)	94	121	159 GWe
Solar PV capacity (grid-connected)	7.6	13.5	21 GWe
Solar hot water capacity	126	149	180 GWth
Ethanol production (annual)	50	69	76 billion liters
Biodiesel production (annual)	10	15	17 billion liters
Countries with policy targets for renewable energy use	68	75	85

Scientists have advanced a plan to power 100% of the world's energy with wind, hydroelectric, and solar power by the year 2030, recommending renewable energy subsidies and a price on carbon reflecting its cost for flood and related expenses.

Economic trends

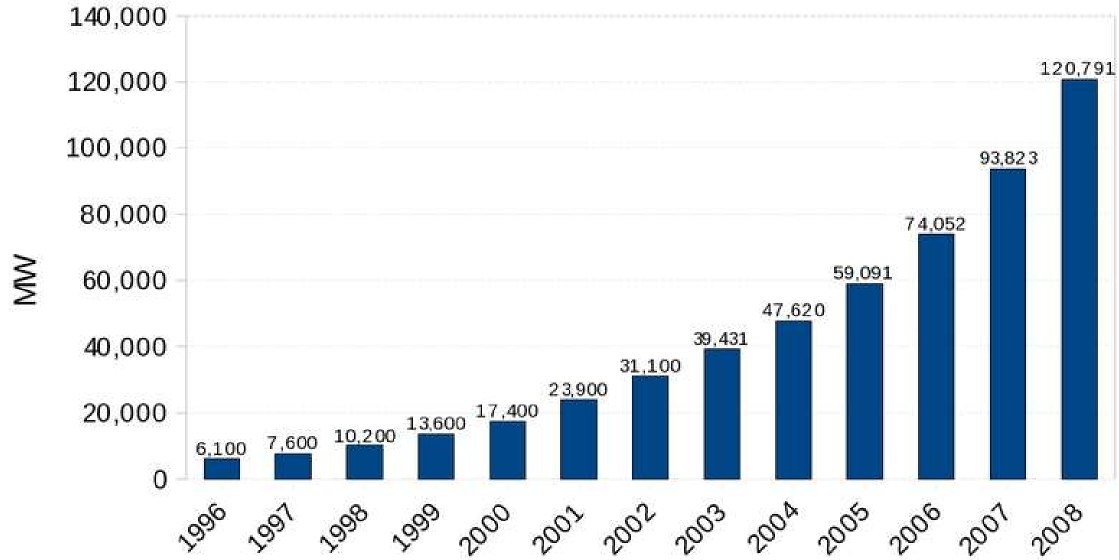
All forms of energy are expensive, but as time progresses, renewable energy generally gets cheaper, while fossil fuels generally get more expensive. Al Gore has explained that renewable energy technologies are declining in price for three main reasons:

First, once the renewable infrastructure is built, the fuel is free forever. Unlike carbon-based fuels, the wind and the sun and the earth itself provide fuel that is free, in amounts that are effectively limitless.

Second, while fossil fuel technologies are more mature, renewable energy technologies are being rapidly improved. So innovation and ingenuity give us the ability to constantly increase the efficiency of renewable energy and continually reduce its cost.

Third, once the world makes a clear commitment to shifting toward renewable energy, the volume of production will itself sharply reduce the cost of each windmill and each solar panel, while adding yet more incentives for additional research and development to further speed up the innovation process.

Wind power market



Wind power: worldwide installed capacity 1996-2008

Fenton Wind Farm at sunrise

Global wind power installations increased by 35,800 MW in 2010, bringing total installed capacity up to 194,400 MW, a 22.5% increase on the 158,700 MW installed at the end of 2009. For the first time more than half of all new wind power was added outside of the traditional markets of Europe and North America, mainly driven, by the continuing boom in China which accounted for nearly half of all of the installations at 16,500 MW. China now has 42,300 MW of wind power installed. Wind power accounts for approximately 19% of electricity generated in Denmark, 9% in Spain and Portugal, and 6% in Germany and the Republic of Ireland.

Top 10 wind power countries

Country	Total capacity end 2009 (MW)	Total capacity June 2010 (MW)
United States	35,159	36,300
China	26,010	33,800
Germany	25,777	26,400
Spain	19,149	19,500
India	10, 925	12,100
Italy	4,850	5,300
France	4,521	5,000
United Kingdom	4,092	4,600
Portugal	3,535	3,800
Denmark	3,497	3,700
Rest of world	21,698	24,500
Total	**159,213**	**175,000**

As of November 2010, the Roscoe Wind Farm (781 MW) is the world's largest wind farm. As of September 2010, the Thanet Offshore Wind Project in United Kingdom is the largest offshore wind farm in the world at 300 MW, followed by Horns Rev II (209 MW) in Denmark. The United Kingdom is the world's leading generator of offshore wind power, followed by Denmark.

New generation of solar thermal plants

Solar Towers from left: PS10, PS20.

Large solar thermal power stations include the 354 megawatt (MW) Solar Energy Generating Systems power plant in the USA, Solnova Solar Power Station (Spain, 150 MW), Andasol solar power station (Spain, 100 MW), Nevada Solar One (USA, 64 MW), PS20 solar power tower (Spain, 20 MW), and the PS10 solar power tower (Spain, 11 MW).

The solar thermal power industry is growing rapidly with 1.2 GW under construction as of April 2009 and another 13.9 GW announced globally through 2014. Spain is the epicenter of solar thermal power development with 22 projects for 1,037 MW under construction, all of which are projected to come online by the end of 2010. In the United States, 5,600 MW of solar thermal power projects have been announced. In developing countries, three World Bank projects for integrated solar thermal/combined-cycle gas-turbine power plants in Egypt, Mexico, and Morocco have been approved.

Photovoltaic market

40 MW PV Array installed in Waldpolenz, Germany

Photovoltaic production has been increasing by an average of some 20 percent each year since 2002, making it a fast-growing energy technology. At the end of 2009, the cumulative global PV installations surpassed 21,000 megawatts.

As of November 2010, the largest photovoltaic (PV) power plants in the world are the Finsterwalde Solar Park (Germany, 80.7 MW), Sarnia Photovoltaic Power Plant (Canada, 80 MW), Olmedilla Photovoltaic Park (Spain, 60 MW), the Strasskirchen Solar Park (Germany, 54 MW), the Lieberose Photovoltaic Park (Germany, 53 MW), and the Puertollano Photovoltaic Park (Spain, 50 MW). Many of these plants are integrated with agriculture and some use innovative tracking systems that follow the sun's daily path across the sky to generate more electricity than conventional fixed-mounted systems. There are no fuel costs or emissions during operation of the power stations.

Topaz Solar Farm is a proposed 550 MW solar photovoltaic power plant which is to be built northwest of California Valley in the USA at a cost of over $1 billion. High Plains Ranch is a proposed 250 MW solar photovoltaic power plant which is to be built on the Carrizo Plain, northwest of California Valley.

However, when it comes to renewable energy systems and PV, it is not just large systems that matter. Building-integrated photovoltaics or "onsite" PV systems use existing land and structures and generate power close to where it is consumed.

Use of ethanol for transportation

E95 trial bus operating in São Paulo, Brazil.

Since the 1970s, Brazil has had an ethanol fuel program which has allowed the country to become the world's second largest producer of ethanol (after the United States) and the world's largest exporter. Brazil's ethanol fuel program uses modern equipment and cheap sugar cane as feedstock, and the residual cane-waste (bagasse) is used to process heat and power. There are no longer light vehicles in Brazil running on pure gasoline. By the end of 2008 there were 35,000 filling stations throughout Brazil with at least one ethanol pump.

Nearly all the gasoline sold in the United States today is mixed with 10 percent ethanol, a mix known as E10, and motor vehicle manufacturers already produce vehicles designed to run on much higher ethanol blends. Ford, DaimlerChrysler, and GM are among the automobile companies that sell "flexible-fuel" cars, trucks, and minivans that can use gasoline and ethanol blends ranging from pure gasoline up to 85% ethanol (E85). By mid-2006, there were approximately six million E85-compatible vehicles on U.S. roads. The challenge is to expand the market for biofuels beyond the farm states where they have been most popular to date. Flex-fuel vehicles are assisting in this transition because they allow drivers to choose different fuels based on price and availability. The *Energy Policy Act of 2005*, which calls for 7.5 billion gallons of biofuels to be used annually by 2012, will also help to expand the market.

Geothermal energy commercialization

The West Ford Flat power plant is one of 22 power plants at The Geysers.

The International Geothermal Association (IGA) has reported that 10,715 megawatts (MW) of geothermal power in 24 countries is online, which is expected to generate 67,246 GWh of electricity in 2010. This represents a 20% increase in geothermal power online capacity since 2005. IGA projects this will grow to 18,500 MW by 2015, due to the large number of projects presently under consideration, often in areas previously assumed to have little exploitable resource.

In 2010, the United States led the world in geothermal electricity production with 3,086 MW of installed capacity from 77 power plants; the largest group of geothermal power plants in the world is located at The Geysers, a geothermal field in California. The

Philippines follows the US as the second highest producer of geothermal power in the world, with 1,904 MW of capacity online; geothermal power makes up approximately 18% of the country's electricity generation.

Geothermal (ground source) heat pumps represented an estimated 30 GWth of installed capacity at the end of 2008, with other direct uses of geothermal heat (i.e., for space heating, agricultural drying and other uses) reaching an estimated 15 GWth. As of 2008, at least 76 countries use direct geothermal energy in some form.

Wave farms expansion

One of 3 Pelamis Wave Energy Converters in the harbor of Peniche, Portugal

Portugal now has the world's first commercial wave farm, the *Agucadoura Wave Park*, officially opened in September 2008. The farm uses three Pelamis P-750 machines generating 2.25 MW. Initial costs are put at € 8.5 million. A second phase of the project is now planned to increase the installed capacity to 21MW using a further 25 Pelamis machines.

Funding for a wave farm in Scotland was announced in February, 2007 by the Scottish Government, at a cost of over 4 million pounds, as part of a UK£13 million funding packages for ocean power in Scotland. The farm will be the world's largest with a capacity of 3MW generated by four Pelamis machines.

Developing country markets

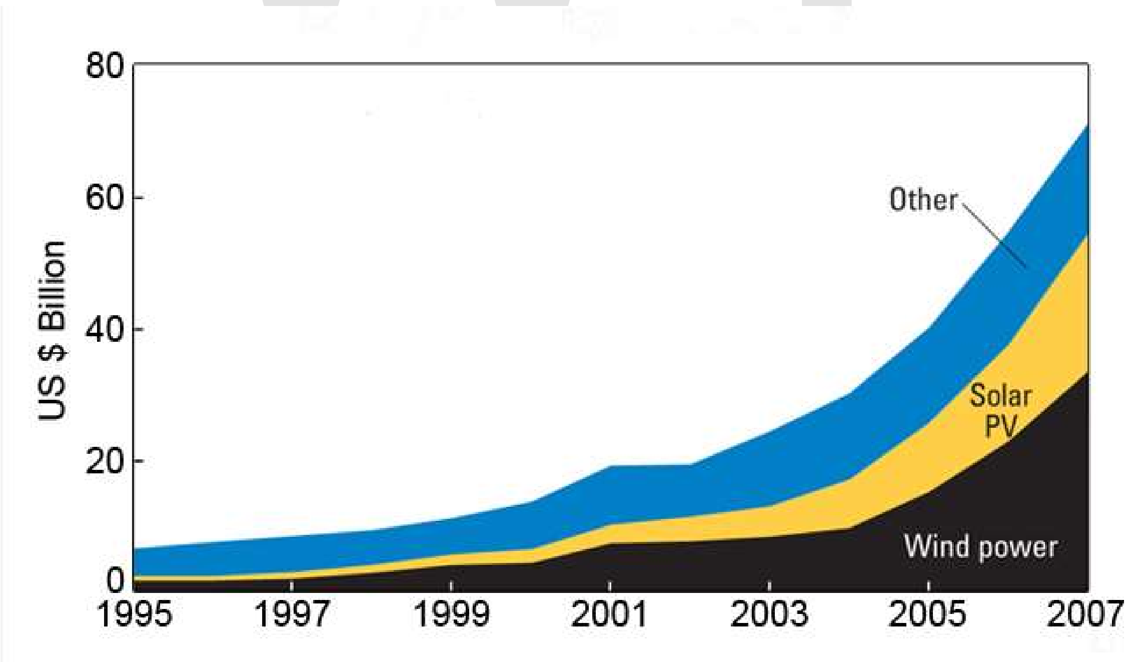
Renewable energy can be particularly suitable for developing countries. In rural and remote areas, transmission and distribution of energy generated from fossil fuels can be difficult and expensive. Producing renewable energy locally can offer a viable alternative.

Biomass cookstoves are used by 40 percent of the world's population. These stoves are being manufactured in factories and workshops worldwide, and more than 160 million households now use them. More than 30 million rural households get lighting and cooking from biogas made in household-scale digesters. An estimated 3 million households get power from small solar PV systems. Micro-hydro systems configured into village-scale or county-scale mini-grids serve many areas.

Kenya is the world leader in the number of solar power systems installed per capita. More than 30,000 very small solar panels, each producing 12 to 30 watts, are sold in Kenya annually.

Renewable energy projects in many developing countries have demonstrated that renewable energy can directly contribute to poverty alleviation by providing the energy needed for creating businesses and employment. Renewable energy technologies can also make indirect contributions to alleviating poverty by providing energy for cooking, space heating, and lighting. Renewable energy can also contribute to education, by providing electricity to schools.

Industry and policy trends



Global renewable energy investment growth (1995-2007)

Global revenues for solar photovoltaics, wind power, and biofuels expanded from $76 billion in 2007 to $115 billion in 2008. New global investments in clean energy technologies expanded by 4.7 percent from $148 billion in 2007 to $155 billion in 2008. U.S. President Barack Obama's American Recovery and Reinvestment Act of 2009 includes more than $70 billion in direct spending and tax credits for clean energy and associated transportation programs. Clean Edge suggests that the commercialization of clean energy will help countries around the world pull out of the current economic malaise. Leading renewable energy companies include First Solar, Gamesa, GE Energy, Q-Cells, Sharp Solar, Siemens, SunOpta, Suntech, and Vestas.

The International Renewable Energy Agency (IRENA) is an intergovernmental organization for promoting the adoption of renewable energy worldwide. It aims to provide concrete policy advice and facilitate capacity building and technology transfer. IRENA was formed on January 26, 2009, by 75 countries signing the charter of IRENA. As of March 2010, IRENA has 143 member states who all are considered as founding members, of which 14 have also ratified the statute.

Renewable energy policy targets exist in some 73 countries around the world, and public policies to promote renewable energy use have become more common in recent years. At least 64 countries have some type of policy to promote renewable power generation. Mandates for solar hot water in new construction are becoming more common at both national and local levels. Mandates for blending biofuels into vehicle fuels have been enacted in 17 countries.

New and emerging renewable energy technologies

New and emerging renewable energy technologies are still under development and include cellulosic ethanol, hot-dry-rock geothermal power, and ocean energy. These technologies are not yet widely demonstrated or have limited commercialization. Many are on the horizon and may have potential comparable to other renewable energy technologies, but still depend on attracting sufficient attention and research, development and demonstration (RD&D) funding.

Cellulosic ethanol

Companies such as Iogen, Broin, and Abengoa are building refineries that can process biomass and turn it into ethanol, while companies such as Diversa, Novozymes, and Dyadic are producing enzymes which could enable a cellulosic ethanol future. The shift from food crop feedstocks to waste residues and native grasses offers significant opportunities for a range of players, from farmers to biotechnology firms, and from project developers to investors.

Selected Commercial Cellulosic Ethanol Plants in the U.S.
(Operational or under construction)

Company	Location	Feedstock
Abengoa Bioenergy	Hugoton, KS	Wheat straw
BlueFire Ethanol	Irvine, CA	Multiple sources
Gulf Coast Energy	Mossy Head, FL	Wood waste
Mascoma	Lansing, MI	Wood
POET LLC	Emmetsburg, IA	Corn cobs
SunOpta	Little Falls, MN	Wood chips
Xethanol	Auburndale, FL	Citrus peels

Ocean energy

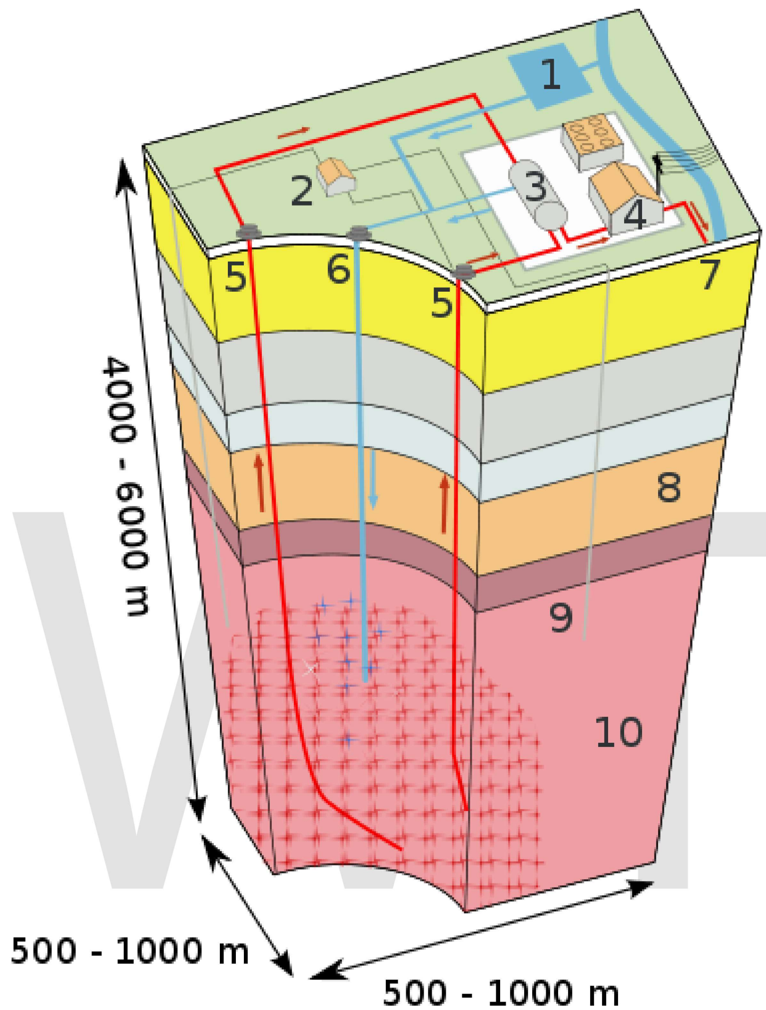
Systems to harvest utility-scale electrical power from ocean waves have recently been gaining momentum as a viable technology. The potential for this technology is considered promising, especially on west-facing coasts with latitudes between 40 and 60 degrees:

In the United Kingdom, for example, the Carbon Trust recently estimated the extent of the economically viable offshore resource at 55 TWh per year, about 14% of current national demand. Across Europe, the technologically achievable resource has been estimated to be at least 280 TWh per year. In 2003, the U.S. Electric Power Research Institute (EPRI) estimated the viable resource in the United States at 255 TWh per year (6% of demand).

The world's first commercial tidal power station was installed in 2007 in the narrows of Strangford Lough in Ireland. The 1.2 megawatt underwater tidal electricity generator, part of Northern Ireland's Environment & Renewable Energy Fund scheme, takes advantage of the fast tidal flow (up to 4 metres per second) in the lough. Although the generator is powerful enough to power a thousand homes, the turbine has minimal environmental impact, as it is almost entirely submerged, and the rotors pose no danger to wildlife as they turn quite slowly.

Ocean thermal energy conversion (OTEC) uses the temperature difference that exists between deep and shallow waters to run a heat engine.

Enhanced Geothermal Systems



Enhanced geothermal system 1:Reservoir 2:Pump house 3:Heat exchanger 4:Turbine hall 5:Production well 6:Injection well 7:Hot water to district heating 8:Porous sediments 9:Observation well 10:Crystalline bedrock

Enhanced Geothermal Systems are a new type of geothermal power technologies that do not require natural convective hydrothermal resources. The vast majority of geothermal energy within drilling reach is in dry and non-porous rock. EGS technologies "enhance" and/or create geothermal resources in this "hot dry rock (HDR)" through hydraulic stimulation.

EGS / HDR technologies, like hydrothermal geothermal, are expected to be baseload resources which produce power 24 hours a day like a fossil plant. Distinct from hydrothermal, HDR / EGS may be feasible anywhere in the world, depending on the economic limits of drill depth. Good locations are over deep granite covered by a thick (3–5 km) layer of insulating sediments which slow heat loss.

There are HDR and EGS systems currently being developed and tested in France, Australia, Japan, Germany, the U.S. and Switzerland. The largest EGS project in the world is a 25 megawatt demonstration plant currently being developed in the Cooper Basin, Australia. The Cooper Basin has the potential to generate 5,000–10,000 MW.

Nanotechnology thin-film solar panels

Solar power panels that use nanotechnology, which can create circuits out of individual silicon molecules, may cost half as much as traditional photovoltaic cells, according to executives and investors involved in developing the products. Nanosolar has secured more than $100 million from investors to build a factory for nanotechnology thin-film solar panels.

Renewable energy debate

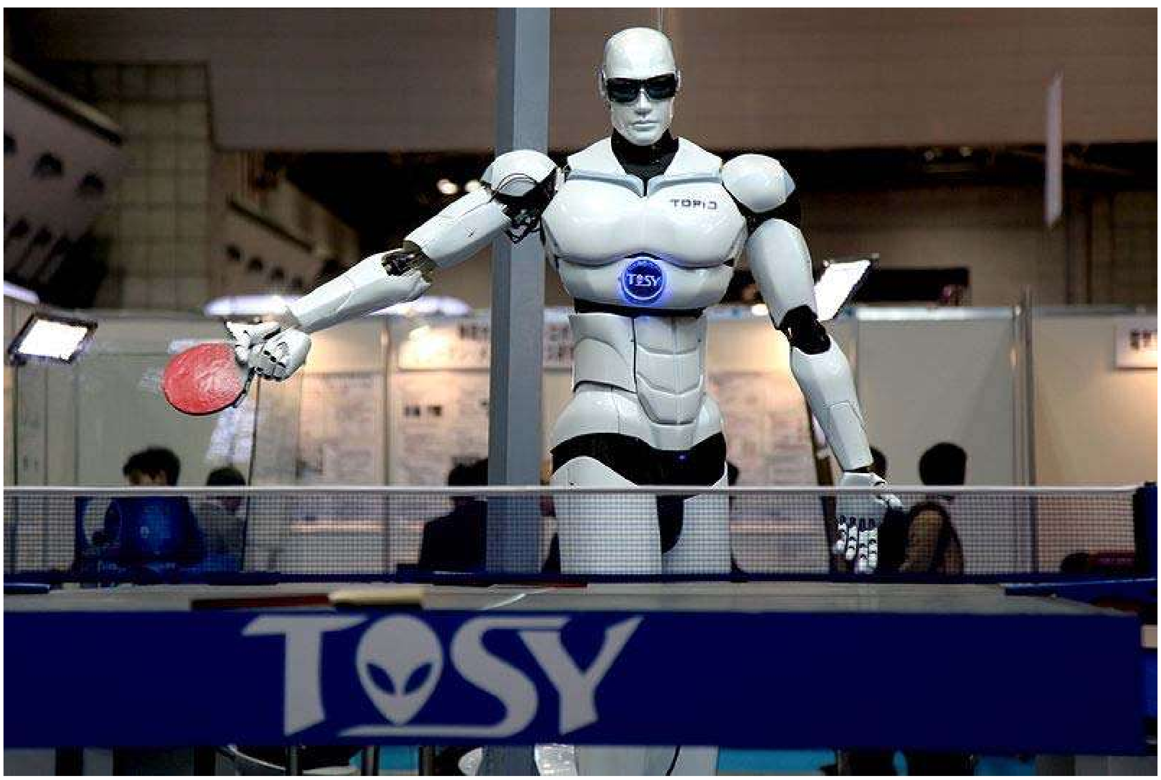
Renewable electricity production, from sources such as wind power and solar power, is sometimes criticized for being variable or intermittent. However, the International Energy Agency has stated that deployment of renewable technologies usually increases the diversity of electricity sources and, through local generation, contributes to the flexibility of the system and its resistance to central shocks.

There have been "not in my back yard" (NIMBY) concerns relating to the visual and other impacts of some wind farms, with local residents sometimes fighting or blocking construction. In the USA, the Massachusetts Cape Wind project was delayed for years partly because of aesthetic concerns. However, residents in other areas have been more positive and there are many examples of community wind farm developments. According to a town councilor, the overwhelming majority of locals believe that the Ardrossan Wind Farm in Scotland has enhanced the area.

The market for renewable energy technologies has continued to grow. Climate change concerns, coupled with high oil prices, peak oil, and increasing government support, are driving increasing renewable energy legislation, incentives and commercialization. New government spending, regulation and policies helped the industry weather the 2009 economic crisis better than many other sectors.

Chapter-4

Artificial Intelligence



TOPIO, a humanoid robot, played table tennis at Tokyo International Robot Exhibition (IREX) 2009.

Artificial intelligence (**AI**) is the intelligence of machines and the branch of computer science that aims to create it. AI textbooks define the field as "the study and design of intelligent agents" where an intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success. John McCarthy, who coined the term in 1956, defines it as "the science and engineering of making intelligent machines."

The field was founded on the claim that a central property of humans, intelligence—the sapience of *Homo sapiens*—can be so precisely described that it can be simulated by a machine. This raises philosophical issues about the nature of the mind and the ethics of creating artificial beings, issues which have been addressed by myth, fiction and philosophy since antiquity. Artificial intelligence has been the subject of optimism, but has also suffered setbacks and, today, has become an essential part of the technology industry, providing the heavy lifting for many of the most difficult problems in computer science.

AI research is highly technical and specialized, deeply divided into subfields that often fail to communicate with each other. Subfields have grown up around particular institutions, the work of individual researchers, the solution of specific problems, longstanding differences of opinion about how AI should be done and the application of widely differing tools. The central problems of AI include such traits as reasoning, knowledge, planning, learning, communication, perception and the ability to move and manipulate objects. General intelligence (or "strong AI") is still among the field's long term goals.

History

Thinking machines and artificial beings appear in Greek myths, such as Talos of Crete, the bronze robot of Hephaestus and Pygmalion's Galatea. Human likenesses believed to have intelligence were built in every major civilization: animated cult images were worshipped in Egypt and Greece and humanoid automatons were built by Yan Shi, Hero of Alexandria and Al-Jazari. It was also widely believed that artificial beings had been created by Jābir ibn Hayyān, Judah Loew and Paracelsus. By the 19th and 20th centuries, artificial beings had become a common feature in fiction, as in Mary Shelley's *Frankenstein* or Karel Čapek's *R.U.R. (Rossum's Universal Robots)*. Pamela McCorduck argues that all of these are examples of an ancient urge, as she describes it, "to forge the gods". Stories of these creatures and their fates discuss many of the same hopes, fears and ethical concerns that are presented by artificial intelligence.

Mechanical or "formal" reasoning has been developed by philosophers and mathematicians since antiquity. The study of logic led directly to the invention of the programmable digital electronic computer, based on the work of mathematician Alan Turing and others. Turing's theory of computation suggested that a machine, by shuffling symbols as simple as "0" and "1", could simulate any conceivable act of mathematical deduction. This, along with recent discoveries in neurology, information theory and cybernetics, inspired a small group of researchers to begin to seriously consider the possibility of building an electronic brain.

The field of AI research was founded at a conference on the campus of Dartmouth College in the summer of 1956. The attendees, including John McCarthy, Marvin Minsky, Allen Newell and Herbert Simon, became the leaders of AI research for many decades. They and their students wrote programs that were, to most people, simply astonishing: computers were solving word problems in algebra, proving logical theorems and speaking English. By the middle of the 1960s, research in the U.S. was heavily funded by the Department of Defense and laboratories had been established around the world. AI's founders were profoundly optimistic about the future of the new field: Herbert Simon predicted that "machines will be capable, within twenty years, of doing any work a man can do" and Marvin Minsky agreed, writing that "within a generation ... the problem of creating 'artificial intelligence' will substantially be solved".

They had failed to recognize the difficulty of some of the problems they faced. In 1974, in response to the criticism of England's Sir James Lighthill and ongoing pressure from Congress to fund more productive projects, the U.S. and British governments cut off all undirected, exploratory research in AI. The next few years, when funding for projects was hard to find, would later be called an "AI winter".

In the early 1980s, AI research was revived by the commercial success of expert systems, a form of AI program that simulated the knowledge and analytical skills of one or more human experts. By 1985 the market for AI had reached over a billion dollars. At the same time, Japan's fifth generation computer project inspired the U.S and British governments to restore funding for academic research in the field. However, beginning with the collapse of the Lisp Machine market in 1987, AI once again fell into disrepute, and a second, longer lasting AI winter began.

In the 1990s and early 21st century, AI achieved its greatest successes, albeit somewhat behind the scenes. Artificial intelligence is used for logistics, data mining, medical diagnosis and many other areas throughout the technology industry. The success was due to several factors: the increasing computational power of computers, a greater emphasis on solving specific subproblems, the creation of new ties between AI and other fields working on similar problems, and a new commitment by researchers to solid mathematical methods and rigorous scientific standards.

The artificial intelligence quiz show contestant "Watson", appearing on the US quiz show *Jeopardy!* in 2011.

On 11 May 1997, Deep Blue became the first computer chess-playing system to beat a reigning world chess champion, Garry Kasparov. In 2005, a Stanford robot won the DARPA Grand Challenge by driving autonomously for 131 miles along an unrehearsed desert trail. In February 2011, in a Jeopardy! quiz show exhibition match, IBM's question answering system, Watson, defeated the two greatest Jeopardy! champions, Brad Rutter and Ken Jennings, by a significant margin.

The leading-edge definition of artificial intelligence research is changing over time. One pragmatic definition is: "AI research is that which computing scientists do not know how to do cost-effectively today." For example, in 1956 optical character recognition (OCR) was considered AI, but today, sophisticated OCR software with a context-sensitive spell checker and grammar checker software comes for free with most image scanners. No one would any longer consider already-solved computing science problems like OCR "artificial intelligence" today.

Low-cost entertaining chess-playing software is commonly available for tablet computers. DARPA no longer provides significant funding for chess-playing computing system development. The child's XBox 360 gaming system November 2010 Kinect 3D-body-motion interface uses algorithms that emerged from lengthy AI research, but few consumers realize the technology source.

AI applications are no longer the exclusive domain of Department of defense R&D, but are now common place consumer items and inexpensive intelligent toys.

In common usage, the term "AI" no longer seems to apply to off-the-shelf solved computing-science problems, which may have originally emerged out of years of AI research.

Problems

The general problem of simulating (or creating) intelligence has been broken down into a number of specific sub-problems. These consist of particular traits or capabilities that researchers would like an intelligent system to display. The traits described below have received the most attention.

Deduction, reasoning, problem solving

Early AI researchers developed algorithms that imitated the step-by-step reasoning that humans were often assumed to use when they solve puzzles, play board games or make logical deductions. By the late 1980s and '90s, AI research had also developed highly successful methods for dealing with uncertain or incomplete information, employing concepts from probability and economics.

For difficult problems, most of these algorithms can require enormous computational resources — most experience a "combinatorial explosion": the amount of memory or computer time required becomes astronomical when the problem goes beyond a certain size. The search for more efficient problem solving algorithms is a high priority for AI research.

Human beings solve most of their problems using fast, intuitive judgments rather than the conscious, step-by-step deduction that early AI research was able to model. AI has made some progress at imitating this kind of "sub-symbolic" problem solving: embodied agent approaches emphasize the importance of sensorimotor skills to higher reasoning; neural net research attempts to simulate the structures inside human and animal brains that give rise to this skill.

Knowledge representation

Knowledge representation and knowledge engineering are central to AI research. Many of the problems machines are expected to solve will require extensive knowledge about the world. Among the things that AI needs to represent are: objects, properties, categories and relations between objects; situations, events, states and time; causes and effects; knowledge about knowledge (what we know about what other people know); and many other, less well researched domains. A complete representation of "what exists" is an ontology (borrowing a word from traditional philosophy), of which the most general are called upper ontologies.

Among the most difficult problems in knowledge representation are:

Default reasoning and the qualification problem

: Many of the things people know take the form of "working assumptions." For example, if a bird comes up in conversation, people typically picture an animal that is fist sized, sings, and flies. None of these things are true about all birds. John McCarthy identified this problem in 1969 as the qualification problem: for any commonsense rule that AI researchers care to represent, there tend to be a huge number of exceptions. Almost nothing is simply true or false in the way that abstract logic requires. AI research has explored a number of solutions to this problem.

The breadth of commonsense knowledge

: The number of atomic facts that the average person knows is astronomical. Research projects that attempt to build a complete knowledge base of commonsense knowledge (e.g., Cyc) require enormous amounts of laborious ontological engineering — they must be built, by hand, one complicated concept at a time. A major goal is to have the computer understand enough concepts to be able to learn by reading from sources like the internet, and thus be able to add to its own ontology.

The subsymbolic form of some commonsense knowledge

: Much of what people know is not represented as "facts" or "statements" that they could express verbally. For example, a chess master will avoid a particular chess position because it "feels too exposed" or an art critic can take one look at a statue and instantly realize that it is a fake. These are intuitions or tendencies that are represented in the brain non-consciously and sub-symbolically. Knowledge like this informs, supports and provides a context for symbolic, conscious knowledge. As with the related problem of sub-symbolic reasoning, it is hoped that situated AI or computational intelligence will provide ways to represent this kind of knowledge.

Planning

Intelligent agents must be able to set goals and achieve them. They need a way to visualize the future (they must have a representation of the state of the world and be able to make predictions about how their actions will change it) and be able to make choices that maximize the utility (or "value") of the available choices.

In classical planning problems, the agent can assume that it is the only thing acting on the world and it can be certain what the consequences of its actions may be. However, if this is not true, it must periodically check if the world matches its predictions and it must change its plan as this becomes necessary, requiring the agent to reason under uncertainty.

Multi-agent planning uses the cooperation and competition of many agents to achieve a given goal. Emergent behavior such as this is used by evolutionary algorithms and swarm intelligence.

Learning

Machine learning has been central to AI research from the beginning. In 1956, at the original Dartmouth AI summer conference, Ray Solomonoff wrote a report on unsupervised probabilistic machine learning: "An Inductive Inference Machine". Unsupervised learning is the ability to find patterns in a stream of input. Supervised learning includes both classification and numerical regression. Classification is used to determine what category something belongs in, after seeing a number of examples of things from several categories. Regression takes a set of numerical input/output examples and attempts to discover a continuous function that would generate the outputs from the inputs. In reinforcement learning the agent is rewarded for good responses and punished for bad ones. These can be analyzed in terms of decision theory, using concepts like utility. The mathematical analysis of machine learning algorithms and their performance is a branch of theoretical computer science known as computational learning theory.

Natural language processing

ASIMO uses sensors and intelligent algorithms to avoid obstacles and navigate stairs.

Natural language processing gives machines the ability to read and understand the languages that humans speak. Many researchers hope that a sufficiently powerful natural language processing system would be able to acquire knowledge on its own, by reading the existing text available over the internet. Some straightforward applications of natural language processing include information retrieval (or text mining) and machine translation.

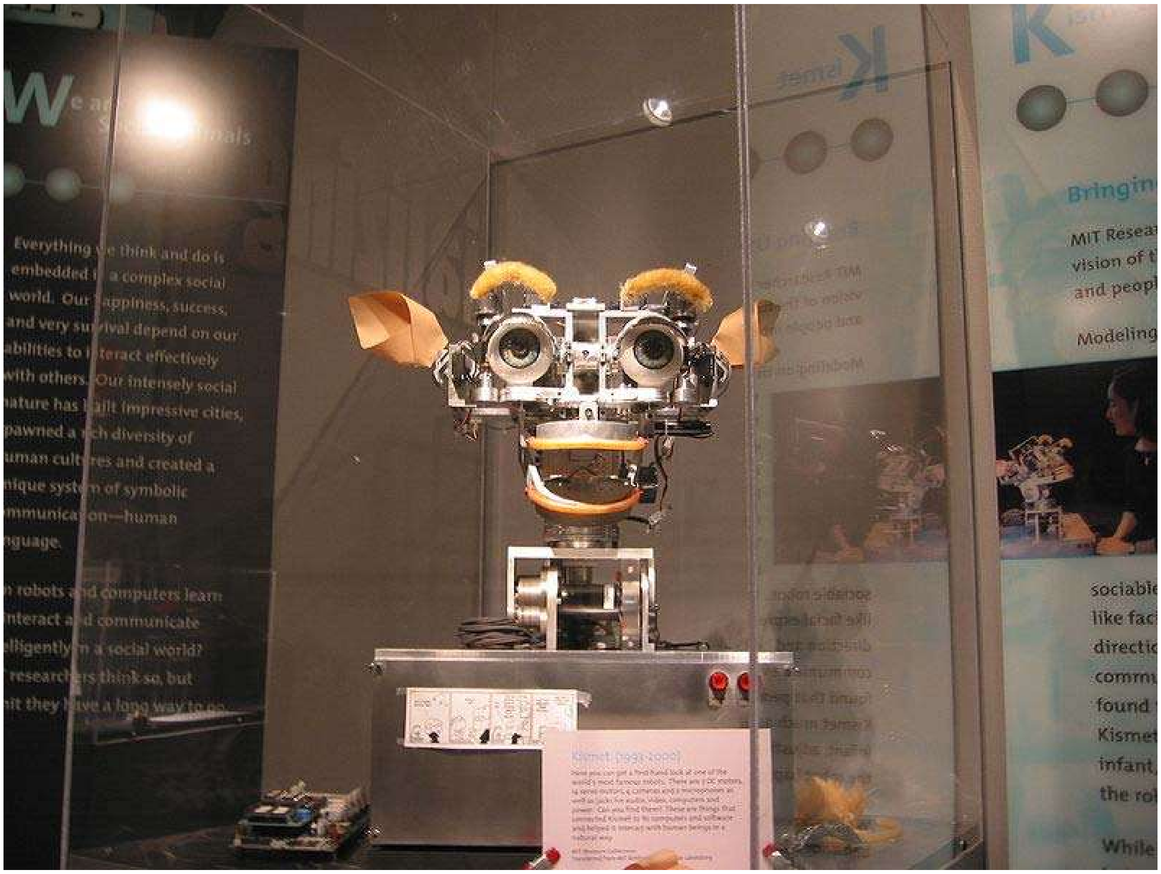
Motion and manipulation

The field of robotics is closely related to AI. Intelligence is required for robots to be able to handle such tasks as object manipulation and navigation, with sub-problems of localization (knowing where you are), mapping (learning what is around you) and motion planning (figuring out how to get there).

Perception

Machine perception is the ability to use input from sensors (such as cameras, microphones, sonar and others more exotic) to deduce aspects of the world. Computer vision is the ability to analyze visual input. A few selected subproblems are speech recognition, facial recognition and object recognition.

Social intelligence



Kismet, a robot with rudimentary social skills

Emotion and social skills play two roles for an intelligent agent. First, it must be able to predict the actions of others, by understanding their motives and emotional states. (This involves elements of game theory, decision theory, as well as the ability to model human emotions and the perceptual skills to detect emotions.) Also, for good human-computer interaction, an intelligent machine also needs to *display* emotions. At the very least it must appear polite and sensitive to the humans it interacts with. At best, it should have normal emotions itself.

Creativity

A sub-field of AI addresses creativity both theoretically (from a philosophical and psychological perspective) and practically (via specific implementations of systems that generate outputs that can be considered creative, or systems that identify and assess creativity). A related area of computational research is Artificial Intuition and Artificial Imagination.

General intelligence

Most researchers hope that their work will eventually be incorporated into a machine with *general* intelligence (known as strong AI), combining all the skills above and exceeding human abilities at most or all of them. A few believe that anthropomorphic features like artificial consciousness or an artificial brain may be required for such a project.

Many of the problems above are considered AI-complete: to solve one problem, you must solve them all. For example, even a straightforward, specific task like machine translation requires that the machine follow the author's argument (reason), know what is being talked about (knowledge), and faithfully reproduce the author's intention (social intelligence). Machine translation, therefore, is believed to be AI-complete: it may require strong AI to be done as well as humans can do it.

Approaches

There is no established unifying theory or paradigm that guides AI research. Researchers disagree about many issues. A few of the most long standing questions that have remained unanswered are these: should artificial intelligence simulate natural intelligence, by studying psychology or neurology? Or is human biology as irrelevant to AI research as bird biology is to aeronautical engineering? Can intelligent behavior be described using simple, elegant principles (such as logic or optimization)? Or does it necessarily require solving a large number of completely unrelated problems? Can intelligence be reproduced using high-level symbols, similar to words and ideas? Or does it require "sub-symbolic" processing? John Haugeland, who coined the term GOFAI, also proposed that AI should more properly be referred to as synthetic intelligence, a term which has since been adopted by some non-GOFAI researchers.

Cybernetics and brain simulation

In the 1940s and 1950s, a number of researchers explored the connection between neurology, information theory, and cybernetics. Some of them built machines that used electronic networks to exhibit rudimentary intelligence, such as W. Grey Walter's turtles and the Johns Hopkins Beast. Many of these researchers gathered for meetings of the Teleological Society at Princeton University and the Ratio Club in England. By 1960, this approach was largely abandoned, although elements of it would be revived in the 1980s.

Symbolic

When access to digital computers became possible in the middle 1950s, AI research began to explore the possibility that human intelligence could be reduced to symbol manipulation. The research was centered in three institutions: CMU, Stanford and MIT, and each one developed its own style of research. John Haugeland named these approaches to AI "good old fashioned AI" or "GOFAI".

Cognitive simulation

Economist Herbert Simon and Allen Newell studied human problem solving skills and attempted to formalize them, and their work laid the foundations of the field of artificial intelligence, as well as cognitive science, operations research and management science. Their research team used the results of psychological experiments to develop programs that simulated the techniques that people used to solve problems. This tradition, centered at Carnegie Mellon University would eventually culminate in the development of the Soar architecture in the middle 80s.

Logic based

Unlike Newell and Simon, John McCarthy felt that machines did not need to simulate human thought, but should instead try to find the essence of abstract reasoning and problem solving, regardless of whether people used the same algorithms. His laboratory at Stanford (SAIL) focused on using formal logic to solve a wide variety of problems, including knowledge representation, planning and learning. Logic was also focus of the work at the University of Edinburgh and elsewhere in Europe which led to the development of the programming language Prolog and the science of logic programming.

"Anti-logic" or "scruffy"

Researchers at MIT (such as Marvin Minsky and Seymour Papert) found that solving difficult problems in vision and natural language processing required ad-hoc solutions – they argued that there was no simple and general principle (like logic) that would capture all the aspects of intelligent behavior. Roger Schank described their "anti-logic" approaches as "scruffy" (as opposed to the "neat" paradigms at CMU and Stanford). Commonsense knowledge bases (such as Doug Lenat's Cyc) are an example of "scruffy" AI, since they must be built by hand, one complicated concept at a time.

Knowledge based

When computers with large memories became available around 1970, researchers from all three traditions began to build knowledge into AI applications. This "knowledge revolution" led to the development and deployment of expert systems (introduced by Edward Feigenbaum), the first truly successful form of AI software. The knowledge revolution was also driven by the realization that enormous amounts of knowledge would be required by many simple AI applications.

Sub-symbolic

During the 1960s, symbolic approaches had achieved great success at simulating high-level thinking in small demonstration programs. Approaches based on cybernetics or neural networks were abandoned or pushed into the background. By the 1980s, however, progress in symbolic AI seemed to stall and many believed that symbolic systems would never be able to imitate all the processes of human cognition, especially perception, robotics, learning and pattern recognition. A number of researchers began to look into "sub-symbolic" approaches to specific AI problems.

Bottom-up, embodied, situated, behavior-based or nouvelle AI

Researchers from the related field of robotics, such as Rodney Brooks, rejected symbolic AI and focused on the basic engineering problems that would allow robots to move and survive. Their work revived the non-symbolic viewpoint of the early cybernetics researchers of the 50s and reintroduced the use of control theory in AI. This coincided with the development of the embodied mind thesis in the related field of cognitive science: the idea that aspects of the body (such as movement, perception and visualization) are required for higher intelligence.

Computational Intelligence

Interest in neural networks and "connectionism" was revived by David Rumelhart and others in the middle 1980s. These and other sub-symbolic approaches, such as fuzzy systems and evolutionary computation, are now studied collectively by the emerging discipline of computational intelligence.

Statistical

In the 1990s, AI researchers developed sophisticated mathematical tools to solve specific subproblems. These tools are truly scientific, in the sense that their results are both measurable and verifiable, and they have been responsible for many of AI's recent successes. The shared mathematical language has also permitted a high level of collaboration with more established fields (like mathematics, economics or operations research). Stuart Russell and Peter Norvig describe this movement as nothing less than a "revolution" and "the victory of the neats."

Integrating the approaches

Intelligent agent paradigm

An intelligent agent is a system that perceives its environment and takes actions which maximizes its chances of success. The simplest intelligent agents are programs that solve specific problems. More complicated agents include human beings and organizations of human beings (such as firms). The paradigm gives researchers license to study isolated problems and find solutions that are both verifiable and useful, without agreeing on one single approach. An agent that solves a specific problem can use any approach that works — some agents are symbolic and logical, some are sub-symbolic neural networks and others may use new approaches. The paradigm also gives researchers a common language to communicate with other fields—such as decision theory and economics—that also use concepts of abstract agents. The intelligent agent paradigm became widely accepted during the 1990s.

Agent architectures and cognitive architectures

Researchers have designed systems to build intelligent systems out of interacting intelligent agents in a multi-agent system. A system with both symbolic and sub-symbolic components is a hybrid intelligent system, and the study of such systems is artificial intelligence systems integration. A hierarchical control system provides a bridge between sub-symbolic AI at its lowest, reactive levels and traditional symbolic AI at its highest levels, where relaxed time constraints permit planning and world modelling. Rodney Brooks' subsumption architecture was an early proposal for such a hierarchical system.

Tools

In the course of 50 years of research, AI has developed a large number of tools to solve the most difficult problems in computer science. A few of the most general of these methods are discussed below.

Search and optimization

Many problems in AI can be solved in theory by intelligently searching through many possible solutions: Reasoning can be reduced to performing a search. For example, logical proof can be viewed as searching for a path that leads from premises to conclusions, where each step is the application of an inference rule. Planning algorithms search through trees of goals and subgoals, attempting to find a path to a target goal, a process called means-ends analysis. Robotics algorithms for moving limbs and grasping objects use local searches in configuration space. Many learning algorithms use search algorithms based on optimization.

Simple exhaustive searches are rarely sufficient for most real world problems: the search space (the number of places to search) quickly grows to astronomical numbers. The result is a search that is too slow or never completes. The solution, for many problems, is to use "heuristics" or "rules of thumb" that eliminate choices that are unlikely to lead to the goal (called "pruning the search tree"). Heuristics supply the program with a "best guess" for what path the solution lies on.

A very different kind of search came to prominence in the 1990s, based on the mathematical theory of optimization. For many problems, it is possible to begin the search with some form of a guess and then refine the guess incrementally until no more refinements can be made. These algorithms can be visualized as blind hill climbing: we begin the search at a random point on the landscape, and then, by jumps or steps, we keep moving our guess uphill, until we reach the top. Other optimization algorithms are simulated annealing, beam search and random optimization.

Evolutionary computation uses a form of optimization search. For example, they may begin with a population of organisms (the guesses) and then allow them to mutate and recombine, selecting only the fittest to survive each generation (refining the guesses). Forms of evolutionary computation include swarm intelligence algorithms (such as ant colony or particle swarm optimization) and evolutionary algorithms (such as genetic algorithms and genetic programming).

Logic

Logic is used for knowledge representation and problem solving, but it can be applied to other problems as well. For example, the satplan algorithm uses logic for planning and inductive logic programming is a method for learning.

Several different forms of logic are used in AI research. Propositional or sentential logic is the logic of statements which can be true or false. First-order logic also allows the use of quantifiers and predicates, and can express facts about objects, their properties, and their relations with each other. Fuzzy logic, is a version of first-order logic which allows the truth of a statement to be represented as a value between 0 and 1, rather than simply True (1) or False (0). Fuzzy systems can be used for uncertain reasoning and have been widely used in modern industrial and consumer product control systems. Subjective logic models uncertainty in a different and more explicit manner than fuzzy-logic: a given binomial opinion satisfies belief + disbelief + uncertainty = 1 within a Beta distribution. By this method, ignorance can be distinguished from probabilistic statements that an agent makes with high confidence.

Default logics, non-monotonic logics and circumscription are forms of logic designed to help with default reasoning and the qualification problem. Several extensions of logic have been designed to handle specific domains of knowledge, such as: description logics; situation calculus, event calculus and fluent calculus (for representing events and time); causal calculus; belief calculus; and modal logics.

Probabilistic methods for uncertain reasoning

Many problems in AI (in reasoning, planning, learning, perception and robotics) require the agent to operate with incomplete or uncertain information. AI researchers have devised a number of powerful tools to solve these problems using methods from probability theory and economics.

Bayesian networks are a very general tool that can be used for a large number of problems: reasoning (using the Bayesian inference algorithm), learning (using the expectation-maximization algorithm), planning (using decision networks) and perception (using dynamic Bayesian networks). Probabilistic algorithms can also be used for filtering, prediction, smoothing and finding explanations for streams of data, helping perception systems to analyze processes that occur over time (e.g., hidden Markov models or Kalman filters).

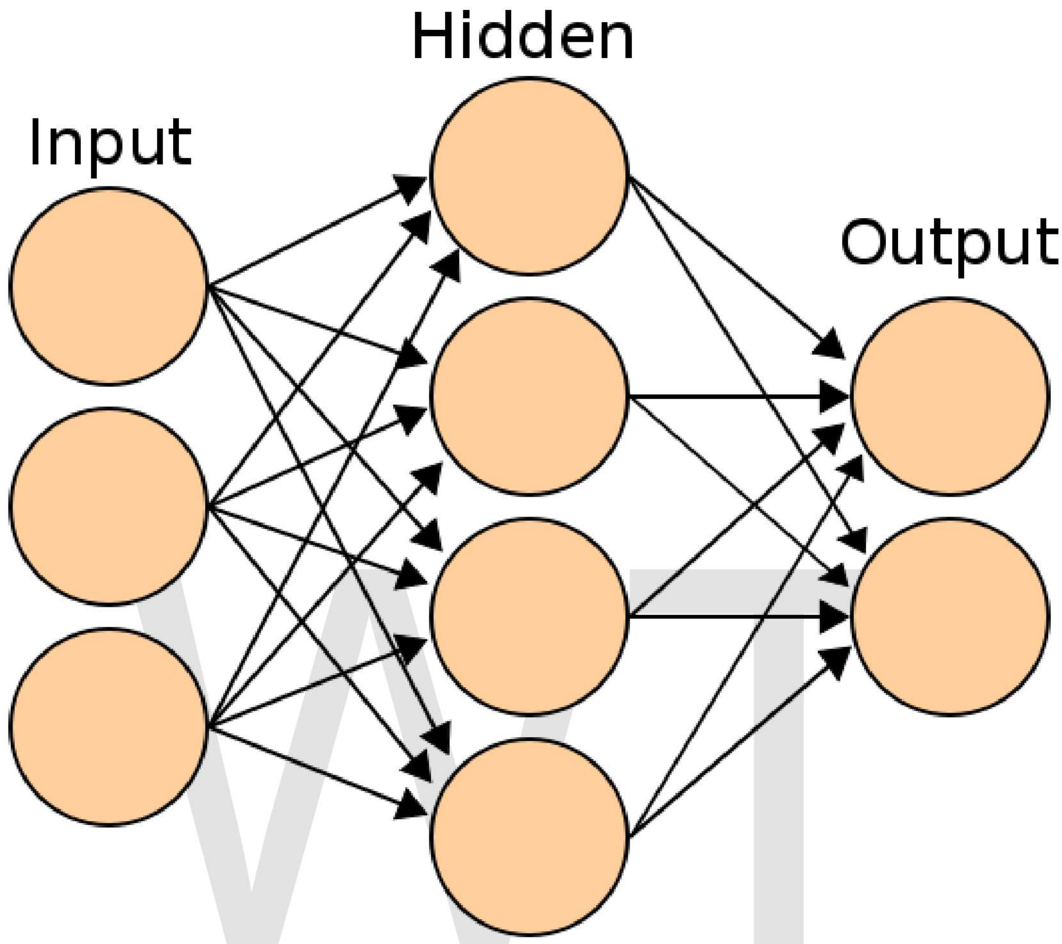
A key concept from the science of economics is "utility": a measure of how valuable something is to an intelligent agent. Precise mathematical tools have been developed that analyze how an agent can make choices and plan, using decision theory, decision analysis, information value theory. These tools include models such as Markov decision processes, dynamic decision networks, game theory and mechanism design.

Classifiers and statistical learning methods

The simplest AI applications can be divided into two types: classifiers ("if shiny then diamond") and controllers ("if shiny then pick up"). Controllers do however also classify conditions before inferring actions, and therefore classification forms a central part of many AI systems. Classifiers are functions that use pattern matching to determine a closest match. They can be tuned according to examples, making them very attractive for use in AI. These examples are known as observations or patterns. In supervised learning, each pattern belongs to a certain predefined class. A class can be seen as a decision that has to be made. All the observations combined with their class labels are known as a data set. When a new observation is received, that observation is classified based on previous experience.

A classifier can be trained in various ways; there are many statistical and machine learning approaches. The most widely used classifiers are the neural network, kernel methods such as the support vector machine, k-nearest neighbor algorithm, Gaussian mixture model, naive Bayes classifier, and decision tree. The performance of these classifiers have been compared over a wide range of tasks. Classifier performance depends greatly on the characteristics of the data to be classified. There is no single classifier that works best on all given problems; this is also referred to as the "no free lunch" theorem. Determining a suitable classifier for a given problem is still more an art than science.

Neural networks



A neural network is an interconnected group of nodes, akin to the vast network of neurons in the human brain.

The study of artificial neural networks began in the decade before the field AI research was founded, in the work of Walter Pitts and Warren McCullough. Other important early researchers were Frank Rosenblatt, who invented the perceptron and Paul Werbos who developed the backpropagation algorithm.

The main categories of networks are acyclic or feedforward neural networks (where the signal passes in only one direction) and recurrent neural networks (which allow feedback). Among the most popular feedforward networks are perceptrons, multi-layer perceptrons and radial basis networks. Among recurrent networks, the most famous is the Hopfield net, a form of attractor network, which was first described by John Hopfield in 1982. Neural networks can be applied to the problem of intelligent control (for robotics) or learning, using such techniques as Hebbian learning and competitive learning.

Jeff Hawkins argues that research in neural networks has stalled because it has failed to model the essential properties of the neocortex, and has suggested a model (Hierarchical Temporal Memory) that is loosely based on neurological research.

Control theory

Control theory, the grandchild of cybernetics, has many important applications, especially in robotics.

Languages

AI researchers have developed several specialized languages for AI research, including Lisp and Prolog.

Evaluating progress

In 1950, Alan Turing proposed a general procedure to test the intelligence of an agent now known as the Turing test. This procedure allows almost all the major problems of artificial intelligence to be tested. However, it is a very difficult challenge and at present all agents fail.

Artificial intelligence can also be evaluated on specific problems such as small problems in chemistry, hand-writing recognition and game-playing. Such tests have been termed subject matter expert Turing tests. Smaller problems provide more achievable goals and there are an ever-increasing number of positive results.

The broad classes of outcome for an AI test are: (1) Optimal: it is not possible to perform better. (2) Strong super-human: performs better than all humans. (3) Super-human: performs better than most humans. (4) Sub-human: performs worse than most humans. For example, performance at draughts is optimal, performance at chess is super-human and nearing strong super-human and performance at many everyday tasks (such as recognizing a face or crossing a room without bumping into something) is sub-human.

A quite different approach measures machine intelligence through tests which are developed from *mathematical* definitions of intelligence. Examples of these kinds of tests start in the late nineties devising intelligence tests using notions from Kolmogorov complexity and data compression. Two major advantages of mathematical definitions are their applicability to nonhuman intelligences and their absence of a requirement for human testers.

Applications

Artificial intelligence techniques are pervasive and are too numerous to list. Frequently, when a technique reaches mainstream use, it is no longer considered artificial intelligence; this phenomenon is described as the AI effect.

Competitions and prizes

There are a number of competitions and prizes to promote research in artificial intelligence. The main areas promoted are: general machine intelligence, conversational behavior, data-mining, driverless cars, robot soccer and games.

Platforms

A platform (or "computing platform")is defined as "some sort of hardware architecture or software framework (including application frameworks), that allows software to run." As Rodney Brooks pointed out many years ago, it is not just the artificial intelligence software that defines the AI features of the platform, but rather the actual platform itself that affects the AI that results, i.e., we need to be working out AI problems on real world platforms rather than in isolation.

A wide variety of platforms has allowed different aspects of AI to develop, ranging from expert systems, albeit PC-based but still an entire real-world system to various robot platforms such as the widely available Roomba with open interface.

Philosophy

Artificial intelligence, by claiming to be able to recreate the capabilities of the human mind, is both a challenge and an inspiration for philosophy. Are there limits to how intelligent machines can be? Is there an essential difference between human intelligence and artificial intelligence? Can a machine have a mind and consciousness? A few of the most influential answers to these questions are given below.

Turing's "polite convention"

> *If a machine acts as intelligently as a human being, then it is as intelligent as a human being.* Alan Turing theorized that, ultimately, we can only judge the intelligence of a machine based on its behavior. This theory forms the basis of the Turing test.

The Dartmouth proposal

> *"Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it."* This conjecture was printed in the proposal for the Dartmouth Conference of 1956, and represents the position of most working AI researchers.

Newell and Simon's physical symbol system hypothesis

> *"A physical symbol system has the necessary and sufficient means of general intelligent action."* Newell and Simon argue that intelligences consists of formal operations on symbols. Hubert Dreyfus argued that, on the contrary, human expertise depends on unconscious instinct rather than conscious symbol manipulation and on having a "feel" for the situation rather than explicit symbolic knowledge.

Gödel's incompleteness theorem

A formal system (such as a computer program) can not prove all true statements. Roger Penrose is among those who claim that Gödel's theorem limits what machines can do.

Searle's strong AI hypothesis

"The appropriately programmed computer with the right inputs and outputs would thereby have a mind in exactly the same sense human beings have minds." John Searle counters this assertion with his Chinese room argument, which asks us to look *inside* the computer and try to find where the "mind" might be.

The artificial brain argument

The brain can be simulated. Hans Moravec, Ray Kurzweil and others have argued that it is technologically feasible to copy the brain directly into hardware and software, and that such a simulation will be essentially identical to the original.

Prediction

Artificial Intelligence is a common topic in both science fiction and projections about the future of technology and society. The existence of an artificial intelligence that rivals human intelligence raises difficult ethical issues, and the potential power of the technology inspires both hopes and fears.

In fiction, Artificial Intelligence has appeared fulfilling many roles, including a servant (R2D2 in *Star Wars*), a law enforcer (K.I.T.T. "Knight Rider"), a comrade (Lt. Commander Data in *Star Trek: The Next Generation*), a conqueror/overlord (*The Matrix*), a dictator (*With Folded Hands*), an assassin (*Terminator*), a sentient race (*Battlestar Galactica*/Transformers), an extension to human abilities (*Ghost in the Shell*) and the savior of the human race (R. Daneel Olivaw in the *Asimov's Robot Series*).

Mary Shelley's *Frankenstein* considers a key issue in the ethics of artificial intelligence: if a machine can be created that has intelligence, could it also *feel*? If it can feel, does it have the same rights as a human? The idea also appears in modern science fiction, including the films *I Robot*, *Blade Runner* and *A.I.: Artificial Intelligence*, in which humanoid machines have the ability to feel human emotions. This issue, now known as "robot rights", is currently being considered by, for example, California's Institute for the Future, although many critics believe that the discussion is premature. The subject is profoundly discussed in the 2010 documentary film *Plug & Pray*.

Martin Ford and others argue that specialized artificial intelligence applications, robotics and other forms of automation will ultimately result in significant unemployment as machines begin to match and exceed the capability of workers to perform most routine and repetitive jobs. Ford predicts that many knowledge-based occupations—and in particular entry level jobs—will be increasingly susceptible to automation via expert systems and other AI-enhanced applications. AI-based applications may also be used to amplify the capabilities of low-wage offshore workers, making it more feasible to outsource knowledge work.

Joseph Weizenbaum wrote that AI applications can not, by definition, successfully simulate genuine human empathy and that the use of AI technology in fields such as customer service or psychotherapy was deeply misguided. Weizenbaum was also bothered that AI researchers (and some philosophers) were willing to view the human mind as nothing more than a computer program (a position now known as computationalism). To Weizenbaum these points suggest that AI research devalues human life.

Many futurists believe that artificial intelligence will ultimately transcend the limits of progress. Ray Kurzweil has used Moore's law (which describes the relentless exponential improvement in digital technology) to calculate that desktop computers will have the same processing power as human brains by the year 2029. He also predicts that by 2045 artificial intelligence will reach a point where it is able to improve *itself* at a rate that far exceeds anything conceivable in the past, a scenario that science fiction writer Vernor Vinge named the "singularity".

Robot designer Hans Moravec, cyberneticist Kevin Warwick and inventor Ray Kurzweil have predicted that humans and machines will merge in the future into cyborgs that are more capable and powerful than either. This idea, called transhumanism, which has roots in Aldous Huxley and Robert Ettinger, has been illustrated in fiction as well, for example in the manga *Ghost in the Shell* and the science-fiction series *Dune*.

Edward Fredkin argues that "artificial intelligence is the next stage in evolution," an idea first proposed by Samuel Butler's "Darwin among the Machines" (1863), and expanded upon by George Dyson in his book of the same name in 1998.

Pamela McCorduck writes that all these scenarios are expressions of the ancient human desire to, as she calls it, "forge the gods".

Chapter-5

3G

3G or **3rd generation mobile telecommunications**, is a generation of standards for mobile phones and mobile telecommunications services fulfilling the **International Mobile Telecommunications-2000 (IMT — 2000)** specifications by the International Telecommunication Union. Application services include wide-area wireless voice telephone, mobile Internet access, video calls and mobile TV, all in a mobile environment. To meet the IMT-2000 standards, a system is required to provide peak data rates of at least 200 kbit/s. Recent 3G releases, often denoted 3.5G and 3.75G, also provide mobile broadband access of several Mbit/s to smartphones and mobile modems in laptop computers.

The following standards are typically branded 3G:

- the UMTS system, first offered in 2001, standardized by 3GPP, used primarily in Europe, Japan, China (however with a different radio interface) and other regions predominated by GSM 2G system infrastructure. The cell phones are typically UMTS and GSM hybrids. Several radio interfaces are offered, sharing the same infrastructure:
    - The original and most widespread radio interface is called W-CDMA.
    - The TD-SCDMA radio interface, was commercialised in 2009 and is only offered in China.
    - The latest UMTS release, HSPA+, can provide peak data rates up to 56 Mbit/s in the downlink in theory (28 Mbit/s in existing services) and 22 Mbit/s in the uplink.

- the CDMA2000 system, first offered in 2002, standardized by 3GPP2, used especially in North America and South Korea, sharing infrastructure with the IS-95 2G standard. The cell phones are typically CDMA2000 and IS-95 hybrids. The latest release EVDO Rev B offers peak rates of 14.7 Mbit/s downstreams.

The above systems and radio interfaces are based on kindred spread spectrum radio transmission technology. While the GSM EDGE standard ("2.9G"), DECT cordless phones and Mobile WiMAX standards formally also fulfill the IMT-2000 requirements and are approved as 3G standards by ITU, these are typically not branded 3G, and are based on completely different technologies.

A new generation of cellular standards has appeared approximately every tenth year since 1G systems were introduced in 1981/1982. Each generation is characterized by new frequency bands, higher data rates and non backwards compatible transmission technology. The first release of the 3GPP Long Term Evolution (LTE) standard does not completely fulfill the ITU 4G requirements called IMT-Advanced. First release LTE is not backwards compatible with 3G, but is a pre-4G or 3.9G technology, however sometimes branded "4G" by the service providers. WiMAX is another technology verging on or marketed as 4G.

Overview

The following common standards comply with the IMT2000/3G standard:

- EDGE, a revision by the 3GPP organization to the older 2G GSM based transmission methods, utilizing the same switching nodes, basestation sites and frequencies as GPRS, but new basestation and cellphone RF circuits. It is based on the three times as efficient 8PSK modulation scheme as supplement to the original GMSK modulation scheme. EDGE is still used extensively due to its ease of upgrade from existing 2G GSM infrastructure and cell-phones.
    - EDGE combined with the GPRS 2.5G technology is called EGPRS, and allows peak data rates in the order of 200 kbit/s, just as the original UMTS WCDMA versions, and thus formally fullfills the IMT2000 requirements on 3G systems. However, in practice EDGE is seldom marketed as a 3G system, but a 2.9G system. EDGE shows slightly better system spectral efficiency than the original UMTS and CDMA2000 systems, but it is difficult to reach much higher peak data rates due to the limited GSM spectral bandwidth of 200 kHz, and it is thus a dead end.
    - EDGE was also a mode in the IS-135 TDMA system, today ceased.
    - Evolved EDGE, the latest revision, has peaks of 1 Mbits/s downstream and 400kbits/s upstream, but is not commercially used.
- The Universal Mobile Telecommunications System, created and revised by the 3GPP. The family is a full revision from GSM in terms of of encoding methods and hardware, although some GSM sites can be retrofitted to broadcast in the UMTS/W-CDMA format.
    - W-CDMA is the most common deployment, commonly operated on the 2100 MHz band. A few others use the 900 and 1850 MHz bands.
    - HSPA is a revision and upgrade to W-CDMA UMTS, used by AT&T Wireless, Telstra and Telecom NZ, typically broadcasting on the 850 MHz band. HSPA requires updates to the
        - HSPA+ a revision and upgrade of HSPA, can provide peak data rates up to 56 Mbit/s in the downlink in theory (28 Mbit/s in existing services) and 22 Mbit/s in the uplink. It utilises multiple

base stations to potentially double the channels available utilising MIMO principles.

- The CDMA2000 system, or IS-2000, standardized by 3GPP2 (*differing* from the 3GPP, updating the IS-95 CDMA system, used especially in North America and South Korea.
    - EVDO Rev. B is the latest update, offering downstream peak rates of 14.7 Mbit/s. It is used primarily by the US carrier Verizon.

While DECT cordless phones and Mobile WiMAX standards formally also fulfill the IMT-2000 requirements, they are not usually considered due to their rarity and unsuitability for usage with mobile phones.

Detailed breakdown of 3G systems

The 3G (UMTS and CDMA2000) research and development projects started in 1992. In 1999, ITU approved five radio interfaces for IMT-2000 as a part of the ITU-R M.1457 Recommendation; WiMAX was added in 2007.

There are **evolutionary standards** (EDGE and CDMA) that are backwards-compatible extensions to pre-existing 2G networks as well as **revolutionary standards** that require all-new network hardware and frequency allocations. The cell phones used utilise UMTS in combination with 2G GSM standards and bandwidths, but *do not support EDGE*. The latter group is the UMTS family, which consists of standards developed for IMT-2000, as well as the independently developed standards DECT and WiMAX, which were included because they fit the IMT-2000 definition.

Overview of 3G/IMT-2000 standards

ITU IMT-2000 compliant standards	common name(s)	bandwidth of data	pre-4G upgrade	duplex	channel	description	geographical areas
TDMA Single-Carrier (IMT-SC)	EDGE (UWC-136)	EDGE Evolution	*likely discontinued*		TDMA	evolutionary upgrade to GSM/GPRS	worldwide, except Japan and South Korea
CDMA Multi-Carrier (IMT-MC)	CDMA2000	EV-DO	UMB[nb 2]	FDD	CDMA	evolutionary upgrade to cdmaOne (IS-95)	Americas, Asia, some others
CDMA Direct Spread	UMTS[nb 3] W-CDMA[nb 4]	HSPA	LTE			family of	worldwide

(IMT-DS)						
CDMA TDD (IMT-TC)	TD-CDMA[nb 5]				revolutionary upgrades to earlier GSM family.	Europe
	TD-SCDMA[nb 6]					China
FDMA/TDMA (IMT-FT)	DECT	*none*	TDD	FDMA/TDMA	short-range; standard for cordless phones	Europe, USA
IP-OFDMA		WiMAX (IEEE 802.16)		OFDMA		worldwide

1. ^ Can also be used as an upgrade to PDC or D-AMPS.
2. ^ development halted in favour of LTE.
3. ^ also known as FOMA; UMTS is the common name for a standard that encompasses multiple air interfaces.
4. ^ also known as UTRA-FDD; W-CDMA is sometimes used as a synonym for UMTS, ignoring the other air interface options.
5. ^ also known as UTRA-TDD 3.84 Mcps high chip rate (HCR)
6. ^ also known as UTRA-TDD 1.28 Mcps low chip rate (LCR)

While EDGE fulfills the 3G specifications, most GSM/UMTS phones report EDGE ("2.75G") and UMTS ("3G") functionality.

History

The first pre-commercial 3G network was launched by NTT DoCoMo in Japan, branded as FOMA. It was first available in May 2001 as a pre-release (test) of W-CDMA technology. The first commercial launch of 3G was also by NTT DoCoMo in Japan on 1 October 2001, although it was initially somewhat limited in scope; broader availability of the system was delayed by apparent concerns over its reliability.

The second network to go commercially live was by SK Telecom in South Korea on the CDMA-based 1xEV-DO technology in January 2002. By May 2002 the second South Korean 3G network was by KT on EV-DO and thus the Koreans were the first to see competition among 3G operators.

The first European pre-commercial network was an UTMS network on the Isle of Man by Manx Telecom, the operator then owned by British Telecom, and the first commercial network (also UTMS based W-CDMA) in Europe was opened for business by Telenor in December 2001 with no commercial handsets and thus no paying customers.

The first commercial United States 3G network was by Monet Mobile Networks, on CDMA2000 1x EV-DO technology, but this network provider later shut down operations. The second 3G network operator in the USA was Verizon Wireless in October 2003 also on CDMA2000 1x EV-DO. AT&T Mobility is also a true 3G UMTS network, having completed its upgrade of the 3G network to HSUPA.

The first pre-commercial demonstration network in the southern hemisphere was built in Adelaide, South Australia by m.Net Corporation in February 2002 using UMTS on 2100 MHz. This was a demonstration network for the 2002 IT World Congress. The first commercial 3G network was launched by Hutchison Telecommunications branded as *Three* in March 2003.

Emtel Launched the first 3G network in Africa.

By June 2007, the 200 millionth 3G subscriber had been connected. Out of 3 billion mobile phone subscriptions worldwide this is only 6.7%. In the countries where 3G was launched first – Japan and South Korea – 3G penetration is over 70%. In Europe the leading country is Italy with a third of its subscribers migrated to 3G. Other leading countries by 3G migration include UK, Austria, Australia and Singapore at the 20% migration level. A confusing statistic is counting CDMA2000 1x RTT customers as if they were 3G customers. If using this definition, then the total 3G subscriber base would be 475 million at June 2007 and 15.8% of all subscribers worldwide.

Adoption

3G was relatively slow to be adopted globally. In some instances, 3G networks do not use the same radio frequencies as 2G so mobile operators must build entirely new networks and license entirely new frequencies, especially so to achieve high-end data transmission rates. Other delays were due to the expenses of upgrading transmission hardware, especially for UMTS, whose deployment required the replacement of most broadcast towers. Due to these issues and difficulties with deployment, many carriers were not able to or delayed acquisition of these updated capabilities.

In December 2007, 190 3G networks were operating in 40 countries and 154 HSDPA networks were operating in 71 countries, according to the Global Mobile Suppliers Association (GSA). In Asia, Europe, Canada and the USA, telecommunication companies use W-CDMA technology with the support of around 100 terminal designs to operate 3G mobile networks.

Roll-out of 3G networks was delayed in some countries by the enormous costs of additional spectrum licensing fees. The license fees in some European countries were particularly high, bolstered by government auctions of a limited number of licenses and sealed bid auctions, and initial excitement over 3G's potential.

The 3G standard is perhaps well known because of a massive expansion of the mobile communications market post-2G and advances of the consumer mophone. An especially notable development during this time is the smartphone (for example, the iPhone, and the Android family), combining the abilities of a PDA with a mobile phone, leading to widespread demand for mobile internet connectivity.3G has also introduced the term "mobile broadband" because its speed and capability makes it a viable alternative for internet browsing, and USB Modems connecting to 3G networks are becoming increasingly common.

Africa

The first African use of 3G technology was a 3G videocall made in Johannesburg on the Vodacom network in November 2004. The first commercial launch was by Emtel-ltd in Mauritius in 2004. In late March 2006, a 3G service was provided by the new company Wana in Morrocco.In East Africa (Tanzania) in 2007 a 3G service was provided by Vodacom Tanzania.

Asia

China

China announced in May 2008, that the telecoms sector was re-organized and three 3G networks would be allocated so that the largest mobile operator, China Mobile, would retain its GSM customer base. China Unicom would retain its GSM customer base but relinquish its CDMA2000 customer base, and launch 3G on the globally leading W-CDMA (UMTS) standard. The CDMA2000 customers of China Unicom would go to China Telecom, which would then launch 3G on the CDMA2000 1x EV-DO standard. This meant that China would have all three main cellular technology 3G standards in commercial use. Finally in January 2009, Ministry of industry and Information Technology of China awarded licenses of all three standards: TD-SCDMA to China Mobile, W-CDMA to China Unicom and CDMA2000 to China Telecom. The launch of 3G occurred on 1 October 2009, to coincide with the 60th Anniversary of the Founding of the People's Republic of China.

India

In 2008, India entered the 3G arena with the launch of 3G enabled Mobile and Data services by Government owned Bharat Sanchar Nigam Ltd. (BSNL). Later, MTNL launched 3G in Delhi and Mumbai. Nationwide auction of 3G wireless spectrum was announced in April 2010.

The first Private-sector service provider that launched 3G services is Tata DoCoMo, on November 5, 2010. And the second is by Reliance Communications, December 13, 2010. Vodafone Launched their 3G by mid of March. Then, Bharti Airtel launched their 3G services on 24 January 2011 in Bangalore and also launched in Delhi & Jaipur on March 4, 2011(not GSM but only USB estick). Aircel also launched 3G in Kolkatta in the month of February. Other providers like Airtel, Idea and are expected to launch 3G services by Q1 2011.

North Korea

North Korea has had a 3G network since 2008, which is called **Koryolink**, a joint venture between Egyptian company Orascom Telecom Holding and the state-owned Korea Post and Telecommunications Corporation (KPTC) is North Korea's only 3G Mobile operator, and one of only two mobile companies in the country. According to Orascom quoted in *BusinessWeek*, the company had 125,661 subscribers in May 2010. The Egyptian company owns 75 percent of Koryolink, and is known to invest in infrastructure for mobile technology in developing nations. It covers Pyongyang, and five additional cities and eight highways and railways. Its only competitor, SunNet, uses GSM technology and suffers from poor call quality and disconnections. Phone numbers on the network are prefixed with +850 (0)192.

Philippines

3G services were made available in the Philippines on December 2008.

Europe

In Europe, mass market commercial 3G services were introduced starting in March 2003 by 3 (Part of Hutchison Whampoa) in the UK and Italy. The European Union Council suggested that the 3G operators should cover 80% of the European national populations by the end of 2005.

North America

In Canada, Bell Mobility, SaskTel and Telus launched a 3G EVDO network in 2005. Rogers Wireless was the first to implement UMTS technology, with HSDPA services in eastern Canada in late 2006. Realizing they would miss out on roaming revenue from the 2010 Winter Olympics, Bell and Telus formed a joint venture and rolled out a shared HSDPA network using Nokia Siemens technology.

Middle East

Mobitel Iraq is the first mobile 3G operator in Iraq. It was launched commercially on February 2007.

MTN Syria is the first mobile 3G operator in Syria. It was launched commercially on May 2010.

Turkey

Turkcell, Avea and Vodafone launched their 3G networks commercially on 30 July 2009 at the same time. Turkcell and Vodafone launched their 3G service on all provincial centres. Avea launched it on 16 provincial centres. It was after Turkey's monopoly mobile operator Turkcell accepted number portability, mobile operators attended frequency band auction and frequencies for 3G usage distributed around mobile operators. Turkcell got A band, Vodafone B and Avea C. Currently Turkcell and Vodafone have 3G networks on most of crowded cities and towns.

Features

Data rates

ITU has not provided a clear definition of the data rate users can expect from 3G equipment or providers. Thus users sold 3G service may not be able to point to a standard and say that the rates it specifies are not being met. While stating in commentary that "it is expected that IMT-2000 will provide higher transmission rates: a minimum data rate of 2 Mbit/s for stationary or walking users, and 384 kbit/s in a moving vehicle," the ITU does not actually clearly specify minimum or average rates or what modes of the interfaces qualify as 3G, so various rates are sold as 3G intended to meet customers expectations of broadband data.

Security

3G networks offer greater security than their 2G predecessors. By allowing the UE (User Equipment) to authenticate the network it is attaching to, the user can be sure the network is the intended one and not an impersonator. 3G networks use the KASUMI block crypto instead of the older A5/1 stream cipher. However, a number of serious weaknesses in the KASUMI cipher have been identified.

In addition to the 3G network infrastructure security, end-to-end security is offered when application frameworks such as IMS are accessed, although this is not strictly a 3G property.

Applications

The bandwidth and location information available to 3G devices gives rise to applications not previously available to mobile phone users. Some of the applications are:

- **Mobile TV** – a provider redirects a TV channel directly to the subscriber's phone where it can be watched.
- **Video on demand** – a provider sends a movie to the subscriber's phone.
- **Video conferencing** – subscribers can see as well as talk to each other.
- **Tele-medicine** – a medical provider monitors or provides advice to the potentially isolated subscriber.

- **Location-based services** – a provider sends localized weather or traffic conditions to the phone, or the phone allows the subscriber to find nearby businesses or friends.

Evolution

Both 3GPP and 3GPP2 are currently working on extensions to 3G standard that are based on an all-IP network infrastructure and using advanced wireless technologies such as MIMO, these specifications already display features characteristic for IMT-Advanced (4G), the successor of 3G. However, falling short of the bandwidth requirements for 4G (which is 1 Gbit/s for stationary and 100 Mbit/s for mobile operation), these standards are classified as 3.9G or Pre-4G.

3GPP plans to meet the 4G goals with LTE Advanced, whereas Qualcomm has halted development of UMB in favour of the LTE family.

On 14 December 2009, Telia Sonera announced in an official press release that "We are very proud to be the first operator in the world to offer our customers 4G services." With the launch of their LTE network, initially they are offering *pre-4G* (or *beyond 3G*) services in Stockholm, Sweden and Oslo, Norway.

Chapter-6

4G

In telecommunications, **4G** is the fourth generation of cellular wireless standards. It is a successor to the 3G and 2G families of standards. In 2008, the ITU-R organization specified the **IMT-Advanced** (International Mobile Telecommunications Advanced) requirements for 4G standards, setting peak speed requirements for 4G service at 100 Mbit/s for high mobility communication (such as from trains and cars) and 1 Gbit/s for low mobility communication (such as pedestrians and stationary users).

A 4G system is expected to provide a comprehensive and secure all-IP based mobile broadband solution to laptop computer wireless modems, smartphones, and other mobile devices. Facilities such as ultra-broadband Internet access, IP telephony, gaming services, and streamed multimedia may be provided to users.

Pre-4G technologies such as mobile WiMAX and first-release 3G Long term evolution (LTE) have been on the market since 2006 and 2009 respectively, and are often branded as 4G. The current versions of these technologies did not fulfill the original ITU-R requirements of data rates approximately up to 1 Gbit/s for 4G systems. Marketing materials use 4G as a description for Mobile-WiMAX and LTE in their current forms.

IMT-Advanced compliant versions of the above two standards are under development and called “LTE Advanced” and “WirelessMAN-Advanced” respectively. ITU has decided that “LTE Advanced” and “WirelessMAN-Advanced” should be accorded the official designation of IMT-Advanced. On December 6, 2010, ITU announced that current versions of LTE, WiMax and other evolved 3G technologies that do not fulfill "IMT-Advanced" requirements could be considered "4G", provided they represent forerunners to IMT-Advanced and "a substantial level of improvement in performance and capabilities with respect to the initial third generation systems now deployed."

In all suggestions for 4G, the CDMA spread spectrum radio technology used in 3G systems and IS-95 is abandoned and replaced by OFDMA and other frequency-domain equalization schemes. This is combined with MIMO (Multiple In Multiple Out), e.g., multiple antennas, dynamic channel allocation and channel-dependent scheduling.

Background

The nomenclature of the generations generally refers to a change in the fundamental nature of the service, non-backwards compatible transmission technology, and new frequency bands. New generations have appeared about every ten years since the first move from 1981 analog (1G) to digital (2G) transmission in 1992. This was followed, in 2001, by 3G multi-media support, spread spectrum transmission and at least 200 kbit/s, in 2011 expected to be followed by 4G, which refers to all-IP packet-switched networks, mobile ultra-broadband (gigabit speed) access and multi-carrier transmission.

The fastest 3G based standard in the WCDMA family is the HSPA+ standard, which was commercially available in 2009 and offers 28 Mbit/s downstreams without MIMO, i.e. only with one antenna (it would offer 56 Mbit/s with 2x2 MIMO), and 22 Mbit/s upstreams. The fastest 3G based standard in the CDMA2000 family is the EV-DO Rev. B, which was available in 2010 and offers 15.67 Mbit/s downstreams.

In mid 1990s, the ITU-R organization specified the IMT-2000 specifications for what standards that should be considered 3G systems. However, the cell phone market only brands some of the IMT-2000 standards as 3G (e.g. WCDMA and CDMA2000), but not all (3GPP EDGE, DECT and mobile-WiMAX all fulfil the IMT-2000 requirements and are formally accepted as 3G standards, but are typically not branded as 3G). In 2008, ITU-R specified the **IMT-Advanced** (*International Mobile Telecommunications Advanced*) requirements for 4G systems.

ITU Requirements and 4G wireless standards

Here we, uses 4G to refer to **IMT-Advanced** (*International Mobile Telecommunications Advanced*), as defined by ITU-R. An IMT-Advanced cellular system must fulfil the following requirements:

- Based on an all-IP packet switched network.
- Peak data rates of up to approximately 100 Mbit/s for high mobility such as mobile access and up to approximately 1 Gbit/s for low mobility such as nomadic/local wireless access, according to the ITU requirements.
- Dynamically share and utilize the network resources to support more simultaneous users per cell.
- Scalable channel bandwidth, between 5 and 20 MHz, optionally up to 40 MHz.
- Peak link spectral efficiency of 15 bit/s/Hz in the downlink, and 6.75 bit/s/Hz in the uplink (meaning that 1 Gbit/s in the downlink should be possible over less than 67 MHz bandwidth).
- System spectral efficiency of up to 3 bit/s/Hz/cell in the downlink and 2.25 bit/s/Hz/cell for indoor usage.
- Smooth handovers across heterogeneous networks.
- Ability to offer high quality of service for next generation multimedia support.

In September 2009, the technology proposals were submitted to the International Telecommunication Union (ITU) as 4G candidates. Basically all proposals are based on two technologies:

- LTE Advanced standardized by the 3GPP
- 802.16m standardized by the IEEE (i.e. WiMAX)

Present implementations of WiMAX and LTE are largely considered a stopgap solution that will offer a considerable boost while WiMAX 2 (based on the 802.16m spec) and LTE Advanced are finalized. Both technologies aim to reach the objectives traced by the ITU, but are still far from being implemented.

The first set of 3GPP requirements on LTE Advanced was approved in June 2008. LTE Advanced will be standardized in 2010 as part of the Release 10 of the 3GPP specification. LTE Advanced will be fully built on the existing LTE specification Release 10 and not be defined as a new specification series. A summary of the technologies that have been studied as the basis for LTE Advanced is included in a technical report.

Current LTE and WiMAX implementations are considered pre-4G, as they don't fully comply with the planned requirements of 1 Gbit/s for stationary reception and 100 Mbit/s for mobile.

Confusion has been caused by some mobile carriers who have launched products advertised as 4G but which are actually current technologies, commonly referred to as '3.9G', which do not follow the ITU-R defined principles for 4G standards. A common argument for branding 3.9G systems as new-generation is that they use different frequency bands to 3G technologies; that they are based on a new radio-interface paradigm; and that the standards are not backwards compatible with 3G, whilst some of the standards are expected to be forwards compatible with "real" 4G technologies.

While the ITU has adopted recommendations for technologies that would be used for future global communications, they do not actually perform the standardization or development work themselves, instead relying on the work of other standards bodies such as IEEE, The WiMAX Forum and 3GPP. Recently, ITU-R Working Party 5D approved two industry-developed technologies (LTE Advanced and WirelessMAN-Advanced) for inclusion in the ITU's International Mobile Telecommunications Advanced (IMT-Advanced program), which is focused on global communication systems that would be available several years from now. This working party's objective was not to comment on today's 4G being rolled out in the United States and in fact, the Working Party itself purposely agreed not to tie their IMT-Advanced work to the term 4G, recognizing its common use in industry already; however, the ITU's PR department ignored that agreement and used term 4G anyway when issuing their press release.

The ITU's purpose is to foster the global use of communications. The ITU is relied upon by developing countries, for example, who want to be assured a technology is standardised and likely to be widely deployed. While the ITU has adopted recommendations for technologies that would be used for future global communications, they do not actually do the standardization or development work themselves, instead relying on the work of other standards bodies such as IEEE, The WiMAX Forum and 3GPP. While the ITU has developed recommendations on IMT-Advanced, those recommendations are not binding on ITU member countries.

4G Predecessors and candidate systems

The wireless telecommunications industry as a whole has early assumed the term 4G as a short hand way to describe those advanced cellular technologies that, among other things, are based on or employ wide channel OFDMA and SC-FDE technologies, MIMO transmission and an all-IP based architecture. Mobile-WiMAX, first release LTE, IEEE 802.20 as well as Flash-OFDM meets these early assumptions, and have been considered as 4G candidate systems, but do not yet meet the more recent ITU-R IMT-Advanced requirements.

4G candidate systems

LTE Advanced

LTE Advanced (Long-term-evolution Advanced) is a candidate for IMT-Advanced standard, formally submitted by the 3GPP organization to ITU-T in the fall 2009, and expected to be released in 2012. The target of 3GPP LTE Advanced is to reach and surpass the ITU requirements. LTE Advanced is essentially an enhancement to LTE. It is not a new technology but rather an improvement on the existing LTE network. This upgrade path makes it more cost effective for vendors to offer LTE and then upgrade to LTE Advanced which is similar to the upgrade from WCDMA to HSPA. LTE and LTE Advanced will also make use of additional spectrum and multiplexing to allow it to achieve higher data speeds. Coordinated Multi-point Transmission will also allow more system capacity to help handle the enhanced data speeds. Release 10 of LTE is expected to achieve the LTE Advanced speeds. Release 8 currently supports up to 300 Mbit/s download speeds which is still short of the IMT-Advanced standards.

Data speeds of LTE Advanced

	LTE Advanced
Peak Download	1 Gbit/s
Peak Upload	500 Mbit/s

IEEE 802.16m or WirelessMAN-Advanced

The IEEE 802.16m or WirelessMAN-Advanced evolution of 802.16e is under development, with the objective to fulfill the IMT-Advanced criteria of 1 Gbit/s for stationary reception and 100 Mbit/s for mobile reception.

4G predecessors and discontinued candidate systems

3GPP Long Term Evolution (LTE)

Telia-branded Samsung LTE modem

The pre-4G technology 3GPP Long Term Evolution (LTE) is often branded "4G", but the first LTE release does not fully comply with the IMT-Advanced requirements. LTE has a theoretical net bit rate capacity of up to 100 Mbit/s in the downlink and 50 Mbit/s in the uplink if a 20 MHz channel is used — and more if multiple-input multiple-output (MIMO), i.e. antenna arrays, are used.

The physical radio interface was at an early stage named *High Speed OFDM Packet Access* (HSOPA), now named Evolved UMTS Terrestrial Radio Access (E-UTRA). The first LTE USB dongles do not support any other radio interface.

The world's first publicly available LTE service was opened in the two Scandinavian capitals Stockholm (Ericsson system) and Oslo (a Huawei system) on 14 December 2009, and branded 4G. The user terminals were manufactured by Samsung. Currently, the two publicly available LTE services in the United States are provided by Metro PCS, and Verizon Wireless. AT&T also has an LTE service in planned for deployment between mid-2011 and end of 2013.

Mobile WiMAX (IEEE 802.16e)

The Mobile WiMAX (IEEE 802.16e-2005) mobile wireless broadband access (MWBA) standard (also known as WiBro in South Korea) is sometimes branded 4G, and offers peak data rates of 128 Mbit/s downlink and 56 Mbit/s uplink over 20 MHz wide channels.

The world's first commercial mobile WiMAX service was opened by KT in Seoul, South Korea on 30 June 2006.

Sprint Nextel has begun using Mobile WiMAX, as of September 29, 2008 branded as a "4G" network even though the current version does not fulfil the IMT Advanced requirements on 4G systems.

In Russia, Belarus and Nicaragua WiMax broadband internet access is offered by a Russian company Scartel, and is also branded 4G, Yota.

UMB (formerly EV-DO Rev. C)

UMB (Ultra Mobile Broadband) was the brand name for a discontinued 4G project within the 3GPP2 standardization group to improve the CDMA2000 mobile phone standard for next generation applications and requirements. In November 2008, Qualcomm, UMB's lead sponsor, announced it was ending development of the technology, favouring LTE instead. The objective was to achieve data speeds over 275 Mbit/s downstream and over 75 Mbit/s upstream.

Flash-OFDM

At an early stage the Flash-OFDM system was expected to be further developed into a 4G standard.

iBurst and MBWA (IEEE 802.20) systems

The iBurst system (or HC-SDMA, High Capacity Spatial Division Multiple Access) was at an early stage considered as a 4G predecessor. It was later further developed into the Mobile Broadband Wireless Access (MBWA) system, also known as IEEE 802.20.

Data rate comparison

The following table shows a comparison of 4G candidate systems as well as other competing technologies.

Comparison of Mobile Internet Access methods (This box: view · talk · edit)

Standard	Family	Primary Use	Radio Tech	Downlink (Mbit/s)	Uplink (Mbit/s)	Notes
WiMAX	802.16	Mobile Internet	MIMO-SOFDMA	128 (in 20MHz bandwidth)	56 (in 20MHz bandwidth)	WiMAX update IEEE 802.16m expected to offer peak rates of at least 1 Gbit/s fixed speeds and 100Mbit/s to mobile users.
LTE	UMTS/4GSM	General 4G	OFDMA/ MIMO/SC-FDMA	100 (in 20MHz bandwidth)	50 (in 20 MHz bandwidth)	LTE-Advanced update expected to offer peak rates up to 1 Gbit/s fixed speeds and 100 Mb/s to mobile users.
Flash-OFDM	Flash-OFDM	Mobile Internet mobility up to 200mph (350km/h)	Flash-OFDM	5.3 10.6 15.9	1.8 3.6 5.4	Mobile range 30km (18 miles) extended range 55 km (34 miles)
HIPERMAN	HIPERMAN	Mobile Internet	OFDM	56.9		
Wi-Fi	802.11 (11n)	Mobile Internet	OFDM/MIMO	300 (using 4x4 configuration in 20MHz bandwidth) or 600 (using 4x4 configuration in 40MHz bandwidth)		Antenna, RF front end enhancements and minor protocol timer tweaks have helped deploy long range P2P networks compromising on radial coverage, throughput and/or spectra efficiency (310km & 382km)
iBurst	802.20	Mobile Internet	HC-SDMA/ TDD/ MIMO	95	36	Cell Radius: 3–12 km Speed: 250km/h Spectral Efficiency: 13 bits/s/Hz/cell Spectrum Reuse

						Factor: "1"
EDGE Evolution	GSM	Mobile Internet	TDMA/FDD	1.6	0.5	3GPP Release 7
UMTS W-CDMA HSDPA+ HSUPA HSPA+	UMTS/3GSM	General 3G	CDMA/FDD CDMA/FDD/ MIMO	0.384 14.4 56	0.384 5.76 22	HSDPA widely deployed. Typical downlink rates today 2 Mbit/s, ~200 kbit/s uplink; HSPA+ downlink up to 56 Mbit/s.
UMTS-TDD	UMTS/3GSM	Mobile Internet	CDMA/TDD	16		Reported speeds according to IPWireless using 16QAM modulation similar to HSDPA+HSUPA
1xRTT	CDMA2000	Mobile phone	CDMA	0.144		Succeeded by EV-DO for data use, but still is used for voice and as a failover for EV-DO
EV-DO 1x Rev. 0 EV-DO 1x Rev.A EV-DO Rev.B	CDMA2000	Mobile Internet	CDMA/FDD	2.45 3.1 4.9xN	0.15 1.8 1.8xN	Rev B note: N is the number of 1.25 MHz chunks of spectrum used. EV-DO is not designed for voice, and requires a fallback to 1xRTT when a voice call is placed or received.

Notes: All speeds are theoretical maximums and will vary by a number of factors, including the use of external antennae, distance from the tower and the ground speed (e.g. communications on a train may be poorer than when standing still). Usually the bandwidth is shared between several terminals. The performance of each technology is determined by a number of constraints, including the spectral efficiency of the technology, the cell sizes used, and the amount of spectrum available. For more information**.**

Objective and approach

Objectives assumed in the literature

4G is being developed to accommodate the quality of service (QoS) and rate requirements set by further development of existing 3G applications like mobile broadband access, Multimedia Messaging Service (MMS), video chat, mobile TV, but also new services like HDTV. 4G may allow roaming with wireless local area networks, and may interact with digital video broadcasting systems.

In the literature, the assumed or expected 4G requirements have changed during the years before IMT-Advanced was specified by the ITU-R. These are examples of objectives stated in various sources:

- A nominal data rate of 100 Mbit/s while the client physically moves at high speeds relative to the station, and 1 Gbit/s while client and station are in relatively fixed positions as defined by the ITU-R
- A data rate of at least 100 Mbit/s between any two points in the world
- Smooth handoff across heterogeneous networks
- Seamless connectivity and global roaming across multiple networks
- High quality of service for next generation multimedia support (real time audio, high speed data, HDTV video content, mobile TV, etc.)
- Interoperability with existing wireless standards
- An all IP, packet switched network
- IP-based femtocells (home nodes connected to fixed Internet broadband infrastructure)

Approaches

Principal technologies

- Physical layer transmission techniques are as follows:
    - MIMO: To attain ultra high spectral efficiency by means of spatial processing including multi-antenna and multi-user MIMO
    - *Frequency-domain-equalization*, for example *Multi-carrier modulation (OFDM) in the downlink or* single-carrier frequency-domain-equalization *(SC-FDE) in the uplink: To exploit the frequency selective channel property without complex equalization.*
    - Frequency-domain statistical multiplexing, for example (OFDMA) or (Single-carrier FDMA) (SC-FDMA, a.k.a. Linearly precoded OFDMA, LP-OFDMA) in the uplink: Variable bit rate by assigning different sub-channels to different users based on the channel conditions
    - Turbo principle error-correcting codes: To minimize the required SNR at the reception side
- Channel-dependent scheduling: To utilize the time-varying channel.
- Link adaptation: Adaptive modulation and error-correcting codes
- Relaying, including fixed relay networks (FRNs), and the cooperative relaying concept, known as multi-mode protocol

4G features assumed in early literature

The 4G system was originally envisioned by the Defense Advanced Research Projects Agency (DARPA). The DARPA selected the distributed architecture, end-to-end Internet protocol (IP), and believed at an early stage in peer-to-peer networking in which every mobile device would be both a transceiver and a router for other devices in the network eliminating the spoke-and-hub weakness of 2G and 3G cellular systems. Since the 2.5G

GPRS system, cellular systems have provided dual infrastructures: packet switched nodes for data services, and circuit switched nodes for voice calls. In 4G systems, the circuit-switched infrastructure is abandoned, and only a packet-switched network is provided, while 2.5G and 3G systems require both packet-switched and circuit-switched network nodes, i.e. two infrastructures in parallel. This means that in 4G, traditional voice calls are replaced by IP telephony.

Cellular systems such as 4G allow seamless mobility; thus a file transfer is not interrupted in case a terminal moves from one cell (one base station coverage area) to another, but handover is carried out. The terminal also keeps the same IP address while moving, meaning that a mobile server is reachable as long as it is within the coverage area of any server. In 4G systems this mobility is provided by the mobile IP protocol, part of IP version 6, while in earlier cellular generations it was only provided by physical layer and datalink layer protocols. In addition to seamless mobility, 4G provides flexible interoperability of the various kinds of existing wireless networks, such as satellite, cellular wirelss, WLAN, PAN and systems for accessing fixed wireless networks.

While maintaining seamless mobility, 4G will offer very high data rates with expectations of 100 Mbit/s wireless service. The increased bandwidth and higher data transmission rates will allow 4G users the ability to utilize high definition video and the video conferencing features of mobile devices attached to a 4G network. The 4G wireless system is expected to provide a comprehensive IP solution where multimedia applications and services can be delivered to the user on an 'Anytime, Anywhere' basis with a satisfactory high data rate, premium quality and high security.

4G is described as MAGIC: mobile multimedia, any-time anywhere, global mobility support, integrated wireless solution, and customized personal service. Some key features (primarily from users' points of view) of 4G mobile networks are:

- High usability: anytime, anywhere, and with any technology
- Support for multimedia services at low transmission cost
- Personalization
- Integrated services

Components

Access schemes

As the wireless standards evolved, the access techniques used also exhibited increase in efficiency, capacity and scalability. The first generation wireless standards used plain TDMA and FDMA. In the wireless channels, TDMA proved to be less efficient in handling the high data rate channels as it requires large guard periods to alleviate the multipath impact. Similarly, FDMA consumed more bandwidth for guard to avoid inter carrier interference. So in second generation systems, one set of standard used the combination of FDMA and TDMA and the other set introduced an access scheme called CDMA. Usage of CDMA increased the system capacity, but as a theoretical drawback placed a soft limit on it rather than the hard limit (i.e. a CDMA network setup does not inherently reject new clients when it approaches its limits, resulting in a denial of service to all clients when the network overloads; though this outcome is avoided in practical implementations by admission control of circuit switched or fixed bitrate communication services). Data rate is also increased as this access scheme (providing the network is not reaching its capacity) is efficient enough to handle the multipath channel. This enabled the third generation systems, such as IS-2000, UMTS, HSXPA, 1xEV-DO, TD-CDMA and TD-SCDMA, to use CDMA as the access scheme. However, the issue with CDMA is that it suffers from poor spectral flexibility and computationally intensive time-domain equalization (high number of multiplications per second) for wideband channels.

Recently, new access schemes like Orthogonal FDMA (OFDMA), Single Carrier FDMA (SC-FDMA), Interleaved FDMA and Multi-carrier CDMA (MC-CDMA) are gaining more importance for the next generation systems. These are based on efficient FFT algorithms and frequency domain equalization, resulting in a lower number of multiplications per second. They also make it possible to control the bandwidth and form the spectrum in a flexible way. However, they require advanced dynamic channel allocation and traffic adaptive scheduling.

WiMax is using OFDMA in the downlink and in the uplink. For the next generation UMTS, OFDMA is used for the downlink. By contrast, IFDMA is being considered for the uplink since OFDMA contributes more to the PAPR related issues and results in nonlinear operation of amplifiers. IFDMA provides less power fluctuation and thus avoids amplifier issues. Similarly, MC-CDMA is in the proposal for the IEEE 802.20 standard. These access schemes offer the same efficiencies as older technologies like CDMA. Apart from this, scalability and higher data rates can be achieved.

The other important advantage of the above mentioned access techniques is that they require less complexity for equalization at the receiver. This is an added advantage especially in the MIMO environments since the spatial multiplexing transmission of MIMO systems inherently requires high complexity equalization at the receiver.

In addition to improvements in these multiplexing systems, improved modulation techniques are being used. Whereas earlier standards largely used Phase-shift keying, more efficient systems such as 64QAM are being proposed for use with the 3GPP Long Term Evolution standards.

IPv6 support

Unlike 3G, which is based on two parallel infrastructures consisting of circuit switched and packet switched network nodes respectively, 4G will be based on packet switching *only*. This will require low-latency data transmission.

By the time that 4G was deployed, the process of IPv4 address exhaustion was expected to be in its final stages. Therefore, in the context of 4G, IPv6 support is essential in order to support a large number of wireless-enabled devices. By increasing the number of IP addresses, IPv6 removes the need for network address translation (NAT), a method of sharing a limited number of addresses among a larger group of devices, although NAT will still be required to communicate with devices that are on existing IPv4 networks.

As of June 2009, Verizon has posted specifications that require any 4G devices on its network to support IPv6.

Advanced antenna systems

The performance of radio communications depends on an antenna system, termed smart or intelligent antenna. Recently, multiple antenna technologies are emerging to achieve the goal of 4G systems such as high rate, high reliability, and long range communications. In the early 1990s, to cater for the growing data rate needs of data communication, many transmission schemes were proposed. One technology, spatial multiplexing, gained importance for its bandwidth conservation and power efficiency. Spatial multiplexing involves deploying multiple antennas at the transmitter and at the receiver. Independent streams can then be transmitted simultaneously from all the antennas. This technology, called MIMO (as a branch of intelligent antenna), multiplies the base data rate by (the smaller of) the number of transmit antennas or the number of receive antennas. Apart from this, the reliability in transmitting high speed data in the fading channel can be improved by using more antennas at the transmitter or at the receiver. This is called *transmit* or *receive diversity*. Both transmit/receive diversity and transmit spatial multiplexing are categorized into the space-time coding techniques, which does not necessarily require the channel knowledge at the transmitter. The other category is closed-loop multiple antenna technologies, which require channel knowledge at the transmitter.

Software-defined radio (SDR)

SDR is one form of open wireless architecture (OWA). Since 4G is a collection of wireless standards, the final form of a 4G device will constitute various standards. This can be efficiently realized using SDR technology, which is categorized to the area of the radio convergence.

History of 4G and pre-4G technologies

- In 2002, the strategic vision for 4G—which ITU designated as IMT-Advanced—was laid out.
- In 2005, OFDMA transmission technology is chosen as candidate for the HSOPA downlink, later renamed 3GPP Long Term Evolution (LTE) air interface E-UTRA.
- In November 2005, KT demonstrated mobile WiMAX service in Busan, South Korea.
- In June 2006, KT started the world's first commercial mobile WiMAX service in Seoul, South Korea.

- In mid-2006, Sprint Nextel announced that it would invest about US$5 billion in a WiMAX technology buildout over the next few years ($5.45 billion in real terms). Since that time Sprint has faced many setbacks, that have resulted in steep quarterly losses. On May 7, 2008, Sprint, Imagine, Google, Intel, Comcast, Bright House, and Time Warner announced a pooling of an average of 120 MHz of spectrum; Sprint merged its Xohm WiMAX division with Clearwire to form a company which will take the name "Clear".
- In February 2007, the Japanese company NTT DoCoMo tested a 4G communication system prototype with 4x4 MIMO called VSF-OFCDM at 100 Mbit/s while moving, and 1 Gbit/s while stationary. NTT DoCoMo completed a trial in which they reached a maximum packet transmission rate of approximately 5 Gbit/s in the downlink with 12x12 MIMO using a 100 MHz frequency bandwidth while moving at 10 km/h, and is planning on releasing the first commercial network in 2010.
- In September 2007, NTT Docomo demonstrated e-UTRA data rates of 200 Mbit/s with power consumption below 100 mW during the test.
- In January 2008, a U.S. Federal Communications Commission (FCC) spectrum auction for the 700 MHz former analog TV frequencies began. As a result, the biggest share of the spectrum went to Verizon Wireless and the next biggest to AT&T. Both of these companies have stated their intention of supporting LTE.
- In January 2008, EU commissioner Viviane Reding suggested re-allocation of 500–800 MHz spectrum for wireless communication, including WiMAX.
- On 15 February 2008 - Skyworks Solutions released a front-end module for e-UTRAN.
- In 2008, ITU-R established the detailed performance requirements of IMT-Advanced, by issuing a Circular Letter calling for candidate Radio Access Technologies (RATs) for IMT-Advanced.
- In April 2008, just after receiving the circular letter, the 3GPP organized a workshop on IMT-Advanced where it was decided that LTE Advanced, an evolution of current LTE standard, will meet or even exceed IMT-Advanced requirements following the ITU-R agenda.
- In April 2008, LG and Nortel demonstrated e-UTRA data rates of 50 Mbit/s while travelling at 110 km/h.
- On 12 November 2008, HTC announced the first WiMAX-enabled mobile phone, the Max 4G
- In December 2008, San Miguel Corporation, Asia's largest food and beverage conglomerate, has signed a memorandum of understanding with Qatar Telecom QSC (Qtel) to build wireless broadband and mobile communications projects in the Philippines. The joint-venture formed wi-tribe Philippines, which offers 4G in the country. Around the same time Globe Telecom rolled out the first WiMAX service in the Philippines.
- On 3 March 2009, Lithuania's LRTC announcing the first operational "4G" mobile WiMAX network in Baltic states.
- In December 2009, Sprint began advertising "4G" service in selected cities in the United States, despite average download speeds of only 3–6 Mbit/s with peak speeds of 10 Mbit/s (not available in all markets).

- On 14 December 2009, the first commercial LTE deployment was in the Scandinavian capitals Stockholm and Oslo by the Swedish-Finnish network operator TeliaSonera and its Norwegian brandname NetCom (Norway). TeliaSonera branded the network "4G". The modem devices on offer were manufactured by Samsung (dongle GT-B3710), and the network infrastructure created by Huawei (in Oslo) and Ericsson (in Stockholm). TeliaSonera plans to roll out nationwide LTE across Sweden, Norway and Finland. TeliaSonera used spectral bandwidth of 10 MHz, and single-in-single-out, which should provide physical layer net bitrates of up to 50 Mbit/s downlink and 25 Mbit/s in the uplink. Introductory tests showed a TCP throughput of 42.8 Mbit/s downlink and 5.3 Mbit/s uplink in Stockholm.
- On 25 February 2010, Estonia's EMT opened LTE "4G" network working in test regime.
- On 4 June 2010, Sprint Nextel released the first WiMAX smartphone in the US, the HTC Evo 4G.
- In July 2010, Uzbekistan's MTS deployed LTE in Tashkent.
- On 25 August 2010, Latvia's LMT opened LTE "4G" network working in test regime 50% of territory.
- On 6 December 2010, at the ITU World Radiocommunication Seminar 2010, the ITU stated that LTE, WiMax and similar "evolved 3G technologies" could be considered "4G".
- On 12 December 2010, VivaCell-MTS launches in Armenia 4G/LTE commercial test network with a live demo conducted in Yerevan.

Deployment plans

In May 2005, Digiweb, an Irish fixed and wireless broadband company, announced that they had received a mobile communications license from the Irish Telecoms regulator, ComReg. This service will be issued the mobile code *088* in Ireland and will be used for the provision of 4G Mobile communications. Digiweb launched a mobile broadband network using FLASH-OFDM technology at 872 MHz.

On September 20, 2007, Verizon Wireless announced plans for a joint effort with the Vodafone Group to transition its networks to the 4G standard LTE. On December 9, 2008, Verizon Wireless announced their intentions to build and begin to roll out an LTE network by the end of 2009. Since then, Verizon Wireless has said that they will start their rollout by the end of 2010.

On July 7, 2008, South Korea announced plans to spend 60 billion won, or US$58,000,000, on developing 4G and even 5G technologies, with the goal of having the highest mobile phone market share by 2012, and the hope of an international standard.

Telus and Bell Canada, the major Canadian cdmaOne and EV-DO carriers, have announced that they will be cooperating towards building a fourth generation (4G) LTE wireless broadband network in Canada. As a transitional measure, they are implementing 3G UMTS that went live in November 2009.

Sprint offers a 3G/4G connection plan, currently available in select cities in the United States. It delivers rates up to 10 Mbit/s.

In the United Kingdom, Telefónica O_2 is to use Slough as a guinea pig in testing the 4G network and has called upon Huawei to install LTE technology in six masts across the town to allow people to talk to each other via HD video conferencing and play PlayStation games while on the move.

Verizon Wireless has announced that it plans to augment its CDMA2000-based EV-DO 3G network in the United States with LTE. AT&T, along with Verizon Wireless, has chosen to migrate toward LTE from 2G/GSM and 3G/HSPA by 2011.

Sprint Nextel has deployed WiMAX technology which it has labeled 4G as of October 2008. It is currently deploying to additional markets and is the first US carrier to offer a WiMAX phone.

The U.S. FCC is exploring the possibility of deployment and operation of a nationwide 4G public safety network which would allow first responders to seamlessly communicate between agencies and across geographies, regardless of devices. In June 2010 the FCC released a comprehensive white paper which indicates that the 10 MHz of dedicated spectrum currently allocated from the 700 MHz spectrum for public safety will provide adequate capacity and performance necessary for normal communications as well as serious emergency situations.

TeliaSonera started deploying LTE (branded "4G") in Stockholm and Oslo November 2009 (as seen above), and in several Swedish, Norwegian, and Finnish cities during 2010. In June 2010, Swedish television companies used 4G to broadcast live television from the Swedish Crown Princess' Royal Wedding.

Safaricom, a telecommunication company in East& Central Africa, began its setup of a 4G network in October 2010 after the now retired& Kenya Tourist Board Chairman, Michael Joseph, regarded their 3G network as a white elephant i.e. it failed to perform to expectations. Huawei was given the contract the network is set to go fully commercial by the end of Q1 of 2011

Telstra announced on 15 February 2011, that it intents to upgrade its current Next G network to 4G with Long Term Evolution (LTE) technology in the central business districts of all Australian capital cities and selected regional centres by the end of 2011.

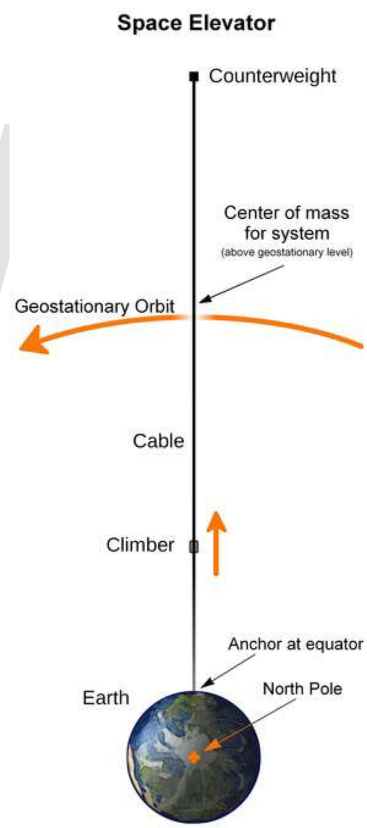
Beyond 4G research

A major issue in 4G systems is to make the high bit rates available in a larger portion of the cell, especially to users in an exposed position in between several base stations. In current research, this issue is addressed by macro-diversity techniques, also known as group cooperative relay, and also by beam-division multiple access.

Pervasive networks are an amorphous and at present entirely hypothetical concept where the user can be simultaneously connected to several wireless access technologies and can seamlessly move between them. These access technologies can be Wi-Fi, UMTS, EDGE, or any other future access technology. Included in this concept is also smart-radio (also known as cognitive radio technology) to efficiently manage spectrum use and transmission power as well as the use of mesh routing protocols to create a pervasive network.

Chapter-7

Space Elevator



A space elevator for Earth would consist of a cable anchored to the Earth's equator, reaching into space. By attaching a counterweight at the end (or by further extending the cable upward for the same purpose), the center of mass is kept above the level of geostationary orbit. Inertia ensures that the cable remains stretched taut, countering the gravitational pull downward. Once above the geostationary level, climbers would have weight in the *upward* direction as the centrifugal force overpowers gravity. (The diagram is to scale with regard to the size of the Earth and the height of the geosynchronous level, and roughly to scale for the height of the counterweight.)

A **space elevator** is a proposed non-rocket spacelaunch structure (a structure designed to transport material from a celestial body's surface into space). Many elevator variants have been suggested, all of which involve travelling along a fixed structure instead of using rocket powered space launch, most often a cable that reaches from the surface of the Earth on or near the Equator to geostationary orbit (GSO) and a counterweight outside of the geostationary orbit.

Discussion of a space elevator dates back to 1895 when Konstantin Tsiolkovsky proposed a free-standing "Tsiolkovsky" tower reaching from the surface of Earth to geostationary orbit 22,236 miles up. Like all buildings, Tsiolkovsky's structure would be under compression, supporting its weight from below. Since 1959, most ideas for space elevators have focused on purely tensile structures, with the weight of the system held up from above. In the tensile concepts, a space tether reaches from a large mass (the counterweight) beyond geostationary orbit to the ground. This structure is held in tension between Earth and the counterweight like a guitar string held taut. Space elevators have also sometimes been referred to as *beanstalks*, *space bridges*, *space lifts*, *space ladders*, *skyhooks*, *orbital towers*, or *orbital elevators*.

While some variants of the space elevator concept are technologically feasible, current technology is not capable of manufacturing practical engineering materials that are sufficiently strong and light to build an Earth-based space elevator of the geostationary orbital tether type. Recent concepts for a space elevator are notable for their plans to use carbon nanotube or boron nitride nanotube based materials as the tensile element in the tether design, since the measured strength of carbon nanotubes appears great enough to make this possible. Technology as of 1978 could produce elevators for locations in the solar system with weaker gravitational fields, such as the Moon or Mars.

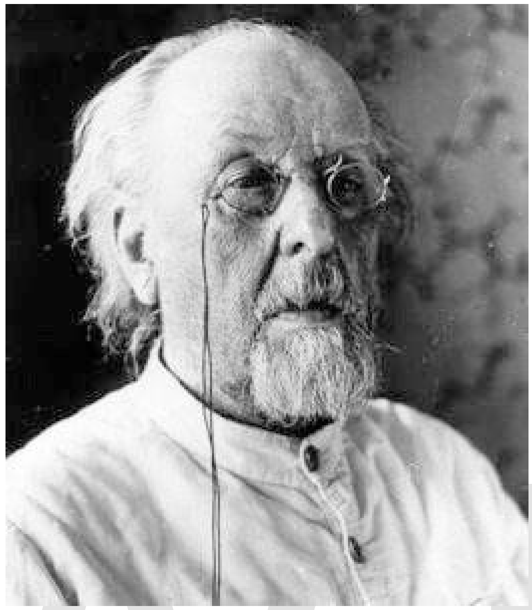
For human riders on an Earth-based elevator, adequate protection against radiation would likely need to be provided, depending on the transit time through the Van Allen belts. At the transit times expected for early systems, radiation due to the Van Allen belts would, if unshielded, give a dose well above permitted levels. This would not be an issue for non-living cargo, however.

Geostationary orbital tethers

This concept, also called an **orbital space elevator**, **geostationary orbital tether**, or a **beanstalk**, is a subset of the skyhook concept, and is what people normally think of when the phrase 'space elevator' is used (although there are variants).

Construction would be a large project: the minimum length of an Earth-based space elevator is well over 38,000 km (24,000 mi) long. The tether would have to be built of a material that could endure tremendous stress while also being light-weight, cost-effective, and manufacturable in great quantities. Materials currently available do not meet these requirements, although carbon nanotube technology shows great promise. As with all leading-edge engineering projects, other novel engineering problems would also have to be solved to make a space elevator practical, and there are problems regarding feasibility that have yet to be addressed.

History



Konstantin Tsiolkovsky

Early concepts

The key concept of the space elevator appeared in 1895 when Russian scientist Konstantin Tsiolkovsky was inspired by the Eiffel Tower in Paris to consider a tower that reached all the way into space, built from the ground up to an altitude of 35,790 kilometers (22,238 mi) above sea level (geostationary orbit). He noted that a "celestial castle" at the top of such a spindle-shaped cable would have the "castle" orbiting Earth in a geostationary orbit (i.e. the castle would remain over the same spot on Earth's surface).

Since the elevator would attain orbital velocity as it rode up the cable, an object released at the tower's top would also have the orbital velocity necessary to remain in geostationary orbit. Unlike more recent concepts for space elevators, Tsiolkovsky's (conceptual) tower was a compression structure, rather than a tension (or "tether") structure.

Twentieth century

Building a compression structure from the ground up proved an unrealistic task as there was no material in existence with enough compressive strength to support its own weight under such conditions. In 1959 another Russian scientist, Yuri N. Artsutanov, suggested a more feasible proposal. Artsutanov suggested using a geostationary satellite as the base from which to deploy the structure downward. By using a counterweight, a cable would be lowered from geostationary orbit to the surface of Earth, while the counterweight was extended from the satellite away from Earth, keeping the center of gravity of the cable motionless relative to Earth. Artsutanov's idea was introduced to the Russian-speaking public in an interview published in the Sunday supplement of *Komsomolskaya Pravda* in 1960, but was not available in English until much later. He also proposed tapering the cable thickness so that the stress in the cable was constant—this gives a thin cable at ground level, thickening up towards GSO.

Both the tower and cable ideas were proposed in the quasi-humorous *Ariadne* column in *New Scientist*, 24 December 1964.

Making a cable over 35,000 kilometers (22,000 miles) long is a difficult task. In 1966, Isaacs, Vine, Bradner and Bachus, four American engineers, reinvented the concept, naming it a "Sky-Hook," and published their analysis in the journal *Science*. They decided to determine what type of material would be required to build a space elevator, assuming it would be a straight cable with no variations in its cross section, and found that the strength required would be twice that of any existing material including graphite, quartz, and diamond.

In 1975 an American scientist, Jerome Pearson, reinvented the concept yet again, publishing his analysis in the journal Acta Astronautica. He designed a tapered cross section that would be better suited to building the elevator. The completed cable would be thickest at the geostationary orbit, where the tension was greatest, and would be narrowest at the tips to reduce the amount of weight per unit area of cross section that any point on the cable would have to bear. He suggested using a counterweight that would be slowly extended out to 144,000 kilometers (90,000 miles, almost half the distance to the Moon) as the lower section of the elevator was built. Without a large counterweight, the upper portion of the cable would have to be longer than the lower due to the way gravitational and centrifugal forces change with distance from Earth. His analysis included disturbances such as the gravitation of the Moon, wind and moving payloads up and down the cable. The weight of the material needed to build the elevator would have required thousands of Space Shuttle trips, although part of the material could be transported up the elevator when a minimum strength strand reached the ground or be manufactured in space from asteroidal or lunar ore.

In 1977, Hans Moravec published an article called "A Non-Synchronous Orbital Skyhook", in which he proposed an alternative space elevator concept, using a rotating cable, in which the rotation speed exactly matches the orbital speed in such a way that the instantaneous velocity at the point where the cable was at the closest point to the Earth was zero. This concept is an early version of a space tether transportation system.

In 1979, space elevators were introduced to a broader audience with the simultaneous publication of Arthur C. Clarke's novel, *The Fountains of Paradise*, in which engineers construct a space elevator on top of a mountain peak in the fictional island country of *Taprobane* (loosely based on Sri Lanka, albeit moved south to the Equator), and Charles Sheffield's first novel, *The Web Between the Worlds*, also featuring the building of a space elevator. Three years later, in Robert A. Heinlein's 1982 novel *Friday* the principal character makes use of the "Nairobi Beanstalk" in the course of her travels. In Kim Stanley Robinson's 1993 novel *Red Mars*, colonists build a space elevator on Mars that allows both for more colonists to arrive and also for natural resources mined there to be able to leave for Earth.

21st century

After the development of carbon nanotubes in the 1990s, engineer David Smitherman of NASA/Marshall's Advanced Projects Office realized that the high strength of these materials might make the concept of an orbital skyhook feasible, and put together a workshop at the Marshall Space Flight Center, inviting many scientists and engineers to discuss concepts and compile plans for an elevator to turn the concept into a reality. The publication he edited, compiling information from the workshop, "Space Elevators: An Advanced Earth-Space Infrastructure for the New Millennium", provides an introduction to the state of the technology at the time, and summarizes the findings.

Another American scientist, Bradley C. Edwards, suggested creating a 100,000 km (62,000 mi) long paper-thin ribbon using a carbon nanotube composite material. He chose a ribbon type structure rather than a cable because that structure might stand a greater chance of surviving impacts by meteoroids. Supported by the NASA Institute for Advanced Concepts, Edwards' work was expanded to cover the deployment scenario, climber design, power delivery system, orbital debris avoidance, anchor system, surviving atomic oxygen, avoiding lightning and hurricanes by locating the anchor in the western equatorial Pacific, construction costs, construction schedule, and environmental hazards. The largest holdup to Edwards' proposed design is the technological limit of the tether material. His calculations call for a fiber composed of epoxy-bonded carbon nanotubes with a minimal tensile strength of 130 GPa (19 million psi) (including a safety factor of 2); however, tests in 2000 of individual single-walled carbon nanotubes (SWCNTs), which should be notably stronger than an epoxy-bonded rope, indicated the strongest measured as 52 GPa (7.5 million psi). Multi-walled carbon nanotubes have been measured with tensile strengths up to 63 GPa (9 million psi).

To speed space elevator development, proponents are planning several competitions, similar to the Ansari X Prize, for relevant technologies. Among them are Elevator:2010, which will organize annual competitions for climbers, ribbons and power-beaming systems, the Robolympics Space Elevator Ribbon Climbing competition, as well as NASA's Centennial Challenges program, which, in March 2005, announced a partnership with the Spaceward Foundation (the operator of Elevator:2010), raising the total value of prizes to US$400,000.

In 2005, "the LiftPort Group of space elevator companies announced that it will be building a carbon nanotube manufacturing plant in Millville, New Jersey, to supply various glass, plastic and metal companies with these strong materials. Although LiftPort hopes to eventually use carbon nanotubes in the construction of a 100,000 km (62,000 mile) space elevator, this move will allow it to make money in the short term and conduct research and development into new production methods. The goal was a space elevator launch in 2010." On February 13, 2006 the LiftPort Group announced that, earlier the same month, they had tested a mile of "space-elevator tether" made of carbon-fiber composite strings and fiberglass tape measuring 5 cm (2 in) wide and 1 mm (approx. 6 sheets of paper) thick, lifted with balloons.

In 2007, Elevator:2010 held the 2007 Space Elevator games, which featured US$500,000 awards for each of the two competitions, (US$1,000,000 total) as well as an additional US$4,000,000 to be awarded over the next five years for space elevator related technologies. No teams won the competition, but a team from MIT entered the first 2-gram (0.07 oz), 100% carbon nanotube entry into the competition. Japan held an international conference in November 2008 to draw up a timetable for building the elevator.

In 2008 the book "Leaving the Planet by Space Elevator", by Dr. Brad Edwards and Philip Ragan, was published in Japanese and entered the Japanese best seller list. This has led to a Japanese announcement of intent to build a Space Elevator at a projected price tag of £5 billion. In a report by Leo Lewis, Tokyo correspondent of The Times newspaper in England, plans by Shuichi Ono, chairman of the Japan Space Elevator Association, are unveiled. Lewis says: "Japan is increasingly confident that its sprawling academic and industrial base can solve those [construction] issues, and has even put the astonishingly low price tag of a trillion yen (£5 billion/ $8 billion) on building the elevator. Japan is renowned as a global leader in the precision engineering and high-quality material production without which the idea could never be possible."

Physics of space elevators

Apparent gravitational field

The space elevator cable rotates along with the rotation of the Earth. Objects fastened to the cable will experience upward centrifugal force that opposes some, all, or more than the downward gravitational force at that point. Along the length of the cable, the *actual* (downward) gravity minus the (upward) centrifugal force is called the *apparent* gravitational field.

The apparent gravitational field can be computed this way:

$g = -G \cdot M/r^2 + \omega^2 \cdot r$, where
g is the acceleration along the radius (m s^{-2}),
G is the gravitational constant (m^3 s^{-2} kg^{-1})
M is the mass of the Earth (kg)
r is the distance from that point to Earth's center (m),
ω is Earth's rotation speed (radians/s).

Near the earth's surface the acceleration g_0 at radius r_0 is given by:

$g_0 = G \cdot M / r_0^2$(the other term is negligible), so that:
$G \cdot M = g_0 \cdot r_0^2$, which gives the $G \cdot M$ constant given the ground acceleration and planet radius.

At some point r_1 above the equator line, the two terms (gravity and centrifugal force) equal each other, objects fixed to the tether there have no weight. This occurs at the level of the stationary orbit:

$r_1 = (g_0 \cdot r_0^2 / \omega^2)^{1/3}$ which is to say $G \cdot M / r_1^2 = \omega^2 \cdot r_1$, which gives the value of r_1.

The same holds true for any planet or satellite.

Seen from a geosynchronous station, any object dropped off the tether from a point closer to Earth will initially accelerate downward. If dropped from any point above a geosynchronous station, the object would initially accelerate up toward space. If a long cable is dropped "down" (toward Earth), it must be balanced by balancing mass being dropped "up" (away from Earth) for the whole system to remain on the geosynchronous orbit. Some designs imagine the balancing mass being another cable (with counterweight) extending upward, other designs elevate the spool itself as the main cable is payed out. When the lower end of the cable is so long as to reach the Earth, it can be anchored at some place. Once anchored, if the center of mass is moved upward to be *above* the level of geosynchronous orbit (by adding mass at the upper end or by paying out more cable), it will add a tension to the whole cable, which can then be used as an elevator cable.

Cable section

Historically, the main technical problem has been considered to be the ability of the cable to hold up its own weight. The cable material combined with its design must be strong enough to hold up 35000 km (22,000 mi) of *itself*. By making any cable larger in cross section at the top compared to the bottom, it can hold up a longer length of itself. For a space elevator cable, an important design factor in addition to the material is how the cross section area tapers down from the maximum at 22,000 miles to the minimum at the surface. To maximize strength to weight of the cable, the cross section area will need to be designed in such a way that at any given point, it is proportional to the force it has to withstand. That is, the section must follow the following differential equation:

$$\sigma \cdot dS = g \cdot \rho \cdot S \cdot dr,$$

where

g is the acceleration along the radius (m·s^{-2}),
S is the cross-area of the cable at any given point r, (m^2) and dS its variation (m^2 as well),
ρ is the density of the material used for the cable (kg·m^{-3}).

σ is the stress a given area can bear without splitting ($N \cdot m^{-2} = kg \cdot m^{-1} \cdot s^{-2}$), its elastic limit

The value of g is given by the first equation, which yields:

$$\Delta \left[\ln(S)\right]_{r1}^{r0} = \rho/\sigma \cdot \Delta \left[G \cdot M/r + w^2 \cdot r^2/2\right]_{r1}^{r0},$$

the variation being taken between r_1 (geostationary) and r_0 (ground).

It turns out that between these two points, this quantity can be expressed simply as:
$\Delta \left[\ln(S)\right] = \rho/\sigma \cdot g_0 \cdot r_0 \cdot (1 + x/2 - 3/2 \cdot x^{1/3})$, or

$$S_0 = S_1.e^{\rho/\sigma \cdot g_0 \cdot r_0 \cdot (1+x/2-3/2 \cdot x^{1/3})}$$

where $x = \omega^2 \cdot r_0/g_0$ is the ratio between the centrifugal force on the equator and the gravitational force.

Thus, the factor with the main influence is $g_0\ r_0$, the combination of the planet's radius and its surface gravity. The rotational speed is slightly influential, but only as a corrective factor. For Earth, it reduces the strength needed by about one third.

Cable material

The second technical problem is that the $g_0\ r_0$ factor is quite large. Since its influence on the maximal cross-section is exponential, one needs to find materials where σ will be large enough to cancel our gravity. On Earth, we have:

$g_0 \cdot r_0 = 62.5 \cdot 10^6\ m^2 s^{-2}$ (or Joules per kg)
$\rho \approx 3 \cdot 10^3\ kg\ m^{-3}$ for most solid materials, so that σ needs to be:
$\sigma \approx 300 \cdot 10^9\ kg\ m^{-1} s^{-2}$

This corresponds to a cable capable of sustaining 30 tons with a cross-section of one square millimeter, under Earth's gravity.

The *free breaking length* can be used to compare materials: it is the length of a cylindrical cable at which it will split under its own weight (under constant gravity). For a given material, that length is $\sigma/\rho/g_0$. The free breaking length needed is given by the equation

$\Delta \left[\ln(S)\right] = \rho/\sigma \cdot g_0 \cdot r_0 \cdot (1 + x/2 - 3/2 \cdot x^{1/3})$, where
$x = w^2 \cdot r_0/g_0$

If one does not take into account the *x* factor (which reduces the strength needed by about 30%), this equation also says that the section ratio equals *e* (exponential one) when:

$$\sigma = \rho \cdot r_0 \cdot g_0$$

In other words, the free breaking length is approximately equal to the planet's radius under its own gravity. Since the section ratio varies exponentially, the free breaking length must be at least of that order of magnitude. If the material is only ten times less resilient, the section needed at a geosynchronous orbit will be e^{10} times the ground section, which is more than a hundredfold in diameter, which is practically impossible.

Structure

One concept for the space elevator has it tethered to a mobile seagoing platform.

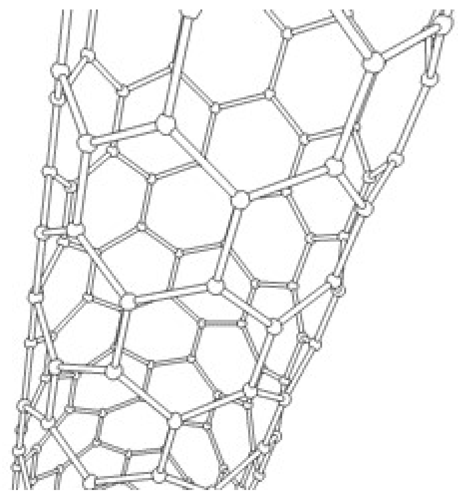
The centrifugal force of earth's rotation is the main principle behind the elevator. As the earth rotates, the centrifugal force tends to align the nanotube in a stretched manner. There are a variety of tether designs. Almost every design includes a base station, a cable, climbers, and a counterweight.

Base station

The base station designs typically fall into two categories—mobile and stationary. Mobile stations are typically large oceangoing vessels. Stationary platforms would generally be located in high-altitude locations, such as on top of mountains, or even potentially on high towers.

Mobile platforms have the advantage of being able to maneuver to avoid high winds, storms, and space debris. While stationary platforms don't have these advantages, they typically would have access to cheaper and more reliable power sources, and require a shorter cable. While the decrease in cable length may seem minimal (no more than a few kilometers), the cable thickness could be reduced over its entire length, significantly reducing the total weight.

Cable



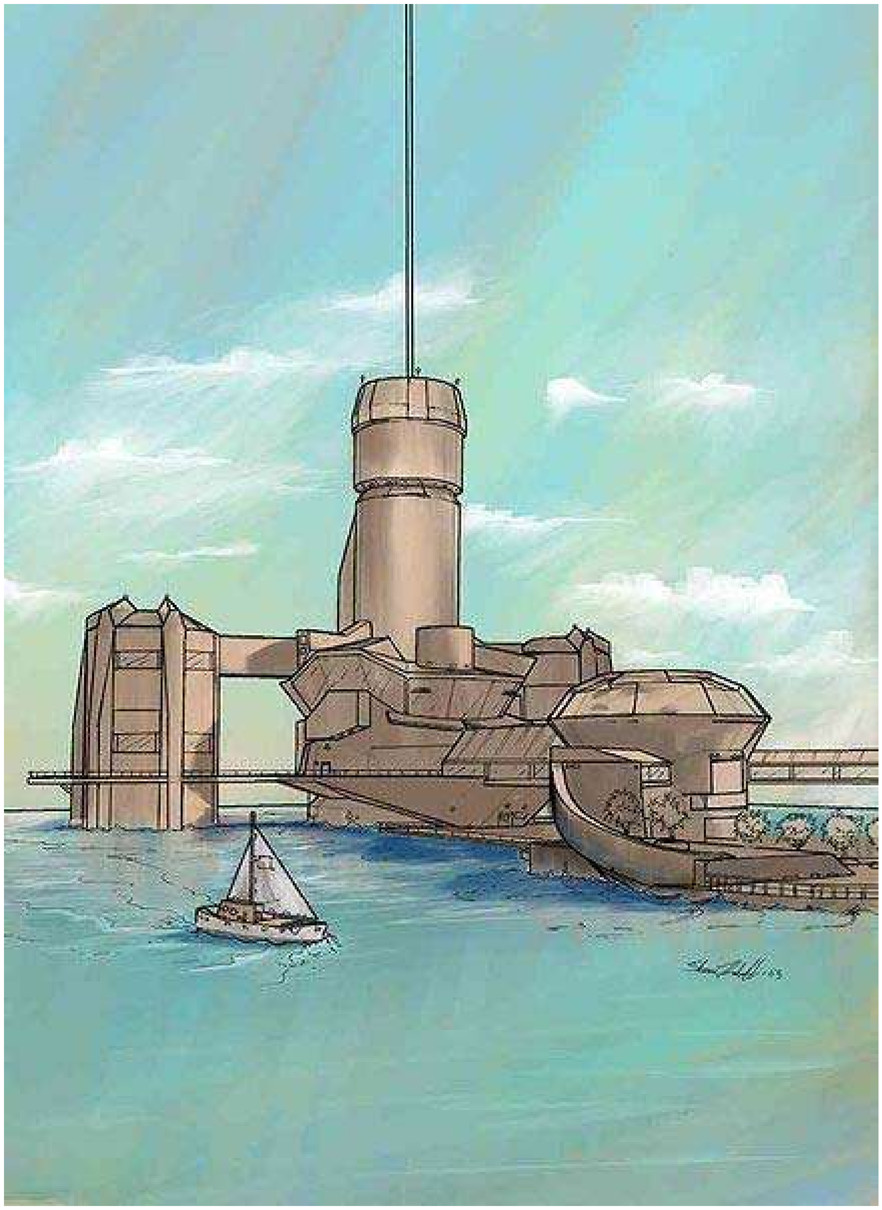
Carbon nanotubes are one of the candidates for a cable material

A space elevator cable must carry its own weight as well as the (smaller) weight of climbers. The required strength of the cable will vary along its length, since at various points it has to carry the weight of the cable below, or provide a centripetal force to retain the cable and counterweight above. In a 1998 report, NASA researchers noted that "maximum stress [*sic*] [on a space elevator cable] is at geosynchronous altitude so the cable must be thickest there and taper exponentially as it approaches Earth. Any potential material may be characterized by the taper factor – the ratio between the cable's radius at geosynchronous altitude and at the Earth's surface."

The cable must be made of a material with a large tensile strength/mass ratio. For example, the Edwards space elevator design assumes a cable material with a specific strength of at least 100,000 kN/(kg/m). This value takes into consideration the entire weight of the space elevator. A space elevator would need a material capable of sustaining a length of 4,960 kilometers (3082 mi) of its own weight *at sea level* to reach a geostationary altitude of 36,000 km (22,300 mi) without tapering and without breaking. Therefore, a material with very high strength and lightness is needed.

For comparison, metals like titanium, steel or aluminium alloys have breaking lengths of only 20–30 km. Modern fibre materials (which tend to achieve greater strength because the microscopic or crystal structure is aligned with the axis of the material and has fewer defects) such as kevlar, fibreglass and carbon/graphite fibre have breaking lengths of 100–400 km. Quartz fibers have an advantage that they can be drawn to a length of hundreds of kilometers even with the present-day technology. Nanoengineered materials such as carbon nanotubes and, more recently discovered, graphene ribbons (perfect two-dimensional sheets of carbon) are expected to have breaking lengths of 5000–6000 km at sea level, and also are able to conduct electrical power.

Carbon is such a good candidate material (for high specific strength) because, as only the 6th element in the periodic table, it has very few of the nucleons which contribute most of the dead weight of any material (whereas most of the interatomic bonding forces are contributed by only the outer few electrons); the challenge now remains to extend to macroscopic sizes the production of such material that are still perfect on the microscopic scale (as microscopic defects are most responsible for material weakness). The current (2009) carbon nanotube technology allows growing tubes up to a few tens of centimeters only.



A seagoing anchor station would incidentally act as a deep-water seaport.

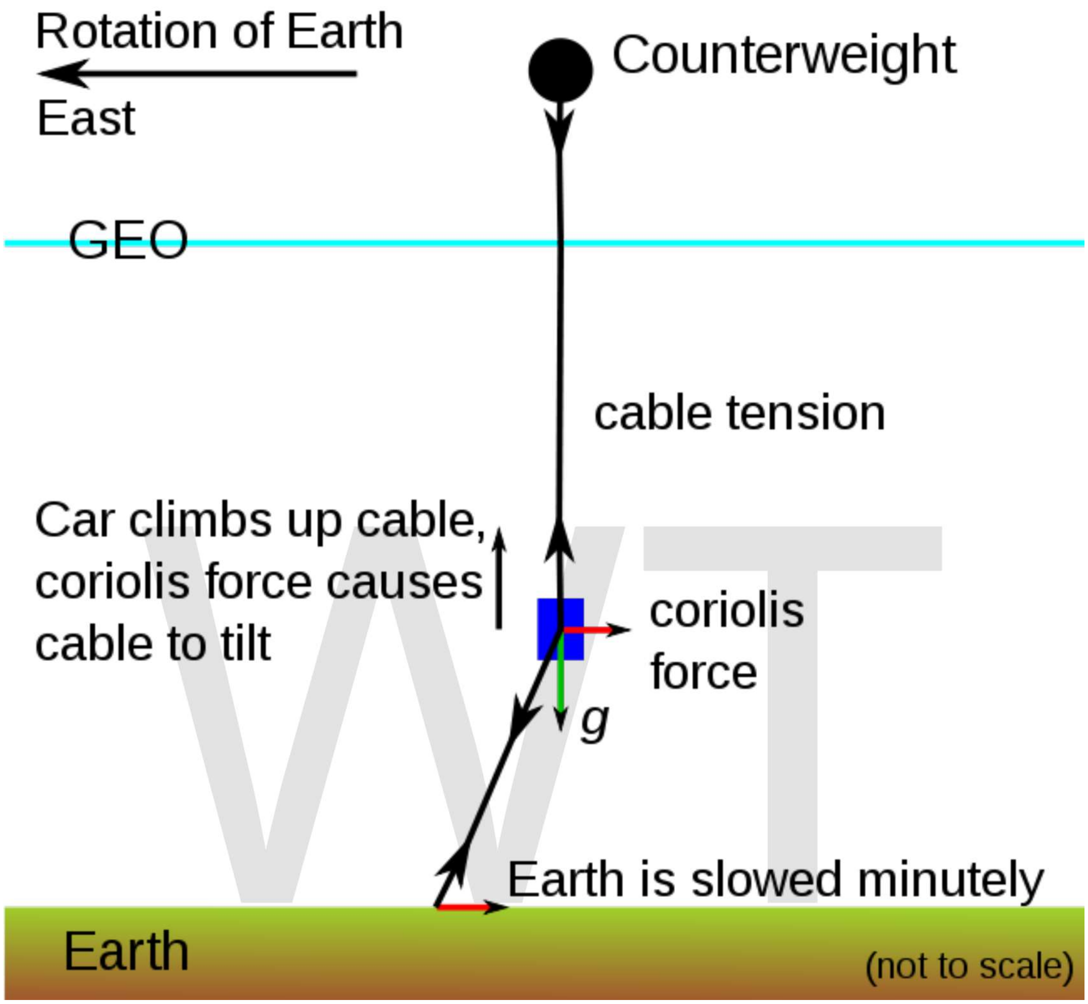
Climbers

A conceptual drawing of a space elevator climbing through the clouds.

A space elevator cannot be an elevator in the typical sense (with moving cables) due to the need for the cable to be significantly wider at the center than the tips. While various designs employing moving cables have been proposed, most cable designs call for the "elevator" to climb up a stationary cable.

Climbers cover a wide range of designs. On elevator designs whose cables are planar ribbons, most propose to use pairs of rollers to hold the cable with friction.

Climbers must be paced at optimal timings so as to minimize cable stress and oscillations and to maximize throughput. Lighter climbers can be sent up more often, with several going up at the same time. This increases throughput somewhat, but lowers the mass of each individual payload.



As the car climbs, the elevator takes on a 1 degree lean, due to the top of the elevator traveling faster than the bottom around the Earth (Coriolis force). This diagram is not to scale.

The horizontal speed of each part of the cable increases with altitude, proportional to distance from the center of the Earth, reaching orbital velocity at geostationary orbit. Therefore as a payload is lifted up a space elevator, it needs to gain not only altitude but angular momentum (horizontal speed) as well. This angular momentum is taken from the Earth's own rotation. As the climber ascends it is initially moving slightly more slowly than the cable that it moves onto (Coriolis force) and thus the climber "drags" on the cable.

The overall effect of the centrifugal force acting on the cable causes it to constantly try to return to the energetically favourable vertical orientation, so after an object has been lifted on the cable the counterweight will swing back towards the vertical like an inverted pendulum. Provided that the space elevator is designed so that the center of weight always stays above geostationary orbit for the maximum climb speed of the climbers, the elevator cannot fall over. Lift and descent operations must be carefully planned so as to keep the pendulum-like motion of the counterweight around the tether point under control.

By the time the payload has reached GEO the angular momentum (horizontal speed) is enough that the payload is in orbit.

The opposite process would occur for payloads descending the elevator, tilting the cable eastwards and insignificantly increasing Earth's rotation speed.

It has also been proposed to use a second cable attached to a platform to lift payload up the main cable, since the lifting device would not have to deal with its own weight against Earth's gravity. Out of the many proposed theories, powering any lifting device also continues to present a challenge.

Another design constraint will be the ascending speed of the climber. As geosynchronous orbit is at 35,786 km (22,236 mi). Assuming the climber can reach the speed of a very fast car or train of 300 km/h (180 mph) it will take 5 days to climb to geosynchronous orbit.

Powering climbers

Both power and energy are significant issues for climbers—the climbers need to gain a large amount of potential energy as quickly as possible to clear the cable for the next payload.

All proposals to get that energy to the climber fall into 3 categories:

- transfer the energy to the climber through wireless energy transfer while it is climbing
- transfer the energy to the climber through some material structure while it is climbing
- store the energy in the climber before it starts—this requires an extremely high specific energy. Nuclear energy and solar power have been proposed, but generating enough energy to reach the top of the elevator in any reasonable time without weighing too much is not feasible.

The proposed method is laser power beaming, using megawatt powered free electron or solid state lasers in combination with adaptive mirrors approximately 10 m (33 ft) wide and a photovoltaic array on the climber tuned to the laser frequency for efficiency. A major obstacle for any climber design is the dissipation of the substantial amount of waste heat generated due to the less than perfect efficiency of any of the power methods.

Yoshio Aoki, a professor of precision machinery engineering at Nihon University and director of the Japan Space Elevator Association, suggested including a second cable and using the conductivity of carbon nanotubes to provide power.

Various mechanical means of applying power have also been proposed; such as moving, looped or vibrating cables.

Counterweight

Several solutions have been proposed to act as a counterweight:

1. a heavy, captured asteroid;
2. a space dock, space station or spaceport positioned past geostationary orbit; or
3. an extension of the cable itself far beyond geostationary orbit.

The third idea has gained more support in recent years due to the relative simplicity of the task and the fact that a payload that went to the end of the counterweight-cable would acquire considerable velocity relative to the Earth, allowing it to be launched into interplanetary space.

Additionally, Brad Edwards has proposed that initially elevators would be up-only, and that the elevator cars that are used to thicken the cable could simply be parked at the top of the cable and act as a counterweight.

Alternative concepts

The original concept envisioned by Tsiolkovsky was a compression structure, a concept similar to an aerial mast. While such structures might reach the agreed altitude for space (100 km—62 mi), they are unlikely to reach geostationary orbit (35,786 km—22,236 mi). The concept of a Tsiolkovsky tower combined with a classic space elevator cable has been suggested.

A mini version of the Space Elevator to access near-space altitudes of 20 km (12 mi) has been proposed by Canadian researchers. The structure would be pneumatically supported and free standing with control systems guiding the structure's center of gravity. Proposed uses include tourism and commerce, communications, wind generation and low-cost space launch.

Other alternatives to a space elevator include an orbital ring, a pneumatic space tower, a space fountain, a launch loop, a Skyhook, a space tether, and a space hoist.

Launching into deep space

The velocities that might be attained at the end of Pearson's 144,000 km (90,000 mi) cable can be determined. The tangential velocity is 10.93 kilometers per second (6.79 mi/s), which is more than enough to escape Earth's gravitational field and send probes at least as far out as Jupiter. Once at Jupiter, a gravitational assist maneuver permits solar escape velocity to be reached.

Extraterrestrial elevators

A space elevator could also be constructed on other planets, asteroids and moons.

A Martian tether could be much shorter than one on Earth. Mars' surface gravity is 38% of Earth's, while it rotates around its axis in about the same time as Earth. Because of this, Martian areostationary orbit is much closer to the surface, and hence the elevator would be much shorter. Current materials are already sufficiently strong to construct such an elevator. However, building a Martian elevator would be complicated by the Martian moon Phobos, which is in a low orbit and intersects the Equator regularly (twice every orbital period of 11 h 6 min).

A lunar space elevator can possibly be built with currently available technology about 50,000 kilometers (31,000 miles) long extending through the Earth-Moon L1 point from an anchor point near the center of the visible part of Earth's moon. However, the lack of an atmosphere allows for other, perhaps better, alternatives to rockets.

On the far side of the moon, a lunar space elevator would need to be very long (more than twice the length of an Earth elevator) but due to the low gravity of the Moon, can be made of existing engineering materials.

Rapidly spinning asteroids or moons could use cables to eject materials to convenient points, such as Earth orbits; or conversely, to eject materials to send the bulk of the mass of the asteroid or moon to Earth orbit or a Lagrangian point. Freeman Dyson, a physicist and mathematician, has suggested using such smaller systems as power generators at points distant from the Sun where solar power is uneconomical. For the purpose of mass ejection, it is not necessary to rely on the asteroid or moon to be rapidly spinning. Instead of attaching the tether to the equator of a rotating body, it can be attached to a rotating hub on the surface. This was suggested in 1980 as a "Rotary Rocket" by Pearson and described very succinctly on the Island One website as a "Tapered Sling".

Construction

The construction of a space elevator would be a vast project requiring advances in engineering, manufacturing, and physical technology.

Safety issues and construction challenges

Depending on transit times through the Van Allen radiation belts passengers will need to be protected from radiation by shielding, which adds mass to the climber and decreases payload. For early systems, transit times are expected to be long enough where, if unshielded, total exposure would be above levels considered safe.

A space elevator would present a navigational hazard, both to aircraft and spacecraft. Aircraft could be diverted by air-traffic control restrictions. All objects in stable orbits that have perigee below the maximum altitude of the cable that are not synchronous with the cable will impact the cable eventually, unless avoiding action is taken. For spacecraft one potential solution proposed by Edwards is to use a movable anchor (a sea anchor) to allow the tether to "dodge" any space debris large enough to track.

Impacts by space objects such as meteoroids, micrometeorites and orbiting man-made debris, pose a more difficult problem, because the potential of a strand break to cause a failure cascade is, according to Tom Nugent, the Research Director of LiftPort Inc., "A potential show-stopper for construction of the space elevator [that] has not yet been adequately addressed.”.

Economics

With a space elevator, materials might be sent into orbit at a fraction of the current cost. As of 2000, conventional rocket designs cost about $11,000 per pound ($25,000 per kilogram) for transfer to geostationary orbit. Current proposals envision payload prices starting as low as $100 per pound ($220 per kilogram), similar to the $5–$300/kg estimates of the Launch loop, although nowhere near the $310/ton to 500 km orbit quoted to Dr. Jerry Pournelle for an orbital airship system.

Philip Ragan, co-author of the book "Leaving the Planet by Space Elevator", states that "The first country to deploy a space elevator will have a 95 percent cost advantage and could potentially control all space activities."