Microelectronics & Macroelectronics

Katharyn Hostetler

Cletus Schilling

First Edition, 2012

ISBN 978-81-323-0954-3

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

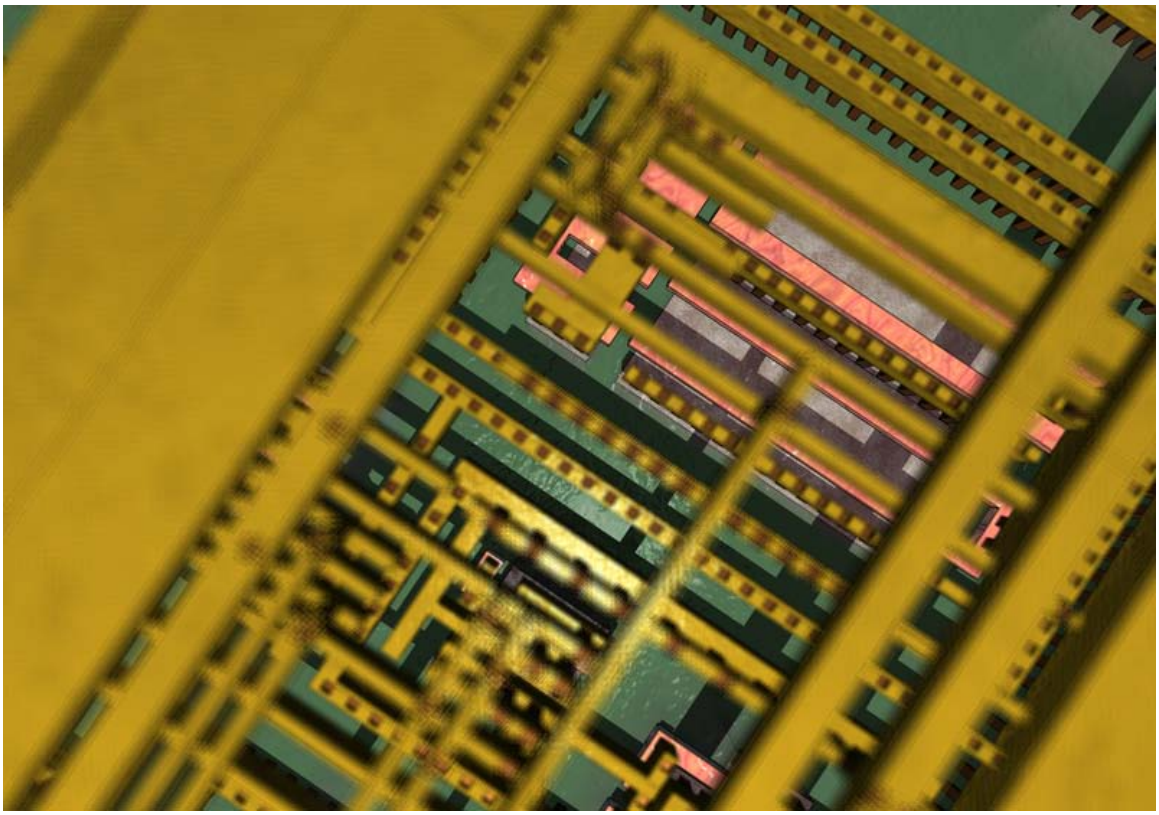
Table of Contents

Chapter-1

Microfabrication

Microfabrication is the term that describes processes of fabrication of miniature structures, of micrometre sizes and smaller. Historically the earliest microfabrication processes were used for integrated circuit fabrication, also known as "semiconductor device fabrication," "semiconductor manufacturing, VLSI technology, microelectronic fabrication". In the last two decades microelectromechanical systems (MEMS), microsystems (European usage), micromachines (Japanese terminology) and their subfields, microfludics/lab-on-a-chip, optical MEMS (also called MOEMS), RF MEMS, PowerMEMS, BioMEMS and their extension into nanoscale (for example NEMS, for nano electro mechanical systems) have re-used, adapted or extended microfabrication methods. Flat-panel displays and solar cells are also using similar techniques.

Miniaturization of various devices presents challenges in many areas of science and engineering: physics, chemistry, material science, computer science, ultra-precision engineering, fabrication processes, and equipment design. It is also giving rise to various kinds of interdisciplinary research.



Synthetic detail of a micromanufactured integrated circuit through four layers of planarized copper interconnect, down to the polysilicon (pink), wells (greyish) and substrate (green).

The major concepts and principles of microfabrication are microlithography, doping, thin films, etching, bonding, polishing,

Fields of Use

Microfabricated devices include:

- Fabrication of integrated circuits (“microchips”)
- Microelectromechanical systems (MEMS), MOEMS,
- microfluidic devices (ink jet print heads)
- solar cells
- Flat Panel Displays
- Sensors (micro-sensors) (biosensors, nanosensors)
- PowerMEMSs, fuel cells, energy harvesters/scavengers

Origins

Microfabrication technologies originate from the microelectronics industry, and the devices are usually made on silicon wafers even though glass, plastics and many other substrate are in use. Micromachining, semiconductor processing, microelectronic fabrication, semiconductor fabrication, MEMS fabrication and integrated circuit technology are terms used instead of microfabrication, but microfabrication is the broad general term.

Traditional machining techniques such as *electro-discharge machining*, *spark erosion machining*, and *laser drilling* have been scaled from the millimeter size range to micrometer range, but they do not share the main idea of microelectronics-originated microfabrication: replication and parallel fabrication of hundreds or millions of identical structures. This parallelism is present in various imprint, casting and moulding techniques which have successfully been applied in the microregime. For example, injection moulding of DVDs involves fabrication of submicrometer-sized spots on the disc.

Microfabrication processes

Microfabrication is actually a collection of technologies which are utilized in making microdevices. Some of them have very old origins, not connected to manufacturing, like lithography or etching. Polishing was borrowed from optics manufacturing, and many of the vacuum techniques come from 19th century physics research. Electroplating is also a 19th century technique adapted to produce micrometre scale structures, as are various stamping and embossing techniques.

To fabricate a microdevice, many processes must be performed, one after the other, many times repeatedly. These processes typically include depositing a film, patterning the film with the desired micro features, and removing (or etching) portions of the film. For example, in memory chip fabrication there are some 30 lithography steps, 10 oxidation steps, 20 etching steps, 10 doping steps, and many others are performed. The complexity of microfabrication processes can be described by their *mask count*. This is the number of different pattern layers that constitute the final device. Modern microprocessors are made with 30 masks while a few masks suffice for a microfluidic device or a laser diode. Microfabrication resembles multiple exposure photography, with many patterns aligned to each other to create the final structure.

Substrates

Microfabricated devices are not generally freestanding devices but are usually formed over or in a thicker support substrate. For electronic applications, semiconducting substrates such as silicon wafers can be used. For optical devices or flat panel displays, transparent substrates such as glass or quartz are common. The substrate enables easy handling of the micro device through the many fabrication steps. Often many individual devices are made together on one substrate and then singulated into separated devices toward the end of fabrication.

Deposition or Growth

Microfabricated devices are typically constructed using one or more thin films. The purpose of these thin films depends on the type of device. Electronic devices may have thin films which are conductors (metals), insulators (dielectrics) or semiconductors. Optical devices may have films which are reflective, transparent, light guiding or scattering. Films may also have a chemical or mechanical purpose as well as for MEMS applications. Examples of deposition techniques include:

- Thermal oxidation
- chemical vapor deposition (CVD)
  - APCVD
  - LPCVD
  - PECVD
- Physical vapor deposition(PVD)
  - sputtering
  - evaporative deposition
  - Electron beam PVD
- epitaxy

Patterning

It is often desirable to pattern a film into distinct features or to form openings (or vias) in some of the layers. These features are on the micrometer or nanometer scale and the patterning technology is what defines microfabrication. The patterning technique typically uses a 'mask' to define portions of the film which will be removed. Examples of patterning techniques include:

- Photolithography
- Shadow Masking

Etching

Etching is the removal of some portion of the thin film or substrate. The substrate is exposed to an etching (such as an acid or plasma) which chemically or physically attacks the film until it is removed. Etching techniques include:

- Dry etching (Plasma etching) such as Reactive-ion etching (RIE) or Deep reactive-ion etching(DRIE)
- Wet etching or Chemical Etching

Other

A wide variety of other processes for cleaning, planarizing, or modifying the chemical properties of the microfabricated devices can also be performed. Some examples include:

- Doping by either thermal diffusion or ion implantation
- Chemical-mechanical planarization (CMP)
- Wafer cleaning, also known as "surface preparation"
- Wire bonding

Micro cutting / microfabrication

Micro cutting/milling is an alternative to lithographic techniques, by downscaling macro processes such as cutting and forming, to tool sizes under 100 μm in diameter.

Cleanliness in wafer fabrication

Microfabrication is carried out in cleanrooms, where air has been filtered of particle contamination and temperature, humidity, vibrations and electrical disturbances are under stringent control. Smoke, dust, bacteria and cells are micrometers in size, and their presence will destroy the functionality of a microfabricated device.

Cleanrooms provide passive cleanliness but the wafers are also actively cleaned before every critical step. RCA-1 clean in ammonia-peroxide solution removes organic contamination and particles; RCA-2 cleaning in hydrogen chloride-peroxide mixture removes metallic impurities. Sulfuric acid-peroxide mixture (a.k.a. Piranha) removes organics. Hydrogen fluoride removes native oxide from silicon surface. These are all wet cleaning steps in solutions. Dry cleaning methods include oxygen and argon plasma treatments to remove unwanted surface layers, or hydrogen bake at elevated temperature to remove native oxide before epitaxy. Pre-gate cleaning is the most critical cleaning step in CMOS fabrication: it ensures that the ca. 2 nm thick oxide of a MOS transistor can be grown in an orderly fashion. Oxidation, and all high temperature steps are very sensitive to contamination, and cleaning steps must precede high temperature steps.

Surface preparation is just a different viewpoint, all the steps are the same as described above: it is about leaving the wafer surface in a controlled and well known state before you start processing. Wafers are contaminated by previous process steps (e.g. metals bombarded from chamber walls by energetic ions during ion implantation), or they may have gathered polymers from wafer boxes, and this might be different depending on wait time.

Wafer cleaning and surface preparation work a little bit like the machines in a bowling alley: first they remove all unwanted bits and pieces, and then they reconstruct the desired pattern so that the game can go on.

Chapter-2

Semiconductor

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

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

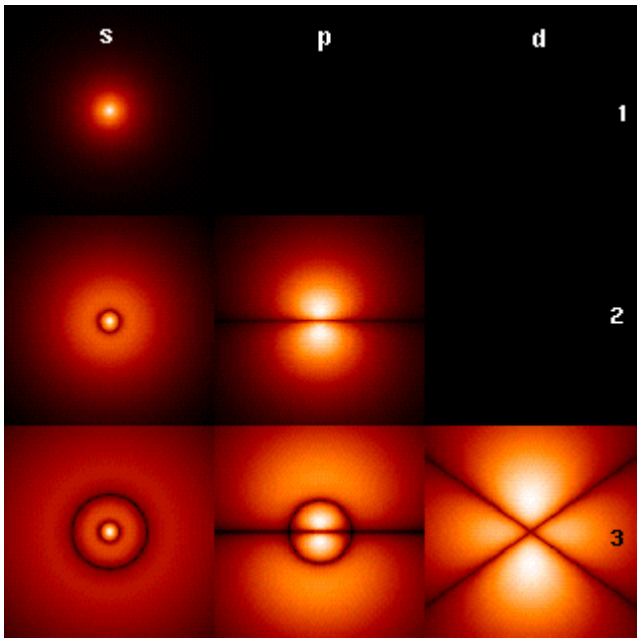
Silicon is used to create most semiconductors commercially. Dozens of other materials are used, including germanium, gallium arsenide, and silicon carbide. A pure semiconductor is often called an “intrinsic” semiconductor. The electronic properties and the conductivity of a semiconductor can be changed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

Explaining semiconductor energy bands

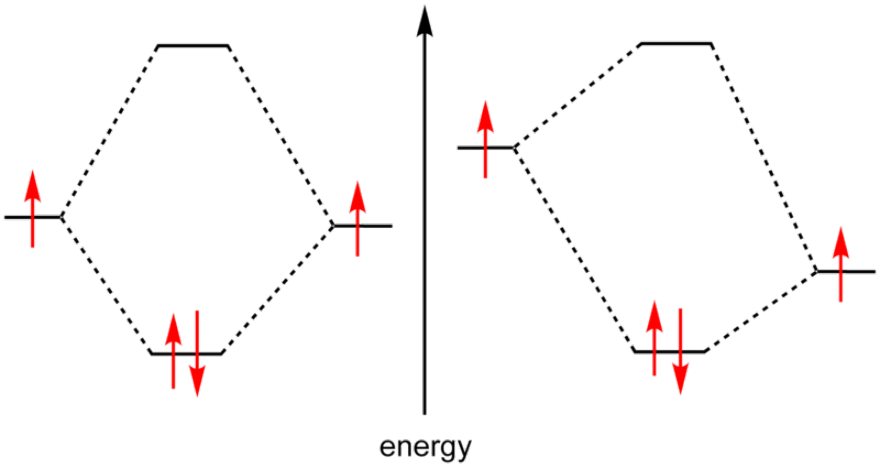
There are three popular ways to classify the electronic structure of a crystal.

- Band structure

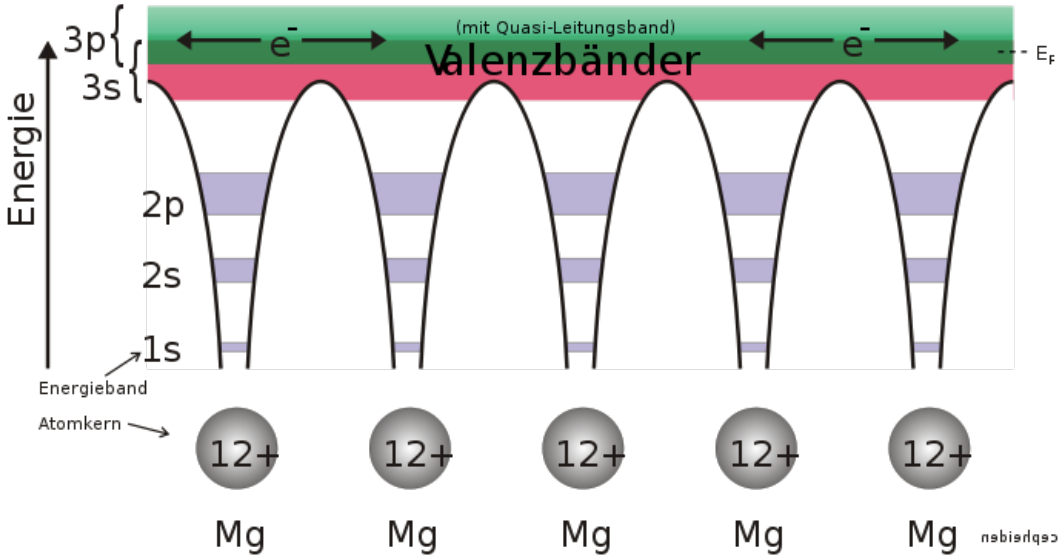
atoms – crystal – vacuum



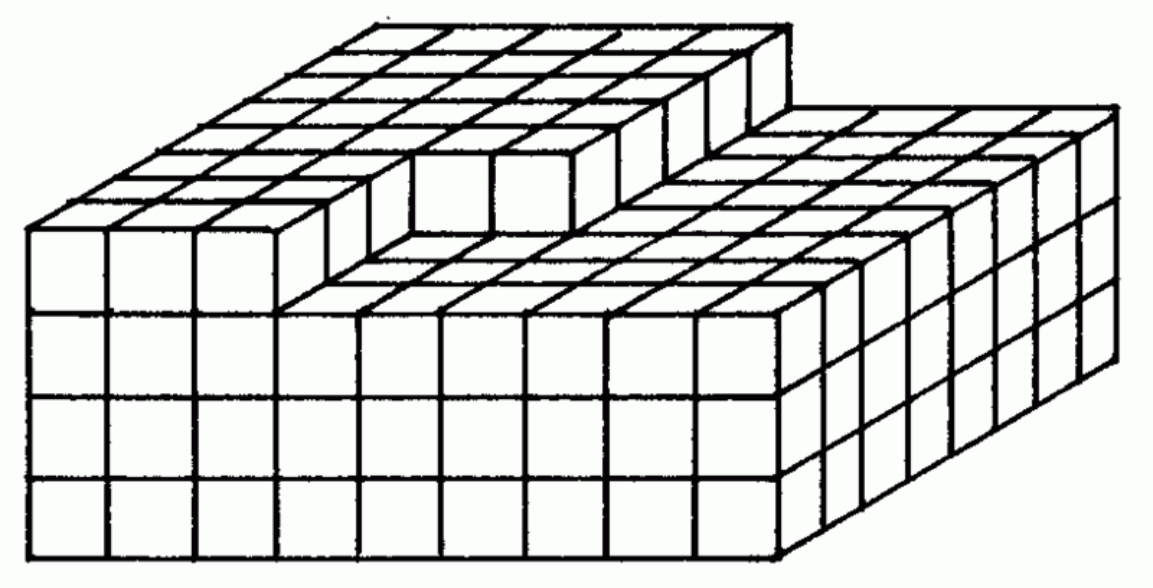
In a single H-atom an electron resides in well known orbitals. Note that the orbitals are called s,p,d in order of increasing circular current.



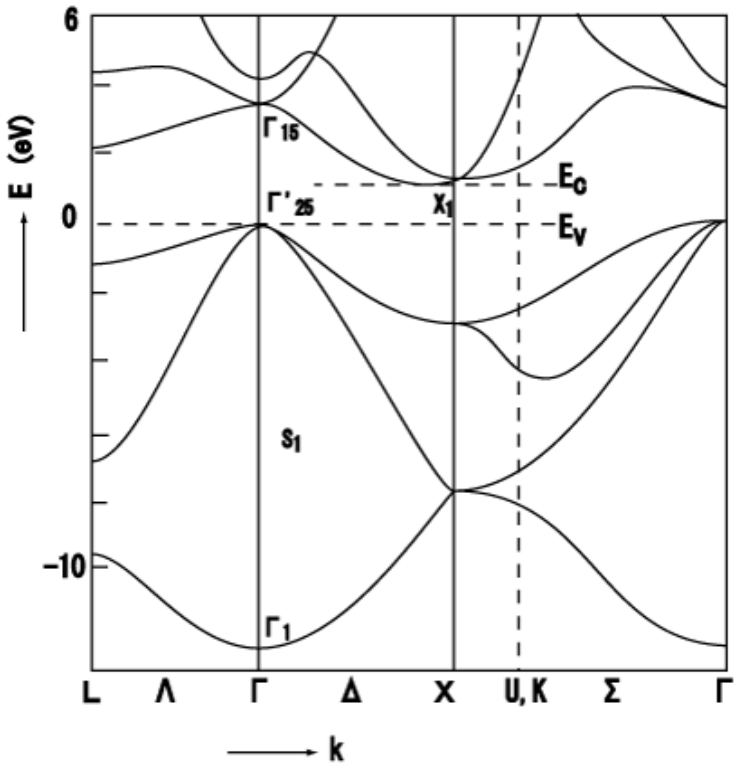
Putting two atoms together leads to delocalized orbitals across two atoms, yielding a covalent bond. Due to the Pauli exclusion principle, every state can contain only one electron.



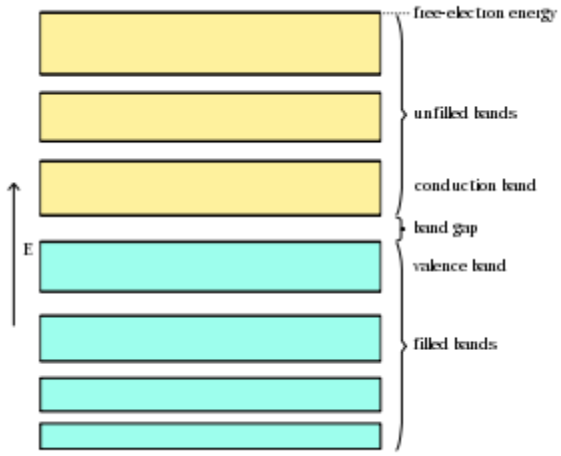
This can be continued with more atoms. Note: This picture shows a metal, not an actual semiconductor.



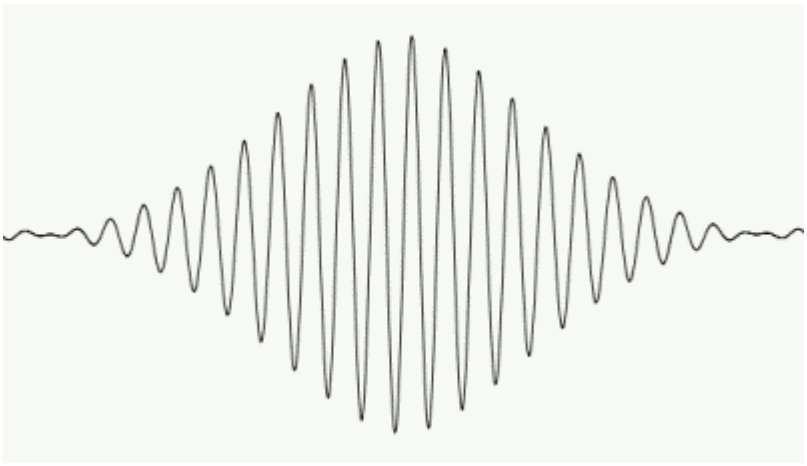
Continuing to add creates a crystal, which may then be cut into a tape and fused together at the ends to allow circular currents.



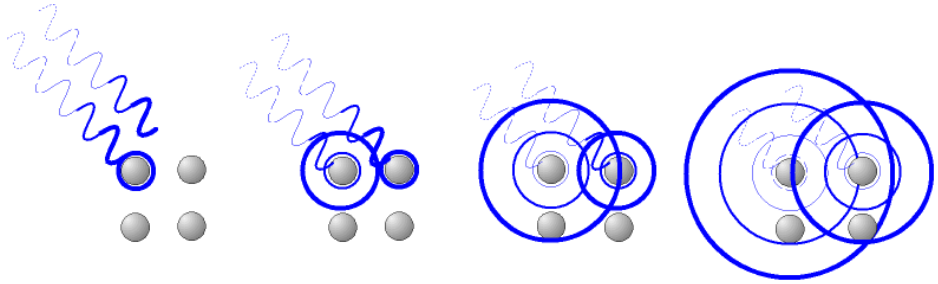
For this regular solid the band structure can be calculated or measured.



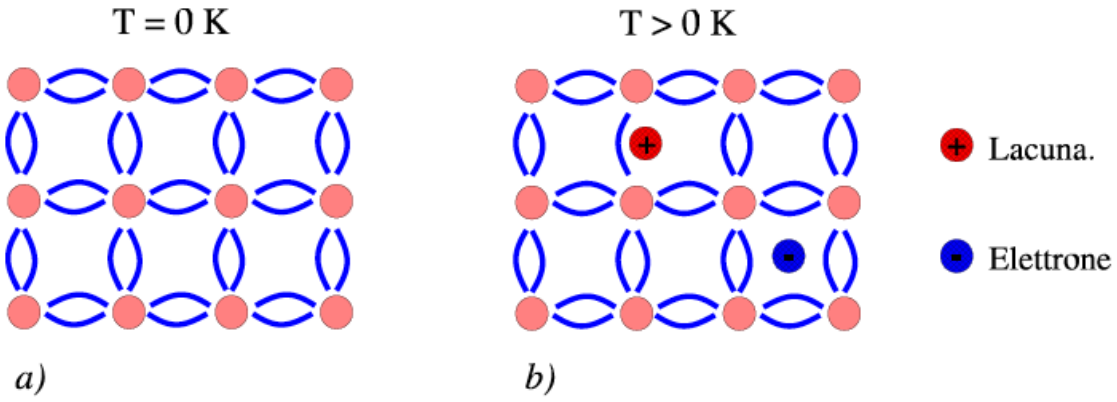
Integrating over the k axis gives the bands of a semiconductor showing a full valence band and an empty conduction band. Generally stopping at the vacuum level is undesirable, because some people want to calculate: photoemission, inverse photoemission



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



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



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

Energy bands and electrical conduction

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

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

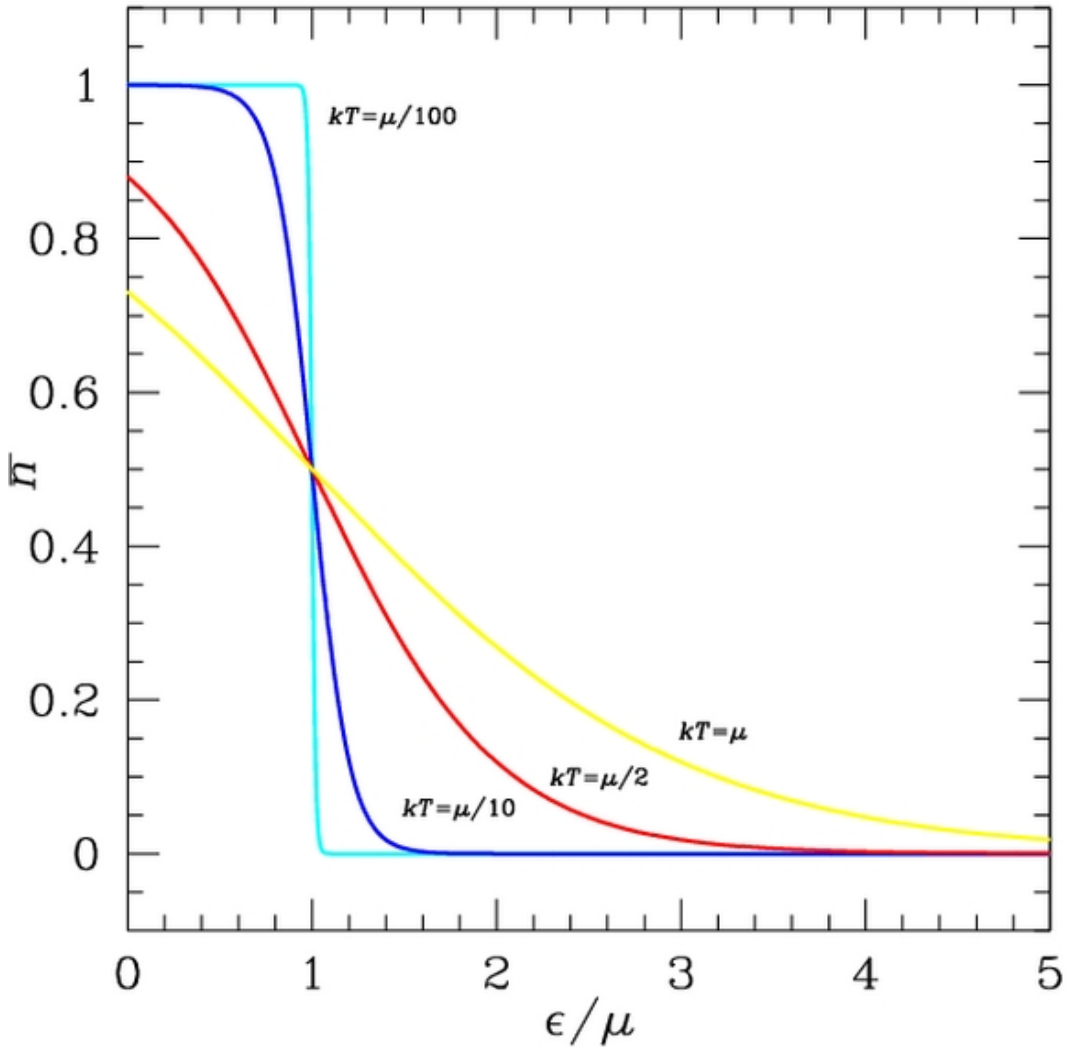
With covalent bonds, an electron moves by hopping to a neighboring bond. The Pauli exclusion principle requires the electron to be lifted into the higher anti-bonding state of that bond. For delocalized states, for example in one dimension – that is in a nanowire, for every energy there is a state with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

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

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

Holes: electron absence as a charge carrier

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



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

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

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

Energy–momentum dispersion

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

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

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

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

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

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

Carrier generation and recombination

When ionizing radiation strikes a semiconductor, it may excite an electron out of its energy level and consequently leave a hole. This process is known as *electron–hole pair*

generation. Electron-hole pairs are constantly generated from thermal energy as well, in the absence of any external energy source.

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

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

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

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

Semi-insulators

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

Doping

The property of semiconductors that makes them most useful for constructing electronic devices is that their conductivity may easily be modified by introducing impurities into their crystal lattice. The process of adding controlled impurities to a semiconductor is known as *doping*. The amount of impurity, or dopant, added to an *intrinsic* (pure) semiconductor varies its level of conductivity. Doped semiconductors are often referred to as *extrinsic*. By adding impurity to pure semiconductors, the electrical conductivity may be varied not only by the number of impurity atoms but also, by the type of impurity atom and the changes may be thousand folds and million folds. For example, 1 cm^3 of a metal or semiconductor specimen has a number of atoms on the order of 10^{22}. Since every atom in metal donates at least one free electron for conduction in metal, 1 cm^3 of

metal contains free electrons on the order of 10^{22}. At the temperature close to 20 °C , 1 cm^3 of pure germanium contains about 4.2×10^{22} atoms and 2.5×10^{13} free electrons and 2.5×10^{13} holes (empty spaces in crystal lattice having positive charge) The addition of 0.001% of arsenic (an impurity) donates an extra 10^{17} free electrons in the same volume and the electrical conductivity increases about 10,000 times."

Dopants

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

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

Carrier concentration

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

$$n = p = n_i.$$

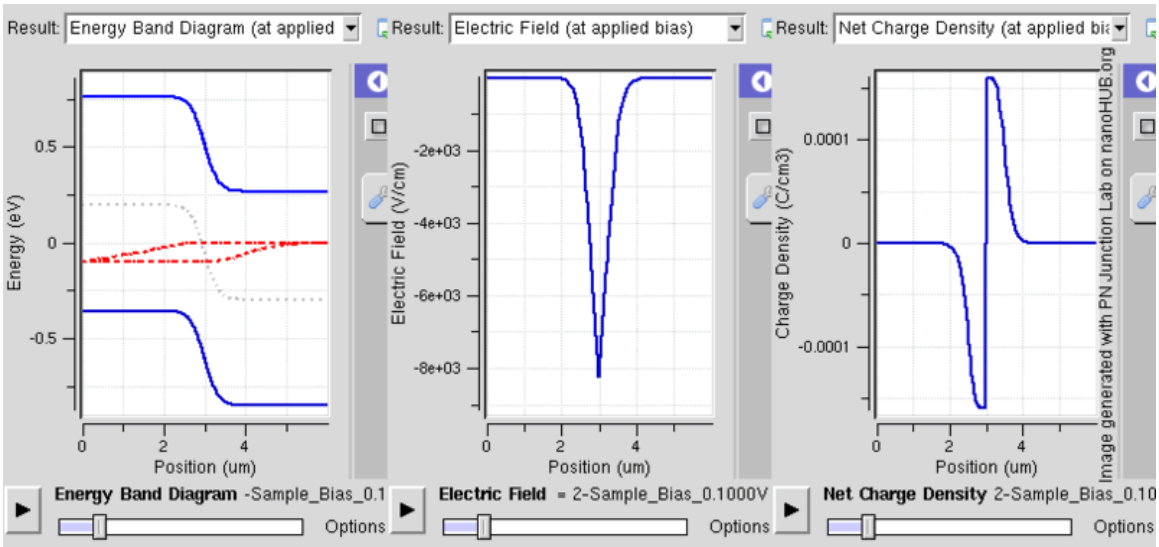
If we have a non-intrinsic semiconductor in thermal equilibrium the relation becomes:

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

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

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

Effect on band structure



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

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

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

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

Preparation of semiconductor materials

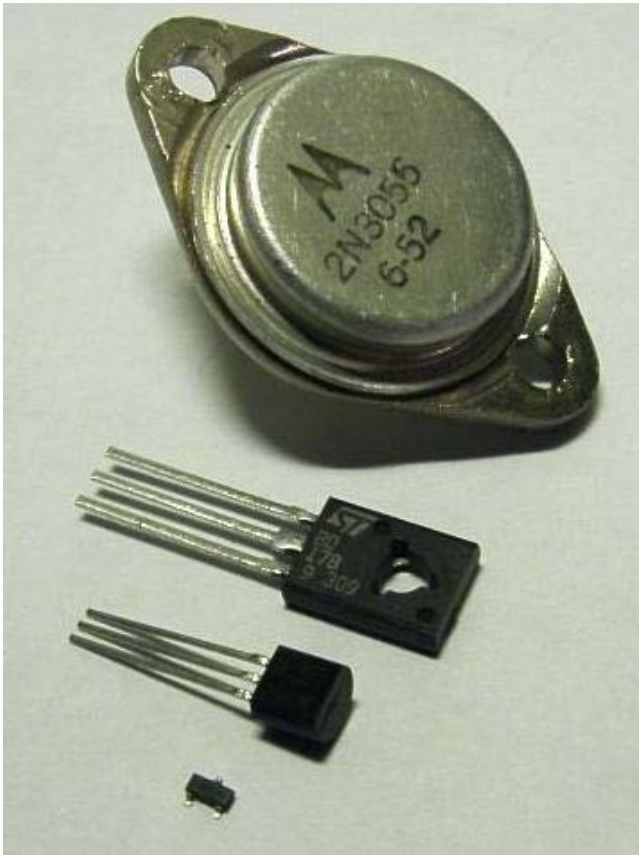
Semiconductors with predictable, reliable electronic properties are necessary for mass production. The level of chemical purity needed is extremely high because the presence of impurities even in very small proportions can have large effects on the properties of the material. A high degree of crystalline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

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

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

Chapter-3

Transistor



Assorted discrete transistors. Packages in order from top to bottom: TO-3, TO-126, TO-92, SOT-23

A **transistor** is a semiconductor device used to amplify and switch electronic signals. It is made of a solid piece of semiconductor material, with at least three terminals for connection to an external circuit. A voltage or current applied to one pair of the transistor's terminals changes the current flowing through another pair of terminals. Because the controlled (output) power can be much more than the controlling (input)

power, the transistor provides amplification of a signal. Today, some transistors are packaged individually, but many more are found embedded in integrated circuits.

The transistor is the fundamental building block of modern electronic devices, and is ubiquitous in modern electronic systems. Following its release in the early 1950s the transistor revolutionized the field of electronics, and paved the way for smaller and cheaper radios, calculators, and computers, among other things.

History

A replica of the first working transistor.

Physicist Julius Edgar Lilienfeld filed the first patent for a transistor in Canada in 1925, describing a device similar to a field-effect transistor or "FET". However, Lilienfeld did not publish any research articles about his devices, nor did his patent cite any examples of devices actually constructed. In 1934, German inventor Oskar Heil patented a similar device.

From 1942 Herbert Mataré experimented with so-called *duodiodes* while working on a detector for a Doppler RADAR system. The duodiodes built by him had two separate but very close metal contacts on the semiconductor substrate. He discovered effects that

could not be explained by two independently operating diodes and thus formed the basic idea for the later point contact transistor.

In 1947, John Bardeen and Walter Brattain at AT&T's Bell Labs in the United States observed that when electrical contacts were applied to a crystal of germanium, the output power was larger than the input. Solid State Physics Group leader William Shockley saw the potential in this, and over the next few months worked to greatly expand the knowledge of semiconductors. The term *transistor* was coined by John R. Pierce. According to physicist/historian Robert Arns, legal papers from the Bell Labs patent show that William Shockley and Gerald Pearson had built operational versions from Lilienfeld's patents, yet they never referenced this work in any of their later research papers or historical articles.

The name *transistor* is a portmanteau of the term "transfer resistor".

The first silicon transistor was produced by Texas Instruments in 1954. This was the work of Gordon Teal, an expert in growing crystals of high purity, who had previously worked at Bell Labs. The first MOS transistor actually built was by Kahng and Atalla at Bell Labs in 1960.

Importance

The transistor is the key active component in practically all modern electronics, and is considered by many to be one of the greatest inventions of the twentieth century. Its importance in today's society rests on its ability to be mass produced using a highly automated process (semiconductor device fabrication) that achieves astonishingly low per-transistor costs.

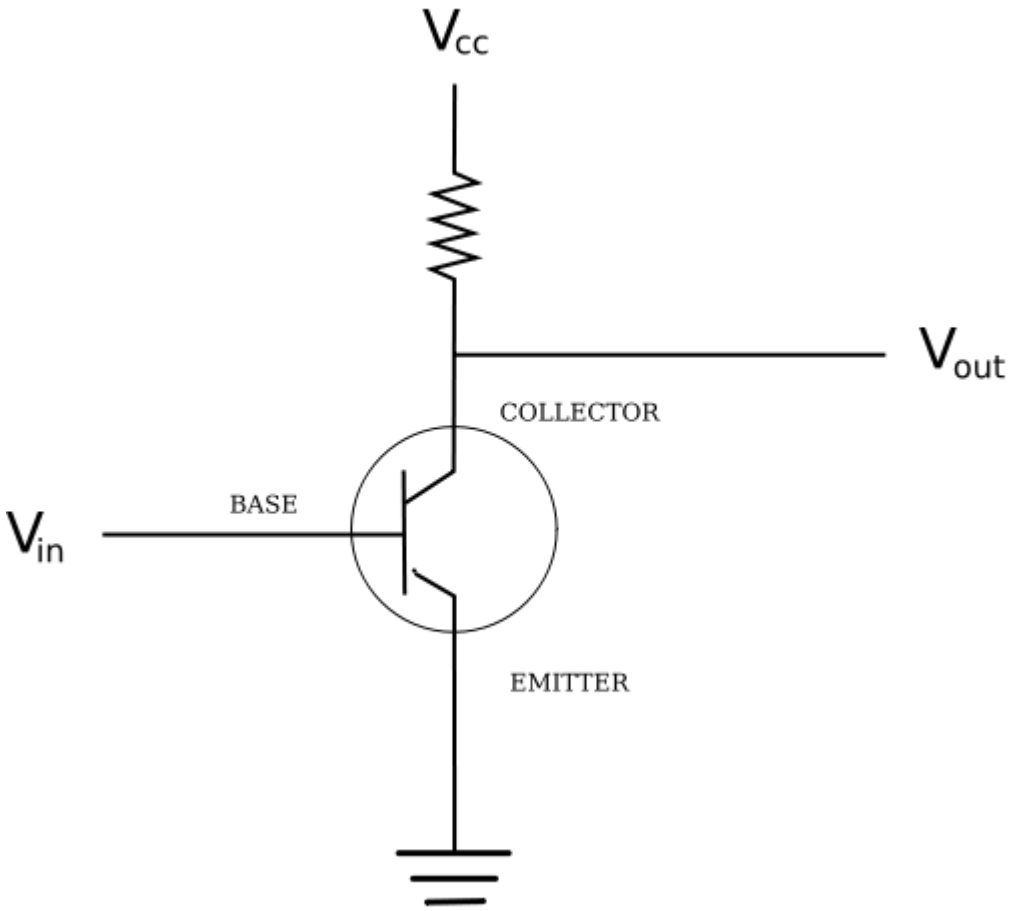
Although several companies each produce over a billion individually packaged (known as *discrete*) transistors every year, the vast majority of transistors now produced are in integrated circuits (often shortened to *IC*, *microchips* or simply *chips*), along with diodes, resistors, capacitors and other electronic components, to produce complete electronic circuits. A logic gate consists of up to about twenty transistors whereas an advanced microprocessor, as of 2011, can use as many as 3 billion transistors (MOSFETs). "About 60 million transistors were built this year [2002] ... for [each] man, woman, and child on Earth."

The transistor's low cost, flexibility, and reliability have made it a ubiquitous device. Transistorized mechatronic circuits have replaced electromechanical devices in controlling appliances and machinery. It is often easier and cheaper to use a standard microcontroller and write a computer program to carry out a control function than to design an equivalent mechanical control function.

Usage

The bipolar junction transistor, or BJT, was the most commonly used transistor in the 1960s and 70s. Even after MOSFETs became widely available, the BJT remained the transistor of choice for many analog circuits such as simple amplifiers because of their greater linearity and ease of manufacture. Desirable properties of MOSFETs, such as their utility in low-power devices, usually in the CMOS configuration, allowed them to capture nearly all market share for digital circuits; more recently MOSFETs have captured most analog and power applications as well, including modern clocked analog circuits, voltage regulators, amplifiers, power transmitters, motor drivers, etc.

Simplified operation



Simple circuit to show the labels of a bipolar transistor.

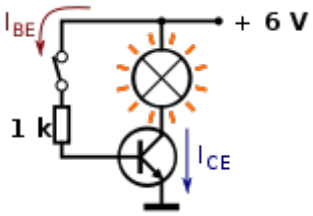
The essential usefulness of a transistor comes from its ability to use a small signal applied between one pair of its terminals to control a much larger signal at another pair of terminals. This property is called gain. A transistor can control its output in proportion to the input signal; that is, it can act as an amplifier. Alternatively, the transistor can be used

to turn current on or off in a circuit as an electrically controlled switch, where the amount of current is determined by other circuit elements.

The two types of transistors have slight differences in how they are used in a circuit. A *bipolar transistor* has terminals labeled **base**, **collector**, and **emitter**. A small current at the base terminal (that is, flowing from the base to the emitter) can control or switch a much larger current between the collector and emitter terminals. For a *field-effect transistor*, the terminals are labeled **gate**, **source**, and **drain**, and a voltage at the gate can control a current between source and drain.

The image to the right represents a typical bipolar transistor in a circuit. Charge will flow between emitter and collector terminals depending on the current in the base. Since internally the base and emitter connections behave like a semiconductor diode, a voltage drop develops between base and emitter while the base current exists. The amount of this voltage depends on the material the transistor is made from, and is referred to as V_{BE}.

Transistor as a switch



BJT used as an electronic switch, in grounded-emitter configuration.

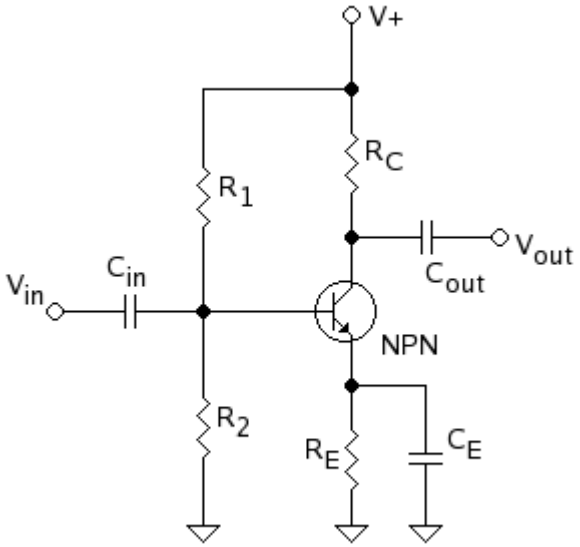
Transistors are commonly used as electronic switches, both for high-power applications such as switched-mode power supplies and for low-power applications such as logic gates.

In a grounded-emitter transistor circuit, such as the light-switch circuit shown, as the base voltage rises the base and collector current rise exponentially, and the collector voltage drops because of the collector load resistor. The relevant equations:

$V_{RC} = I_{CE} \times R_C$, the voltage across the load (the lamp with resistance R_C)
$V_{RC} + V_{CE} = V_{CC}$, the supply voltage shown as 6V

If V_{CE} could fall to 0 (perfect closed switch) then Ic could go no higher than V_{CC} / R_C, even with higher base voltage and current. The transistor is then said to be saturated. Hence, values of input voltage can be chosen such that the output is either completely off, or completely on. The transistor is acting as a switch, and this type of operation is common in digital circuits where only "on" and "off" values are relevant.

Transistor as an amplifier



Amplifier circuit, common-emitter configuration.

The common-emitter amplifier is designed so that a small change in voltage in (V_{in}) changes the small current through the base of the transistor and the transistor's current amplification combined with the properties of the circuit mean that small swings in V_{in} produce large changes in V_{out}.

Various configurations of single transistor amplifier are possible, with some providing current gain, some voltage gain, and some both.

From mobile phones to televisions, vast numbers of products include amplifiers for sound reproduction, radio transmission, and signal processing. The first discrete transistor audio amplifiers barely supplied a few hundred milliwatts, but power and audio fidelity gradually increased as better transistors became available and amplifier architecture evolved.

Modern transistor audio amplifiers of up to a few hundred watts are common and relatively inexpensive.

Comparison with vacuum tubes

Prior to the development of transistors, vacuum (electron) tubes (or in the UK "thermionic valves" or just "valves") were the main active components in electronic equipment.

Advantages

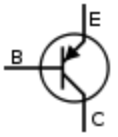
The key advantages that have allowed transistors to replace their vacuum tube predecessors in most applications are

- Small size and minimal weight, allowing the development of miniaturized electronic devices.
- Highly automated manufacturing processes, resulting in low per-unit cost.
- Lower possible operating voltages, making transistors suitable for small, battery-powered applications.
- No warm-up period for cathode heaters required after power application.
- Lower power dissipation and generally greater energy efficiency.
- Higher reliability and greater physical ruggedness.
- Extremely long life. Some transistorized devices have been in service for more than 50 years.
- Complementary devices available, facilitating the design of complementary-symmetry circuits, something not possible with vacuum tubes.
- Insensitivity to mechanical shock and vibration, thus avoiding the problem of microphonics in audio applications.

Limitations

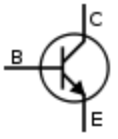
- Silicon transistors do not operate at voltages higher than about 1,000 volts (SiC devices can be operated as high as 3,000 volts). In contrast, vacuum tubes have been developed that can be operated at tens of thousands of volts.
- High-power, high-frequency operation, such as that used in over-the-air television broadcasting, is better achieved in vacuum tubes due to improved electron mobility in a vacuum.
- Silicon transistors are much more vulnerable than vacuum tubes to an electromagnetic pulse generated by a high-altitude nuclear explosion.

Types

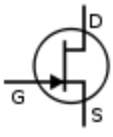


PNP

P-channel



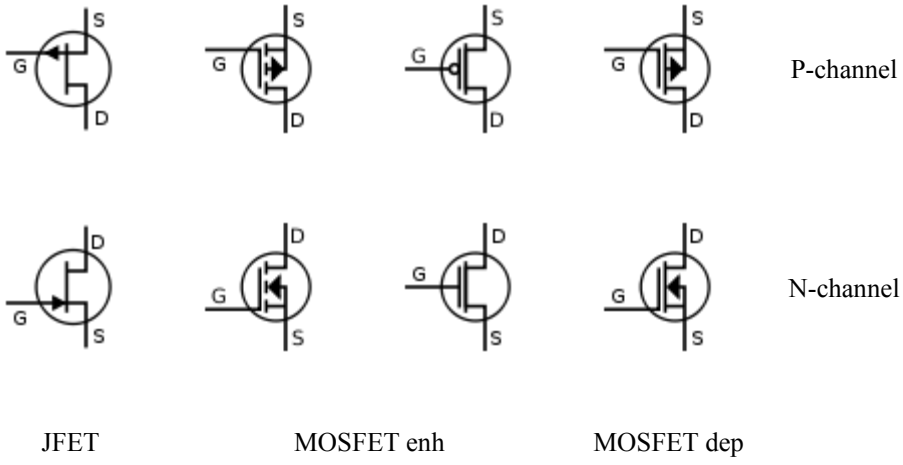
NPN



N-channel

BJT JFET

BJT and JFET symbols



JFET and IGFET symbols

Transistors are categorized by

- Semiconductor material: germanium, silicon, gallium arsenide, silicon carbide, etc.
- Structure: BJT, JFET, IGFET (MOSFET), IGBT, "other types"
- Polarity: NPN, PNP (BJTs); N-channel, P-channel (FETs)
- Maximum power rating: low, medium, high
- Maximum operating frequency: low, medium, high, radio frequency (RF), microwave (The maximum effective frequency of a transistor is denoted by the term f_T, an abbreviation for "frequency of transition". The frequency of transition is the frequency at which the transistor yields unity gain).
- Application: switch, general purpose, audio, high voltage, super-beta, matched pair
- Physical packaging: through hole metal, through hole plastic, surface mount, ball grid array, power modules
- Amplification factor h_{fe} (transistor beta)

Thus, a particular transistor may be described as *silicon, surface mount, BJT, NPN, low power, high frequency switch*.

Bipolar junction transistor

Bipolar transistors are so named because they conduct by using both majority and minority carriers. The bipolar junction transistor (BJT), the first type of transistor to be mass-produced, is a combination of two junction diodes, and is formed of either a thin layer of p-type semiconductor sandwiched between two n-type semiconductors (an n-p-n transistor), or a thin layer of n-type semiconductor sandwiched between two p-type

semiconductors (a p-n-p transistor). This construction produces two p-n junctions: a base–emitter junction and a base–collector junction, separated by a thin region of semiconductor known as the base region (two junction diodes wired together without sharing an intervening semiconducting region will not make a transistor).

The BJT has three terminals, corresponding to the three layers of semiconductor – an *emitter*, a *base*, and a *collector*. It is useful in amplifiers because the currents at the emitter and collector are controllable by a relatively small base current." In an NPN transistor operating in the active region, the emitter-base junction is forward biased (electrons and holes recombine at the junction), and electrons are injected into the base region. Because the base is narrow, most of these electrons will diffuse into the reverse-biased (electrons and holes are formed at, and move away from the junction) base-collector junction and be swept into the collector; perhaps one-hundredth of the electrons will recombine in the base, which is the dominant mechanism in the base current. By controlling the number of electrons that can leave the base, the number of electrons entering the collector can be controlled. Collector current is approximately β (common-emitter current gain) times the base current. It is typically greater than 100 for small-signal transistors but can be smaller in transistors designed for high-power applications.

Unlike the FET, the BJT is a low–input-impedance device. Also, as the base–emitter voltage (V_{be}) is increased the base–emitter current and hence the collector–emitter current (I_{ce}) increase exponentially according to the Shockley diode model and the Ebers-Moll model. Because of this exponential relationship, the BJT has a higher transconductance than the FET.

Bipolar transistors can be made to conduct by exposure to light, since absorption of photons in the base region generates a photocurrent that acts as a base current; the collector current is approximately β times the photocurrent. Devices designed for this purpose have a transparent window in the package and are called phototransistors.

Field-effect transistor

The *field-effect transistor* (FET), sometimes called a *unipolar transistor*, uses either electrons (in *N-channel FET*) or holes (in *P-channel FET*) for conduction. The four terminals of the FET are named *source*, *gate*, *drain*, and *body* (*substrate*). On most FETs, the body is connected to the source inside the package, and this will be assumed for the following description.

In FETs, the drain-to-source current flows via a conducting channel that connects the *source* region to the *drain* region. The conductivity is varied by the electric field that is produced when a voltage is applied between the gate and source terminals; hence the current flowing between the drain and source is controlled by the voltage applied between the gate and source. As the gate–source voltage (V_{gs}) is increased, the drain–source current (I_{ds}) increases exponentially for V_{gs} below threshold, and then at a roughly quadratic rate ($I_{ds} \propto (V_{gs} - V_T)^2$) (where V_T is the threshold voltage at which drain

current begins) in the "space-charge-limited" region above threshold. A quadratic behavior is not observed in modern devices, for example, at the 65 nm technology node.

For low noise at narrow bandwidth the higher input resistance of the FET is advantageous.

FETs are divided into two families: *junction FET* (JFET) and *insulated gate FET* (IGFET). The IGFET is more commonly known as a *metal–oxide–semiconductor FET* (MOSFET), reflecting its original construction from layers of metal (the gate), oxide (the insulation), and semiconductor. Unlike IGFETs, the JFET gate forms a PN diode with the channel which lies between the source and drain. Functionally, this makes the N-channel JFET the solid state equivalent of the vacuum tube triode which, similarly, forms a diode between its grid and cathode. Also, both devices operate in the *depletion mode*, they both have a high input impedance, and they both conduct current under the control of an input voltage.

Metal–semiconductor FETs (MESFETs) are JFETs in which the reverse biased PN junction is replaced by a metal–semiconductor Schottky-junction. These, and the HEMTs (high electron mobility transistors, or HFETs), in which a two-dimensional electron gas with very high carrier mobility is used for charge transport, are especially suitable for use at very high frequencies (microwave frequencies; several GHz).

Unlike bipolar transistors, FETs do not inherently amplify a photocurrent. Nevertheless, there are ways to use them, especially JFETs, as light-sensitive devices, by exploiting the photocurrents in channel–gate or channel–body junctions.

FETs are further divided into *depletion-mode* and *enhancement-mode* types, depending on whether the channel is turned on or off with zero gate-to-source voltage. For enhancement mode, the channel is off at zero bias, and a gate potential can "enhance" the conduction. For depletion mode, the channel is on at zero bias, and a gate potential (of the opposite polarity) can "deplete" the channel, reducing conduction. For either mode, a more positive gate voltage corresponds to a higher current for N-channel devices and a lower current for P-channel devices. Nearly all JFETs are depletion-mode as the diode junctions would forward bias and conduct if they were enhancement mode devices; most IGFETs are enhancement-mode types.

Other transistor types

- Point-contact transistor, first kind of transistor ever constructed
- Bipolar junction transistor (BJT)
    - Heterojunction bipolar transistor, up to several hundred GHz, common in modern ultrafast and RF circuits
    - Grown-junction transistor, first kind of BJT
    - Alloy-junction transistor, improvement of grown-junction transistor
        - Micro-alloy transistor (MAT), speedier than alloy-junction transistor

        - Micro-alloy diffused transistor (MADT), speedier than MAT, a diffused-base transistor
        - Post-alloy diffused transistor (PADT), speedier than MAT, a diffused-base transistor
        - Schottky transistor
        - Surface barrier transistor
    - Drift-field transistor
    - Avalanche transistor
    - Darlington transistors are two BJTs connected together to provide a high current gain equal to the product of the current gains of the two transistors.
    - Insulated gate bipolar transistors (IGBTs) use a medium power IGFET, similarly connected to a power BJT, to give a high input impedance. Power diodes are often connected between certain terminals depending on specific use. IGBTs are particularly suitable for heavy-duty industrial applications. The Asea Brown Boveri (ABB) *5SNA2400E170100* illustrates just how far power semiconductor technology has advanced. Intended for three-phase power supplies, this device houses three NPN IGBTs in a case measuring 38 by 140 by 190 mm and weighing 1.5 kg. Each IGBT is rated at 1,700 volts and can handle 2,400 amperes.
    - Photo transistor
- Field-effect transistor
    - Carbon nanotube field-effect transistor (CNFET)
    - JFET, where the gate is insulated by a reverse-biased PN junction
    - MESFET, similar to JFET with a Schottky junction instead of PN one
        - High Electron Mobility Transistor (HEMT, HFET, MODFET)
    - MOSFET, where the gate is insulated by a shallow layer of insulator
    - Inverted-T field effect transistor (ITFET)
    - FinFET, source/drain region shapes fins on the silicon surface.
    - FREDFET, fast-reverse epitaxial diode field-effect transistor
    - Thin film transistor, in LCDs.
    - OFET Organic Field-Effect Transistor, in which the semiconductor is an organic compound
    - Ballistic transistor
    - Floating-gate transistor, for non-volatile storage.
    - FETs used to sense environment
        - Ion-sensitive field effect transistor, to measure ion concentrations in solution.
        - EOSFET, electrolyte-oxide-semiconductor field effect transistor (Neurochip)
        - DNAFET, deoxyribonucleic acid field-effect transistor
- Spacistor
- Diffusion transistor, formed by diffusing dopants into semiconductor substrate; can be both BJT and FET
- Unijunction transistors can be used as simple pulse generators. They comprise a main body of either P-type or N-type semiconductor with ohmic contacts at each end (terminals *Base1* and *Base2*). A junction with the opposite semiconductor

type is formed at a point along the length of the body for the third terminal (*Emitter*).

- Single-electron transistors (SET) consist of a gate island between two tunnelling junctions. The tunnelling current is controlled by a voltage applied to the gate through a capacitor.
- Nanofluidic transistor, controls the movement of ions through sub-microscopic, water-filled channels. Nanofluidic transistor, the basis of future chemical processors
- Multigate devices
  - Tetrode transistor
  - Pentode transistor
  - Multigate device
  - Trigate transistors (Prototype by Intel)
  - **Dual gate FETs** have a single channel with two gates in cascode; a configuration optimized for *high frequency amplifiers*, *mixers*, and oscillators.
- Junctionless Nanowire Transistor (JNT), developed at Tyndall National Institute in Ireland, was the first transistor successfully fabricated without junctions. (Even MOSFETs have junctions, although its gate is electrically insulated from the region the gate controls.) Junctions are difficult and expensive to fabricate, and, because they are a significant source of current leakage, they waste significant power and generate significant waste heat. Eliminating them held the promise of cheaper and denser microchips. The JNT uses a simple nanowire of silicon surrounded by an electrically isolated "wedding ring" that acts to gate the flow of electrons through the wire. This method has been described as akin to squeezing a garden hose to gate the flow of water through the hose. The nanowire is heavily n-doped, making it an excellent conductor. Crucially the gate, comprising silicon, is heavily p-doped; and its presence depletes the underlying silicon nanowire thereby preventing carrier flow past the gate.

Part numbers

The types of some transistors can be parsed from the part number. There are three major semiconductor naming standards; in each the alphanumeric prefix provides clues to type of the device:

Japanese Industrial Standard (JIS) has a standard for transistor part numbers. They begin with "2S", e.g. 2SD965, but sometimes the "2S" prefix is not marked on the package – a 2SD965 might only be marked "D965"; a 2SC1815 might be listed by a supplier as simply "C1815". This series sometimes has suffixes (such as "R", "O", "BL"... standing for "Red", "Orange", "Blue" etc.) to denote variants, such as tighter h_{FE} (gain) groupings.

Beginning of Part Number	Type of Transistor
2SA	high frequency PNP BJTs
2SB	audio frequency PNP BJTs

2SC	high frequency NPN BJTs
2SD	audio frequency NPN BJTs
2SJ	P-channel FETs (both JFETs and MOSFETs)
2SK	N-channel FETs (both JFETs and MOSFETs)

The Pro Electron part numbers begin with two letters: the first gives the semiconductor type (A for Germanium, B for Silicon, and C for materials like GaAs); the second letter denotes the intended use (A for diode, C for general-purpose transistor, etc.). A 3-digit sequence number (or one letter then 2 digits, for industrial types) follows (and, with early devices, indicated the case type – just as the older system for vacuum tubes used the last digit or two to indicate the number of pins, and the first digit or two for the filament voltage). Suffixes may be used, such as a letter (e.g. "C" often means high h_{FE}, such as in: BC549C) or other codes may follow to show gain (e.g. BC327-25) or voltage rating (e.g. BUK854-800A). The more common prefixes are:

Prefix class	Usage	Example
AC	Germanium small signal transistor	AC126
AF	Germanium RF transistor	AF117
BC	Silicon, small signal transistor ("allround")	BC548B
BD	Silicon, power transistor	BD139
BF	Silicon, RF (high frequency) BJT or FET	BF245
BS	Silicon, switching transistor (BJT or MOSFET)	BS170
BL	Silicon, high frequency, high power (for transmitters)	BLW34
BU	Silicon, high voltage (for CRT horizontal deflection circuits)	BU508

The JEDEC transistor device numbers usually start with 2N, indicating a three-terminal device (dual-gate field-effect transistors are four-terminal devices, so begin with 3N), then a 2, 3 or 4-digit sequential number with no significance as to device properties (although low numbers tend to be Germanium devices, because early transistors were mainly Germanium). For example 2N3055 is a silicon NPN power transistor, 2N1301 is a PNP germanium switching transistor. A letter suffix (such as "A") is sometimes used to indicate a newer variant, but rarely gain groupings.

Other schemes

Manufacturers of devices may have their own proprietary numbering system, for example CK722. Note that a manufacturer's prefix (like "MPF" in MPF102, which originally would denote a Motorola FET) now is an unreliable indicator of who made the device. Some proprietary naming schemes adopt parts of other naming schemes, for example a PN2222A is a (possibly Fairchild Semiconductor) 2N2222A in a plastic case (but a PN108 is a plastic version of a BC108, not a 2N108, while the PN100 is unrelated to other xx100 devices).

Military part numbers sometimes are assigned their own codes, such as the British Military CV Naming System.

Manufacturers buying large numbers of similar parts may have them supplied with "house numbers", identifying a particular purchasing specification and not necessarily a device with a standardized registered number. For example, an HP part 1854,0053 is a (JEDEC) 2N2218 transistor which is also assigned the CV number: CV7763

Naming problems

With so many independent naming schemes, and the abbreviation of part numbers when printed on the devices, ambiguity sometimes occurs. For example two different devices may be marked "J176" (one the J176 low-power Junction FET, the other the higher-powered MOSFET 2SJ176).

As older "through-hole" transistors are given Surface-Mount packaged counterparts, they tend to be assigned many different part numbers because manufacturers have their own systems to cope with the variety in pinout arrangements and options for dual or matched NPN+PNP devices in one pack. So even when the original device (such as a 2N3904) may have been assigned by a standards authority, and well known by engineers over the years, the new versions are far from standardised in their naming.

Construction

Semiconductor material

The first BJTs were made from germanium (Ge). Silicon (Si) types currently predominate but certain advanced microwave and high performance versions now employ the *compound semiconductor* material gallium arsenide (GaAs) and the *semiconductor alloy* silicon germanium (SiGe). Single element semiconductor material (Ge and Si) is described as *elemental*.

Rough parameters for the most common semiconductor materials used to make transistors are given in the table below; it must be noted that these parameters will vary with increase in temperature, electric field, impurity level, strain, and sundry other factors:

Semiconductor material characteristics

Semiconductor material	Junction forward voltage V @ 25 °C	Electron mobility $m^2/(V \cdot s)$ @ 25 °C	Hole mobility $m^2/(V \cdot s)$ @ 25 °C	Max. junction temp. °C
Ge	0.27	0.39	0.19	70 to 100
Si	0.71	0.14	0.05	150 to 200
GaAs	1.03	0.85	0.05	150 to 200
Al-Si junction	0.3	—	—	150 to 200

The *junction forward voltage* is the voltage applied to the emitter-base junction of a BJT in order to make the base conduct a specified current. The current increases exponentially as the junction forward voltage is increased. The values given in the table are typical for a current of 1 mA (the same values apply to semiconductor diodes). The lower the junction forward voltage the better, as this means that less power is required to "drive" the transistor. The junction forward voltage for a given current decreases with increase in temperature. For a typical silicon junction the change is −2.1 mV/°C. In some circuits special compensating elements (sensistors) must be used to compensate for such changes.

The density of mobile carriers in the channel of a MOSFET is a function of the electric field forming the channel and of various other phenomena such as the impurity level in the channel. Some impurities, called dopants, are introduced deliberately in making a MOSFET, to control the MOSFET electrical behavior.

The *electron mobility* and *hole mobility* columns show the average speed that electrons and holes diffuse through the semiconductor material with an electric field of 1 volt per meter applied across the material. In general, the higher the electron mobility the speedier the transistor. The table indicates that Ge is a better material than Si in this respect. However, Ge has four major shortcomings compared to silicon and gallium arsenide:

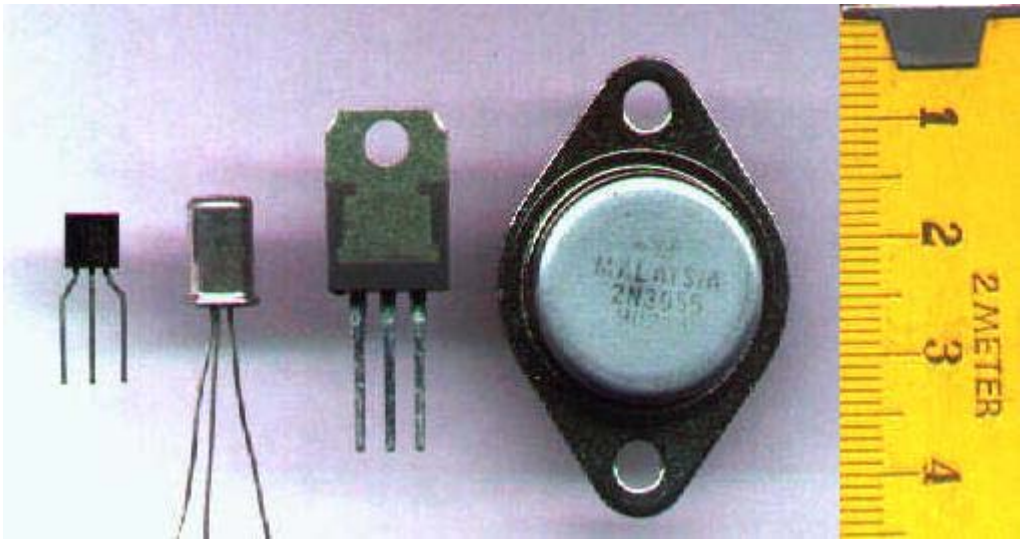
- Its maximum temperature is limited;
- it has relatively high leakage current;
- it cannot withstand high voltages;
- it is less suitable for fabricating integrated circuits.

Because the electron mobility is higher than the hole mobility for all semiconductor materials, a given bipolar NPN transistor tends to be swifter than an equivalent PNP transistor type. GaAs has the highest electron mobility of the three semiconductors. It is for this reason that GaAs is used in high frequency applications. A relatively recent FET development, the *high electron mobility transistor* (HEMT), has a heterostructure (junction between different semiconductor materials) of aluminium gallium arsenide (AlGaAs)-gallium arsenide (GaAs) which has twice the electron mobility of a GaAs-metal barrier junction. Because of their high speed and low noise, HEMTs are used in satellite receivers working at frequencies around 12 GHz.

Max. junction temperature values represent a cross section taken from various manufacturers' data sheets. This temperature should not be exceeded or the transistor may be damaged.

Al–Si junction refers to the high-speed (aluminum–silicon) semiconductor–metal barrier diode, commonly known as a Schottky diode. This is included in the table because some silicon power IGFETs have a *parasitic* reverse Schottky diode formed between the source and drain as part of the fabrication process. This diode can be a nuisance, but sometimes it is used in the circuit.

Packaging



Through-hole transistors (tape measure marked in centimetres)

Transistors come in many different packages (semiconductor packages). The two main categories are *through-hole* (or *leaded*), and *surface-mount*, also known as *surface mount device* (SMD). The *ball grid array* (BGA) is the latest surface mount package (currently only for large *transistor arrays*). It has solder "balls" on the underside in place of leads. Because they are smaller and have shorter interconnections, SMDs have better high frequency characteristics but lower power rating.

Transistor packages are made of glass, metal, ceramic, or plastic. The package often dictates the power rating and frequency characteristics. Power transistors have larger packages that can be clamped to heat sinks for enhanced cooling. Additionally, most power transistors have the collector or drain physically connected to the metal can/metal plate. At the other extreme, some surface-mount *microwave* transistors are as small as grains of sand.

Often a given transistor type is available in sundry packages. Transistor packages are mainly standardized, but the assignment of a transistor's functions to the terminals is not: other transistor types can assign other functions to the package's terminals. Even for the same transistor type the terminal assignment can vary (normally indicated by a suffix letter to the part number, q.e. BC212L and BC212K).

Chapter-4

Capacitor

Capacitor

Modern capacitors, by a cm rule

Type Passive

Invented Ewald Georg von Kleist (October 1745)

Electronic symbol

A typical electrolytic capacitor

A **capacitor** (formerly known as **condenser**) is a device for storing electric charge. The forms of practical capacitors vary widely, but all contain at least two conductors separated by a non-conductor. Capacitors used as parts of electrical systems, for example, consist of metal foils separated by a layer of insulating film.

Capacitors are widely used in electronic circuits for blocking direct current while allowing alternating current to pass, in filter networks, for smoothing the output of power supplies, in the resonant circuits that tune radios to particular frequencies and for many other purposes.

A capacitor is a passive electronic component consisting of a pair of conductors separated by a dielectric (insulator). When there is a potential difference (voltage) across the conductors, a static electric field develops in the dielectric that stores energy and produces a mechanical force between the conductors. An ideal capacitor is characterized by a single constant value, capacitance, measured in farads. This is the ratio of the electric charge on each conductor to the potential difference between them.

The capacitance is greatest when there is a narrow separation between large areas of conductor, hence capacitor conductors are often called "plates", referring to an early

means of construction. In practice the dielectric between the plates passes a small amount of leakage current and also has an electric field strength limit, resulting in a breakdown voltage, while the conductors and leads introduce an undesired inductance and resistance.

History

Battery of four Leyden jars in Museum Boerhaave, Leiden, the Netherlands.

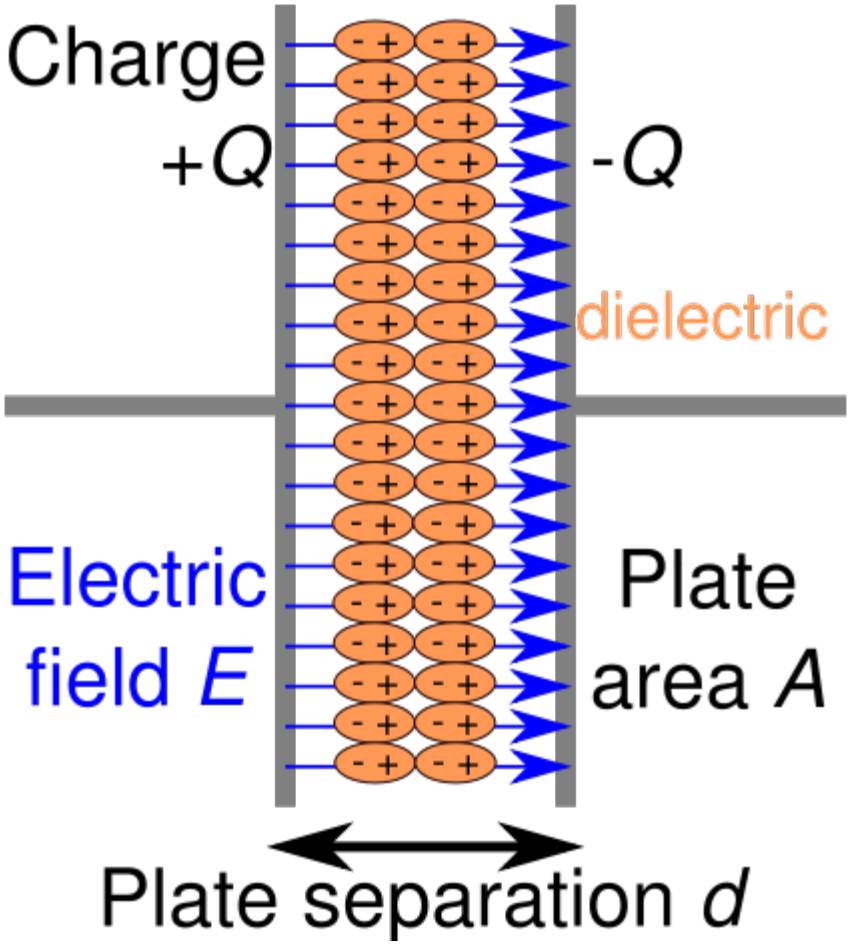
In October 1745, Ewald Georg von Kleist of Pomerania in Germany found that charge could be stored by connecting a high voltage electrostatic generator by a wire to a volume of water in a hand-held glass jar. Von Kleist's hand and the water acted as conductors and the jar as a dielectric (although details of the mechanism were incorrectly identified at the time). Von Kleist found, after removing the generator, that touching the wire resulted in a painful spark. In a letter describing the experiment, he said "I would not take a second shock for the kingdom of France." The following year, the Dutch physicist Pieter van Musschenbroek invented a similar capacitor, which was named the Leyden jar, after the University of Leiden where he worked.

Daniel Gralath was the first to combine several jars in parallel into a "battery" to increase the charge storage capacity. Benjamin Franklin investigated the Leyden jar and "proved" that the charge was stored on the glass, not in the water as others had assumed. He also adopted the term "battery", (denoting the increasing of power with a row of similar units as in a battery of cannon), subsequently applied to clusters of electrochemical cells. Leyden jars were later made by coating the inside and outside of jars with metal foil, leaving a space at the mouth to prevent arcing between the foils. The earliest unit of capacitance was the 'jar', equivalent to about 1 nanofarad.

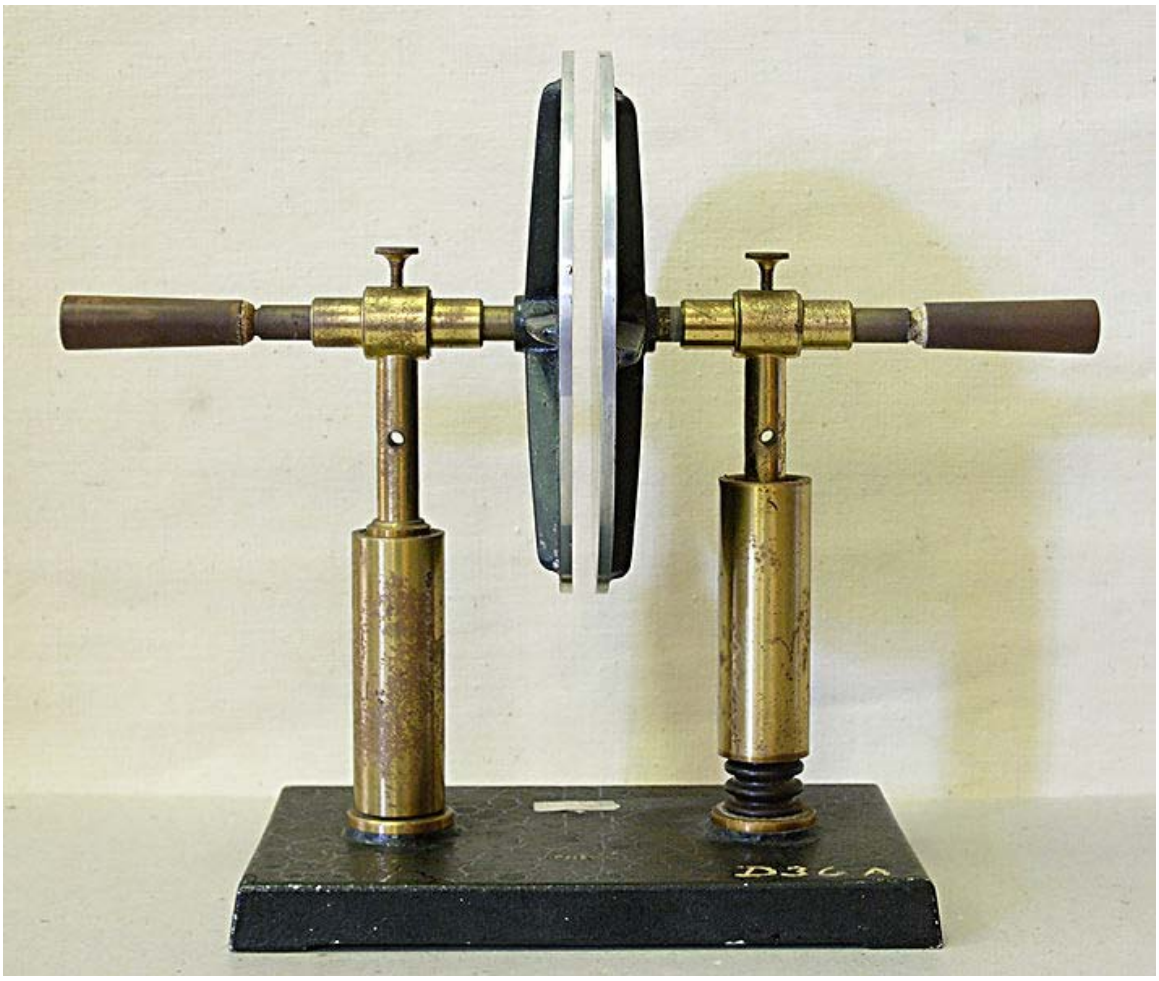
Leyden jars or more powerful devices employing flat glass plates alternating with foil conductors were used exclusively up until about 1900, when the invention of wireless (radio) created a demand for standard capacitors, and the steady move to higher frequencies required capacitors with lower inductance. A more compact construction began to be used of a flexible dielectric sheet such as oiled paper sandwiched between sheets of metal foil, rolled or folded into a small package.

Early capacitors were also known as *condensers*, a term that is still occasionally used today. The term was first used for this purpose by Alessandro Volta in 1782, with reference to the device's ability to store a higher density of electric charge than a normal isolated conductor.

Theory of operation



Charge separation in a parallel-plate capacitor causes an internal electric field. A dielectric (orange) reduces the field and increases the capacitance.



A simple demonstration of a parallel-plate capacitor

A capacitor consists of two conductors separated by a non-conductive region. The non-conductive region is called the dielectric or sometimes the dielectric medium. In simpler terms, the dielectric is just an electrical insulator. Examples of dielectric mediums are glass, air, paper, vacuum, and even a semiconductor depletion region chemically identical to the conductors. A capacitor is assumed to be self-contained and isolated, with no net electric charge and no influence from any external electric field. The conductors thus hold equal and opposite charges on their facing surfaces, and the dielectric develops an electric field. In SI units, a capacitance of one farad means that one coulomb of charge on each conductor causes a voltage of one volt across the device.

The capacitor is a reasonably general model for electric fields within electric circuits. An ideal capacitor is wholly characterized by a constant capacitance C, defined as the ratio of charge $\pm Q$ on each conductor to the voltage V between them:

$$C = \frac{Q}{V}$$

Sometimes charge build-up affects the capacitor mechanically, causing its capacitance to vary. In this case, capacitance is defined in terms of incremental changes:

$$C = \frac{\mathrm{d}q}{\mathrm{d}v}$$

Energy storage

Work must be done by an external influence to "move" charge between the conductors in a capacitor. When the external influence is removed the charge separation persists in the electric field and energy is stored to be released when the charge is allowed to return to its equilibrium position. The work done in establishing the electric field, and hence the amount of energy stored, is given by:

$$W = \int_{q=0}^{Q} V\,\mathrm{d}q = \int_{q=0}^{Q} \frac{q}{C}\mathrm{d}q = \frac{1}{2}\frac{Q^2}{C} = \frac{1}{2}CV^2 = \frac{1}{2}VQ.$$

Current-voltage relation

The current $i(t)$ through any component in an electric circuit is defined as the rate of flow of a charge $q(t)$ passing through it, but actual charges, electrons, cannot pass through the dielectric layer of a capacitor, rather an electron accumulates on the negative plate for each one that leaves the positive plate, resulting in an electron depletion and consequent positive charge on one electrode that is equal and opposite to the accumulated negative charge on the other. Thus the charge on the electrodes is equal to the integral of the current as well as proportional to the voltage as discussed above. As with any antiderivative, a constant of integration is added to represent the initial voltage $v\,(t_0)$. This is the integral form of the capacitor equation,

$$v(t) = \frac{q(t)}{C} = \frac{1}{C}\int_{t_0}^{t} i(\tau)\mathrm{d}\tau + v(t_0)$$

.

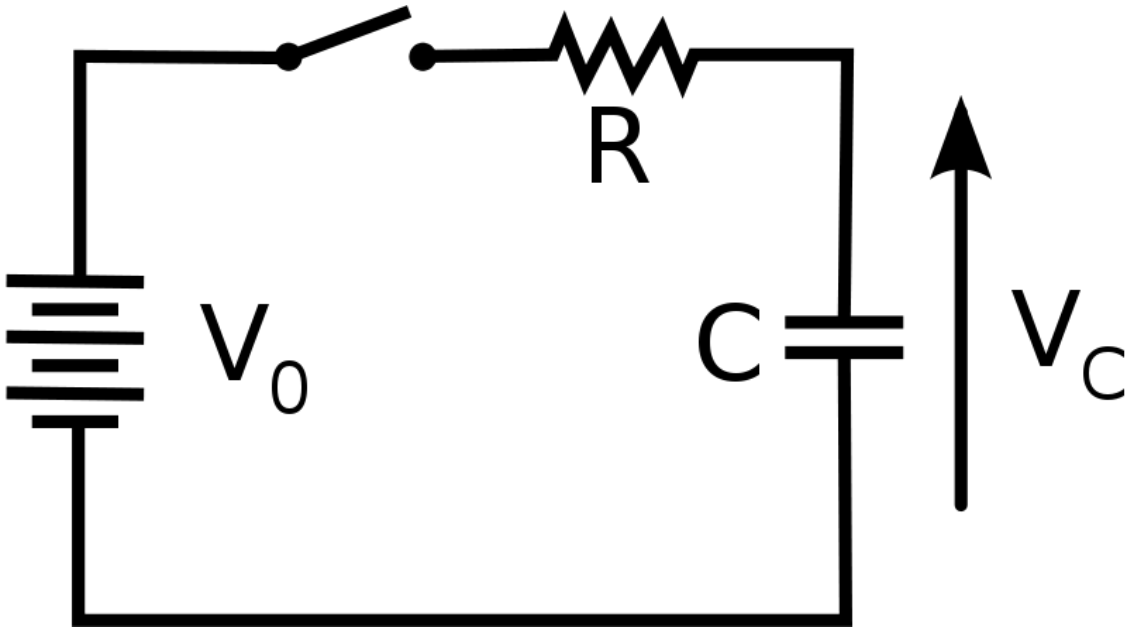
Taking the derivative of this, and multiplying by C, yields the derivative form,

$$i(t) = \frac{\mathrm{d}q(t)}{\mathrm{d}t} = C\frac{\mathrm{d}v(t)}{\mathrm{d}t}$$

.

The dual of the capacitor is the inductor, which stores energy in the magnetic field rather than the electric field. Its current-voltage relation is obtained by exchanging current and voltage in the capacitor equations and replacing C with the inductance L.

DC circuits



A simple resistor-capacitor circuit demonstrates charging of a capacitor.

A series circuit containing only a resistor, a capacitor, a switch and a constant DC source of voltage V_0 is known as a *charging circuit*. If the capacitor is initially uncharged while the switch is open, and the switch is closed at $t = 0$, it follows from Kirchhoff's voltage law that

$$V_0 = v_{\text{resistor}}(t) + v_{\text{capacitor}}(t) = i(t)R + \frac{1}{C}\int_0^t i(\tau)\mathrm{d}\tau.$$

Taking the derivative and multiplying by C, gives a first-order differential equation,

$$RC\frac{\mathrm{d}i(t)}{\mathrm{d}t} + i(t) = 0.$$

At $t = 0$, the voltage across the capacitor is zero and the voltage across the resistor is V_0. The initial current is then $i\,(0) = V_0\,/R$. With this assumption, the differential equation yields

$$i(t) = \frac{V_0}{R}e^{-t/\tau_0}$$
$$v(t) = V_0\left(1 - e^{-t/\tau_0}\right),$$

where $\tau_0 = RC$ is the *time constant* of the system.

As the capacitor reaches equilibrium with the source voltage, the voltage across the resistor and the current through the entire circuit decay exponentially. The case of *discharging* a charged capacitor likewise demonstrates exponential decay, but with the initial capacitor voltage replacing V_0 and the final voltage being zero.

AC circuits

Impedance, the vector sum of reactance and resistance, describes the phase difference and the ratio of amplitudes between sinusoidally varying voltage and sinusoidally varying current at a given frequency. Fourier analysis allows any signal to be constructed from a spectrum of frequencies, whence the circuit's reaction to the various frequencies may be found. The reactance and impedance of a capacitor are respectively

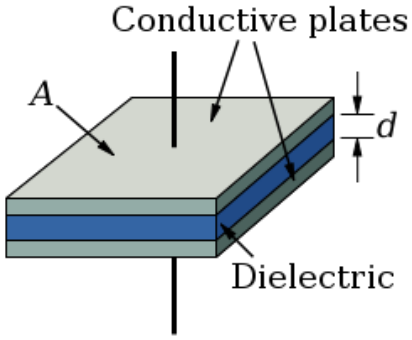
$$X = -\frac{1}{\omega C} = -\frac{1}{2\pi f C}$$
$$Z = \frac{1}{j\omega C} = -\frac{j}{\omega C} = -\frac{j}{2\pi f C}$$

where j is the imaginary unit and ω is the angular velocity of the sinusoidal signal. The $-j$ phase indicates that the AC voltage $V = Z\,I$ lags the AC current by 90°: the positive current phase corresponds to increasing voltage as the capacitor charges; zero current corresponds to instantaneous constant voltage, etc.

Note that impedance decreases with increasing capacitance and increasing frequency. This implies that a higher-frequency signal or a larger capacitor results in a lower voltage amplitude per current amplitude—an AC "short circuit" or AC coupling. Conversely, for very low frequencies, the reactance will be high, so that a capacitor is nearly an open circuit in AC analysis—those frequencies have been "filtered out".

Capacitors are different from resistors and inductors in that the impedance is *inversely* proportional to the defining characteristic, i.e. capacitance.

Parallel plate model



Dielectric is placed between two conducting plates, each of area A and with a separation of d.

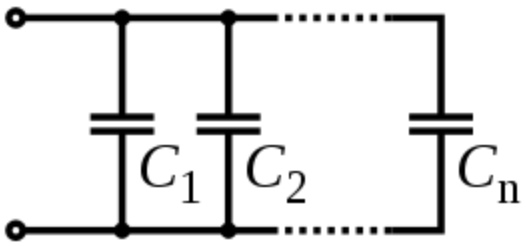
The simplest capacitor consists of two parallel conductive plates separated by a dielectric with permittivity ε (such as air). The model may also be used to make qualitative predictions for other device geometries. The plates are considered to extend uniformly over an area A and a charge density $\pm\rho = \pm Q/A$ exists on their surface. Assuming that the width of the plates is much greater than their separation d, the electric field near the centre of the device will be uniform with the magnitude $E = \rho/\varepsilon$. The voltage is defined as the line integral of the electric field between the plates

$$V = \int_0^d E\mathrm{d}z = \int_0^d \frac{\rho}{\varepsilon}\mathrm{d}z = \frac{\rho d}{\varepsilon} = \frac{Qd}{\varepsilon A}.$$

Solving this for $C = Q/V$ reveals that capacitance increases with area and decreases with separation

$$C = \frac{\varepsilon A}{d}.$$

The capacitance is therefore greatest in devices made from materials with a high permittivity.



Several capacitors in parallel.

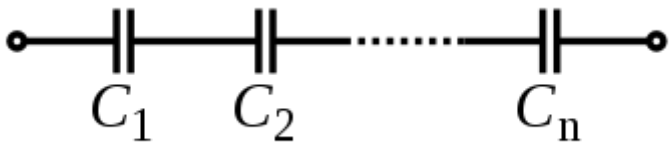
Networks

For capacitors in parallel

Capacitors in a parallel configuration each have the same applied voltage. Their capacitances add up. Charge is apportioned among them by size. Using the schematic diagram to visualize parallel plates, it is apparent that each capacitor contributes to the total surface area.

$$C_{eq} = C_1 + C_2 + \cdots + C_n$$

For capacitors in series



Several capacitors in series.

Connected in series, the schematic diagram reveals that the separation distance, not the plate area, adds up. The capacitors each store instantaneous charge build-up equal to that of every other capacitor in the series. The total voltage difference from end to end is apportioned to each capacitor according to the inverse of its capacitance. The entire series acts as a capacitor *smaller* than any of its components.

$$\frac{1}{C_{eq}} = \frac{1}{C_1} + \frac{1}{C_2} + \cdots + \frac{1}{C_n}$$

Capacitors are combined in series to achieve a higher working voltage, for example for smoothing a high voltage power supply. The voltage ratings, which are based on plate separation, add up. In such an application, several series connections may in turn be connected in parallel, forming a matrix. The goal is to maximize the energy storage utility of each capacitor without overloading it.
Series connection is also used to adapt electrolytic capacitors for AC use.

Non-ideal behaviour

Capacitors deviate from the ideal capacitor equation in a number of ways. Some of these, such as leakage current and parasitic effects are linear, or can be assumed to be linear, and can be dealt with by adding virtual components to the equivalent circuit of the capacitor. The usual methods of network analysis can then be applied. In other cases, such as with breakdown voltage, the effect is non-linear and normal (i.e., linear) network analysis cannot be used, the effect must be dealt with separately. There is yet another group, which may be linear but invalidate the assumption in the analysis that capacitance is a constant. Such an example is temperature dependence.

Breakdown voltage

Above a particular electric field, known as the dielectric strength E_{ds}, the dielectric in a capacitor becomes conductive. The voltage at which this occurs is called the breakdown voltage of the device, and is given by the product of the dielectric strength and the separation between the conductors,

$$V_{\mathrm{bd}} = E_{\mathrm{ds}} d$$

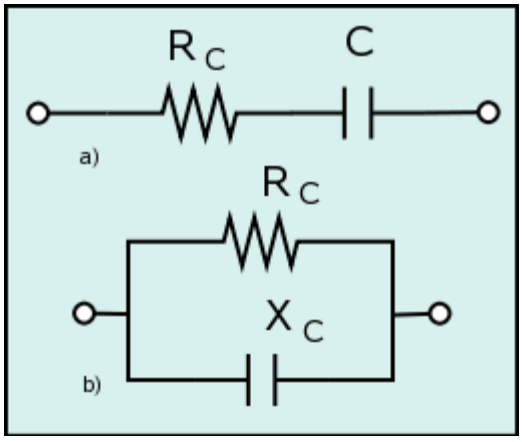
The maximum energy that can be stored safely in a capacitor is limited by the breakdown voltage. Due to the scaling of capacitance and breakdown voltage with dielectric

thickness, all capacitors made with a particular dielectric have approximately equal maximum energy density, to the extent that the dielectric dominates their volume.

For air dielectric capacitors the breakdown field strength is of the order 2 to 5 MV/m; for mica the breakdown is 100 to 300 MV/m, for oil 15 to 25 MV/m, and can be much less when other materials are used for the dielectric. The dielectric is used in very thin layers and so absolute breakdown voltage of capacitors is limited. Typical ratings for capacitors used for general electronics applications range from a few volts to 100V or so. As the voltage increases, the dielectric must be thicker, making high-voltage capacitors larger than those rated for lower voltages. The breakdown voltage is critically affected by factors such as the geometry of the capacitor conductive parts; sharp edges or points increase the electric field strength at that point and can lead to a local breakdown. Once this starts to happen, the breakdown will quickly "track" through the dielectric till it reaches the opposite plate and cause a short circuit.

The usual breakdown route is that the field strength becomes large enough to pull electrons in the dielectric from their atoms thus causing conduction. Other scenarios are possible, such as impurities in the dielectric, and, if the dielectric is of a crystalline nature, imperfections in the crystal structure can result in an avalanche breakdown as seen in semi-conductor devices. Breakdown voltage is also affected by pressure, humidity and temperature.

Equivalent circuit



Two different circuit models of a real capacitor

An ideal capacitor only stores and releases electrical energy, without dissipating any. In reality, all capacitors have imperfections within the capacitor's material that create resistance. This is specified as the *equivalent series resistance* or **ESR** of a component. This adds a real component to the impedance:

$$R_{\mathrm{C}} = Z + R_{\mathrm{ESR}} = \frac{1}{j\omega C} + R_{\mathrm{ESR}}$$

As frequency approaches infinity, the capacitive impedance (or reactance) approaches zero and the ESR becomes significant. As the reactance becomes negligible, power dissipation approaches $P_{RMS} = V_{RMS}^2 / R_{ESR}$.

Similarly to ESR, the capacitor's leads add *equivalent series inductance* or **ESL** to the component. This is usually significant only at relatively high frequencies. As inductive reactance is positive and increases with frequency, above a certain frequency capacitance will be canceled by inductance. High-frequency engineering involves accounting for the inductance of all connections and components.

If the conductors are separated by a material with a small conductivity rather than a perfect dielectric, then a small leakage current flows directly between them. The capacitor therefore has a finite parallel resistance, and slowly discharges over time (time may vary greatly depending on the capacitor material and quality).

Ripple current

Ripple current is the AC component of an applied source (often a switched-mode power supply) whose frequency may be constant or varying. Certain types of capacitors, such as electrolytic tantalum capacitors, usually have a rating for maximum ripple current (both in frequency and magnitude). This ripple current can cause damaging heat to be generated within the capacitor due to the current flow across resistive imperfections in the materials used within the capacitor, more commonly referred to as equivalent series resistance (ESR). For example electrolytic tantalum capacitors are limited by ripple current and generally have the highest ESR ratings in the capacitor family, while ceramic capacitors generally have no ripple current limitation and have some of the lowest ESR ratings.

Capacitance instability

The capacitance of certain capacitors decreases as the component ages. In ceramic capacitors, this is caused by degradation of the dielectric. The type of dielectric and the ambient operating and storage temperatures are the most significant aging factors, while the operating voltage has a smaller effect. The aging process may be reversed by heating the component above the Curie point. Aging is fastest near the beginning of life of the component, and the device stabilizes over time. Electrolytic capacitors age as the electrolyte evaporates. In contrast with ceramic capacitors, this occurs towards the end of life of the component.

Temperature dependence of capacitance is usually expressed in parts per million (ppm) per °C. It can usually be taken as a broadly linear function but can be noticeably non-linear at the temperature extremes. The temperature coefficient can be either positive or negative, sometimes even amongst different samples of the same type. In other words, the spread in the range of temperature coefficients can encompass zero.

Capacitors, especially ceramic capacitors, and older designs such as paper capacitors, can absorb sound waves resulting in a microphonic effect. Vibration moves the plates, causing the capacitance to vary, in turn inducing AC current. Some dielectrics also generate piezoelectricity. The resulting interference is especially problematic in audio applications, potentially causing feedback or unintended recording. In the reverse microphonic effect, the varying electric field between the capacitor plates exerts a physical force, moving them as a speaker. This can generate audible sound, but drains energy and stresses the dielectric and the electrolyte, if any.

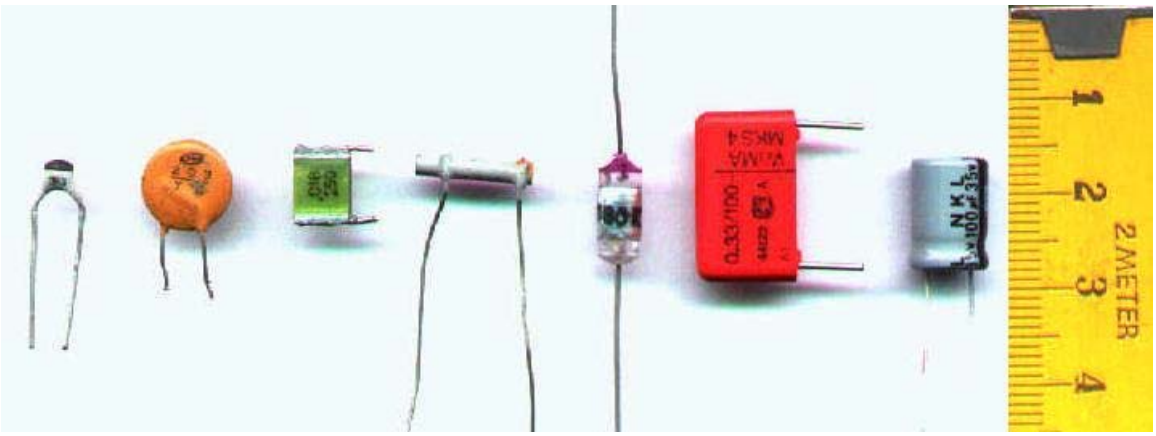
Capacitor types

Practical capacitors are available commercially in many different forms. The type of internal dielectric, the structure of the plates and the device packaging all strongly affect the characteristics of the capacitor, and its applications.

Values available range from very low (picofarad range; while arbitrarily low values are in principle possible, stray (parasitic) capacitance in any circuit is the limiting factor) to about 5 kF supercapacitors.

Above approximately 1 microfarad electrolytic capacitors are usually used because of their small size and low cost compared with other technologies, unless their relatively poor stability, life and polarised nature make them unsuitable. Very high capacity supercapacitors use a porous carbon-based electrode material.

Dielectric materials



Capacitor materials. From left: multilayer ceramic, ceramic disc, multilayer polyester film, tubular ceramic, polystyrene, metalized polyester film, aluminum electrolytic. Major scale divisions are in centimetres.

Most types of capacitor include a dielectric spacer, which increases their capacitance. These dielectrics are most often insulators. However, low capacitance devices are available with a vacuum between their plates, which allows extremely high voltage

operation and low losses. Variable capacitors with their plates open to the atmosphere were commonly used in radio tuning circuits. Later designs use polymer foil dielectric between the moving and stationary plates, with no significant air space between them.

In order to maximise the charge that a capacitor can hold, the dielectric material needs to have as high a permittivity as possible, while also having as high a breakdown voltage as possible.

Several solid dielectrics are available, including paper, plastic, glass, mica and ceramic materials. Paper was used extensively in older devices and offers relatively high voltage performance. However, it is susceptible to water absorption, and has been largely replaced by plastic film capacitors. Plastics offer better stability and aging performance, which makes them useful in timer circuits, although they may be limited to low operating temperatures and frequencies. Ceramic capacitors are generally small, cheap and useful for high frequency applications, although their capacitance varies strongly with voltage and they age poorly. They are broadly categorized as class 1 dielectrics, which have predictable variation of capacitance with temperature or class 2 dielectrics, which can operate at higher voltage. Glass and mica capacitors are extremely reliable, stable and tolerant to high temperatures and voltages, but are too expensive for most mainstream applications. Electrolytic capacitors and supercapacitors are used to store small and larger amounts of energy, respectively, ceramic capacitors are often used in resonators, and parasitic capacitance occurs in circuits wherever the simple conductor-insulator-conductor structure is formed unintentionally by the configuration of the circuit layout.

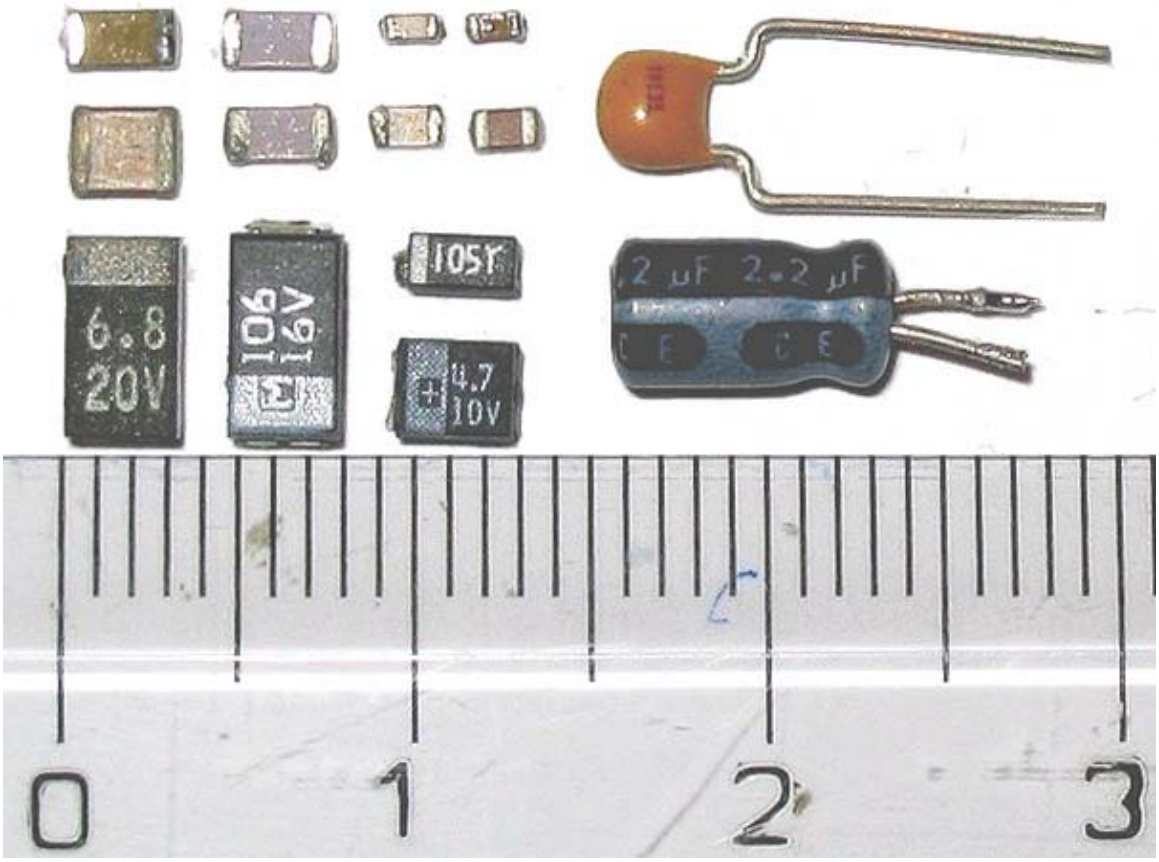
Electrolytic capacitors use an aluminum or tantalum plate with an oxide dielectric layer. The second electrode is a liquid electrolyte, connected to the circuit by another foil plate. Electrolytic capacitors offer very high capacitance but suffer from poor tolerances, high instability, gradual loss of capacitance especially when subjected to heat, and high leakage current. Poor quality capacitors may leak electrolyte, which is harmful to printed circuit boards. The conductivity of the electrolyte drops at low temperatures, which increases equivalent series resistance. While widely used for power-supply conditioning, poor high-frequency characteristics make them unsuitable for many applications. Electrolytic capacitors will self-degrade if unused for a period (around a year), and when full power is applied may short circuit, permanently damaging the capacitor and usually blowing a fuse or causing arcing in rectifier tubes. They can be restored before use (and damage) by gradually applying the operating voltage, often done on antique vacuum tube equipment over a period of 30 minutes by using a variable transformer to supply AC power. Unfortunately, the use of this technique may be less satisfactory for some solid state equipment, which may be damaged by operation below its normal power range, requiring that the power supply first be isolated from the consuming circuits. Such remedies may not be applicable to modern high-frequency power supplies as these produce full output voltage even with reduced input.

Tantalum capacitors offer better frequency and temperature characteristics than aluminum, but higher dielectric absorption and leakage. OS-CON (or OC-CON)

capacitors are a polymerized organic semiconductor solid-electrolyte type that offer longer life at higher cost than standard electrolytic capacitors.

Several other types of capacitor are available for specialist applications. Supercapacitors store large amounts of energy. Supercapacitors made from carbon aerogel, carbon nanotubes, or highly porous electrode materials offer extremely high capacitance (up to 5 kF as of 2010) and can be used in some applications instead of rechargeable batteries. Alternating current capacitors are specifically designed to work on line (mains) voltage AC power circuits. They are commonly used in electric motor circuits and are often designed to handle large currents, so they tend to be physically large. They are usually ruggedly packaged, often in metal cases that can be easily grounded/earthed. They also are designed with direct current breakdown voltages of at least five times the maximum AC voltage.

Structure



Capacitor packages: SMD ceramic at top left; SMD tantalum at bottom left; through-hole tantalum at top right; through-hole electrolytic at bottom right. Major scale divisions are cm.

The arrangement of plates and dielectric has many variations depending on the desired ratings of the capacitor. For small values of capacitance (microfarads and less), ceramic

disks use metallic coatings, with wire leads bonded to the coating. Larger values can be made by multiple stacks of plates and disks. Larger value capacitors usually use a metal foil or metal film layer deposited on the surface of a dielectric film to make the plates, and a dielectric film of impregnated paper or plastic – these are rolled up to save space. To reduce the series resistance and inductance for long plates, the plates and dielectric are staggered so that connection is made at the common edge of the rolled-up plates, not at the ends of the foil or metalized film strips that comprise the plates.

The assembly is encased to prevent moisture entering the dielectric – early radio equipment used a cardboard tube sealed with wax. Modern paper or film dielectric capacitors are dipped in a hard thermoplastic. Large capacitors for high-voltage use may have the roll form compressed to fit into a rectangular metal case, with bolted terminals and bushings for connections. The dielectric in larger capacitors is often impregnated with a liquid to improve its properties.

Capacitors may have their connecting leads arranged in many configurations, for example axially or radially. "Axial" means that the leads are on a common axis, typically the axis of the capacitor's cylindrical body – the leads extend from opposite ends. Radial leads might more accurately be referred to as tandem; they are rarely actually aligned along radii of the body's circle, so the term is inexact, although universal. The leads (until bent) are usually in planes parallel to that of the flat body of the capacitor, and extend in the same direction; they are often parallel as manufactured.

Small, cheap discoidal ceramic capacitors have existed since the 1930s, and remain in widespread use. Since the 1980s, surface mount packages for capacitors have been widely used. These packages are extremely small and lack connecting leads, allowing them to be soldered directly onto the surface of printed circuit boards. Surface mount components avoid undesirable high-frequency effects due to the leads and simplify automated assembly, although manual handling is made difficult due to their small size.

Mechanically controlled variable capacitors allow the plate spacing to be adjusted, for example by rotating or sliding a set of movable plates into alignment with a set of stationary plates. Low cost variable capacitors squeeze together alternating layers of aluminum and plastic with a screw. Electrical control of capacitance is achievable with varactors (or varicaps), which are reverse-biased semiconductor diodes whose depletion region width varies with applied voltage. They are used in phase-locked loops, amongst other applications.

Capacitor markings

Most capacitors have numbers printed on their bodies to indicate their electrical characteristics. Larger capacitors like electrolytics usually display the actual capacitance together with the unit (for example, **220 µF**). Smaller capacitors like ceramics, however, use a shorthand consisting of three numbers and a letter, where the numbers show the capacitance in pF (calculated as XY x 10^Z for the numbers XYZ) and the letter indicates the tolerance (J, K or M for ±5%, ±10% and ±20% respectively).

Additionally, the capacitor may show its working voltage, temperature and other relevant characteristics.

Example

A capacitor with the text **473K 330V** on its body has a capacitance of 47 x 10^3 pF = 47 nF (±10%) with a working voltage of 330 V.

Applications

Capacitors have many uses in electronic and electrical systems. They are so common that it is a rare electrical product that does not include at least one for some purpose.

Energy storage

A capacitor can store electric energy when disconnected from its charging circuit, so it can be used like a temporary battery. Capacitors are commonly used in electronic devices to maintain power supply while batteries are being changed. (This prevents loss of information in volatile memory.)

Conventional capacitors provide less than 360 joules per kilogram of energy density, while capacitors using developing technologies could provide more than 2.52 kilojoules per kilogram.

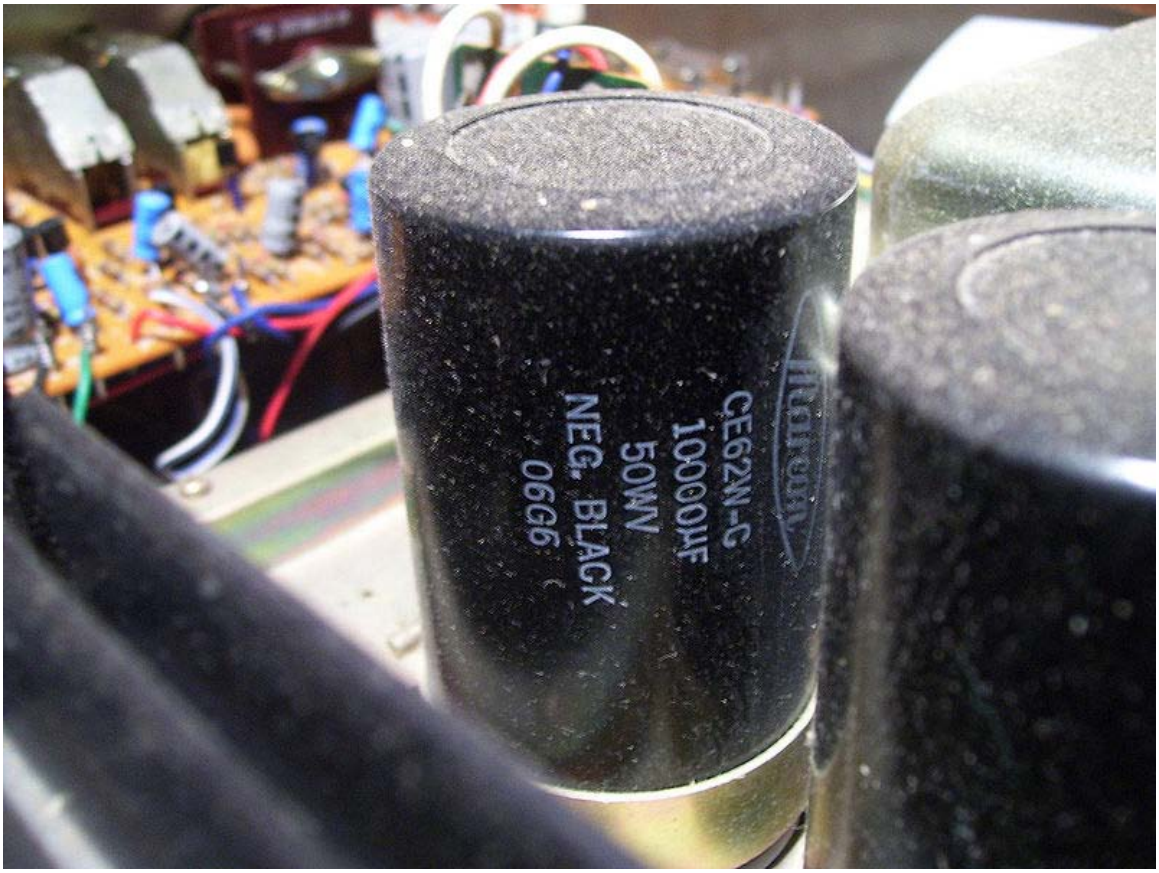
In car audio systems, large capacitors store energy for the amplifier to use on demand. Also for a flash tube a capacitor is used to hold the high voltage.

Pulsed power and weapons

Groups of large, specially constructed, low-inductance high-voltage capacitors (*capacitor banks*) are used to supply huge pulses of current for many pulsed power applications. These include electromagnetic forming, Marx generators, pulsed lasers (especially TEA lasers), pulse forming networks, radar, fusion research, and particle accelerators.

Large capacitor banks (reservoir) are used as energy sources for the exploding-bridgewire detonators or slapper detonators in nuclear weapons and other specialty weapons. Experimental work is under way using banks of capacitors as power sources for electromagnetic armour and electromagnetic railguns and coilguns.

Power conditioning



A 10,000 microfarad capacitor in a TRM-800 amplifier

Reservoir capacitors are used in power supplies where they smooth the output of a full or half wave rectifier. They can also be used in charge pump circuits as the energy storage element in the generation of higher voltages than the input voltage.

Capacitors are connected in parallel with the power circuits of most electronic devices and larger systems (such as factories) to shunt away and conceal current fluctuations from the primary power source to provide a "clean" power supply for signal or control circuits. Audio equipment, for example, uses several capacitors in this way, to shunt away power line hum before it gets into the signal circuitry. The capacitors act as a local reserve for the DC power source, and bypass AC currents from the power supply. This is used in car audio applications, when a stiffening capacitor compensates for the inductance and resistance of the leads to the lead-acid car battery.

Power factor correction

In electric power distribution, capacitors are used for power factor correction. Such capacitors often come as three capacitors connected as a three phase load. Usually, the values of these capacitors are given not in farads but rather as a reactive power in volt-

amperes reactive (VAr). The purpose is to counteract inductive loading from devices like electric motors and transmission lines to make the load appear to be mostly resistive. Individual motor or lamp loads may have capacitors for power factor correction, or larger sets of capacitors (usually with automatic switching devices) may be installed at a load center within a building or in a large utility substation.

Supression and coupling

Signal coupling

Because capacitors pass AC but block DC signals (when charged up to the applied dc voltage), they are often used to separate the AC and DC components of a signal. This method is known as *AC coupling* or "capacitive coupling". Here, a large value of capacitance, whose value need not be accurately controlled, but whose reactance is small at the signal frequency, is employed.

Decoupling

A decoupling capacitor is a capacitor used to protect one part of a circuit from the effect of another, for instance to suppress noise or transients. Noise caused by other circuit elements is shunted through the capacitor, reducing the effect they have on the rest of the circuit. It is most commonly used between the power supply and ground. An alternative name is *bypass capacitor* as it is used to bypass the power supply or other high impedance component of a circuit.

Noise filters and snubbers

When an inductive circuit is opened, the current through the inductance collapses quickly, creating a large voltage across the open circuit of the switch or relay. If the inductance is large enough, the energy will generate a spark, causing the contact points to oxidize, deteriorate, or sometimes weld together, or destroying a solid-state switch. A snubber capacitor across the newly opened circuit creates a path for this impulse to bypass the contact points, thereby preserving their life; these were commonly found in contact breaker ignition systems, for instance. Similarly, in smaller scale circuits, the spark may not be enough to damage the switch but will still radiate undesirable radio frequency interference (RFI), which a filter capacitor absorbs. Snubber capacitors are usually employed with a low-value resistor in series, to dissipate energy and minimize RFI. Such resistor-capacitor combinations are available in a single package.

Capacitors are also used in parallel to interrupt units of a high-voltage circuit breaker in order to equally distribute the voltage between these units. In this case they are called grading capacitors.

In schematic diagrams, a capacitor used primarily for DC charge storage is often drawn vertically in circuit diagrams with the lower, more negative, plate drawn as an arc. The straight plate indicates the positive terminal of the device, if it is polarized.

Motor starters

In single phase squirrel cage motors, the primary winding within the motor housing is not capable of starting a rotational motion on the rotor, but is capable of sustaining one. To start the motor, a secondary winding is used in series with a non-polarized *starting capacitor* to introduce a lag in the sinusoidal current through the starting winding. When the secondary winding is placed at an angle with respect to the primary winding, a rotating electric field is created. The force of the rotational field is not constant, but is sufficient to start the rotor spinning. When the rotor comes close to operating speed, a centrifugal switch (or current-sensitive relay in series with the main winding) disconnects the capacitor. The start capacitor is typically mounted to the side of the motor housing. These are called capacitor-start motors, that have relatively high starting torque.

There are also capacitor-run induction motors which have a permanently connected phase-shifting capacitor in series with a second winding. The motor is much like a two-phase induction motor.

Motor-starting capacitors are typically non-polarized electrolytic types, while running capacitors are conventional paper or plastic film dielectric types.

Signal processing

The energy stored in a capacitor can be used to represent information, either in binary form, as in DRAMs, or in analogue form, as in analog sampled filters and CCDs. Capacitors can be used in analog circuits as components of integrators or more complex filters and in negative feedback loop stabilization. Signal processing circuits also use capacitors to integrate a current signal.

Tuned circuits

Capacitors and inductors are applied together in tuned circuits to select information in particular frequency bands. For example, radio receivers rely on variable capacitors to tune the station frequency. Speakers use passive analog crossovers, and analog equalizers use capacitors to select different audio bands.

The resonant frequency *f* of a tuned circuit is a function of the inductance (*L*) and capacitance (*C*) in series, and is given by:

$$f = \frac{1}{2\pi\sqrt{LC}}$$

where *L* is in henries and *C* is in farads.

Sensing

Most capacitors are designed to maintain a fixed physical structure. However, various factors can change the structure of the capacitor, and the resulting change in capacitance can be used to sense those factors.

Changing the dielectric:

The effects of varying the physical and/or electrical characteristics of the **dielectric** can be used for sensing purposes. Capacitors with an exposed and porous dielectric can be used to measure humidity in air. Capacitors are used to accurately measure the fuel level in airplanes; as the fuel covers more of a pair of plates, the circuit capacitance increases.

Changing the distance between the plates:

Capacitors with a flexible plate can be used to measure strain or pressure. Industrial pressure transmitters used for process control use pressure-sensing diaphragms, which form a capacitor plate of an oscillator circuit. Capacitors are used as the sensor in condenser microphones, where one plate is moved by air pressure, relative to the fixed position of the other plate. Some accelerometers use MEMS capacitors etched on a chip to measure the magnitude and direction of the acceleration vector. They are used to detect changes in acceleration, e.g. as tilt sensors or to detect free fall, as sensors triggering airbag deployment, and in many other applications. Some fingerprint sensors use capacitors. Additionally, a user can adjust the pitch of a theremin musical instrument by moving his hand since this changes the effective capacitance between the user's hand and the antenna.

Changing the effective area of the plates:

Capacitive touch switches are now used on many consumer electronic products.

Hazards and safety

Capacitors may retain a charge long after power is removed from a circuit; this charge can cause dangerous or even potentially fatal shocks or damage connected equipment. For example, even a seemingly innocuous device such as a disposable camera flash unit powered by a 1.5 volt AA battery contains a capacitor which may be charged to over 300 volts. This is easily capable of delivering a shock. Service procedures for electronic devices usually include instructions to discharge large or high-voltage capacitors. Capacitors may also have built-in discharge resistors to dissipate stored energy to a safe level within a few seconds after power is removed. High-voltage capacitors are stored with the terminals shorted, as protection from potentially dangerous voltages due to dielectric absorption.

Some old, large oil-filled capacitors contain polychlorinated biphenyls (PCBs). It is known that waste PCBs can leak into groundwater under landfills. Capacitors containing

PCB were labelled as containing "Askarel" and several other trade names. PCB-filled capacitors are found in very old (pre-1975) fluorescent lamp ballasts, and other applications.

Capacitors may catastrophically fail when subjected to voltages or currents beyond their rating, or as they reach their normal end of life. Dielectric or metal interconnection failures may create arcing that vaporizes the dielectric fluid, resulting in case bulging, rupture, or even an explosion. Capacitors used in RF or sustained high-current applications can overheat, especially in the center of the capacitor rolls. Capacitors used within high-energy capacitor banks can violently explode when a short in one capacitor causes sudden dumping of energy stored in the rest of the bank into the failing unit. High voltage vacuum capacitors can generate soft X-rays even during normal operation. Proper containment, fusing, and preventive maintenance can help to minimize these hazards.

High-voltage capacitors can benefit from a pre-charge to limit in-rush currents at power-up of high voltage direct current (HVDC) circuits. This will extend the life of the component and may mitigate high-voltage hazards.

Chapter-5

Inductor

Inductor

A selection of low-value inductors

Type	Passive
Working principle	Electromagnetic induction
First production	Michael Faraday (1831)

Electronic symbol

An **inductor** (or **reactor**) is a passive electrical component that can store energy in a magnetic field created by the electric current passing through it. An inductor's ability to store magnetic energy is measured by its inductance, in units of henries. Typically an inductor is a conducting wire shaped as a coil; the loops help to create a strong magnetic field inside the coil due to Ampere's Law. Due to the time-varying magnetic field inside the coil, a voltage is induced, according to Faraday's law of electromagnetic induction, which by Lenz's Law opposes the change in current that created it. Inductors are one of the basic components used in electronics where current and voltage change with time, due to the ability of inductors to delay and reshape alternating currents. Inductors called chokes are used as parts of filters in power supplies or to block AC signals from passing through a circuit.

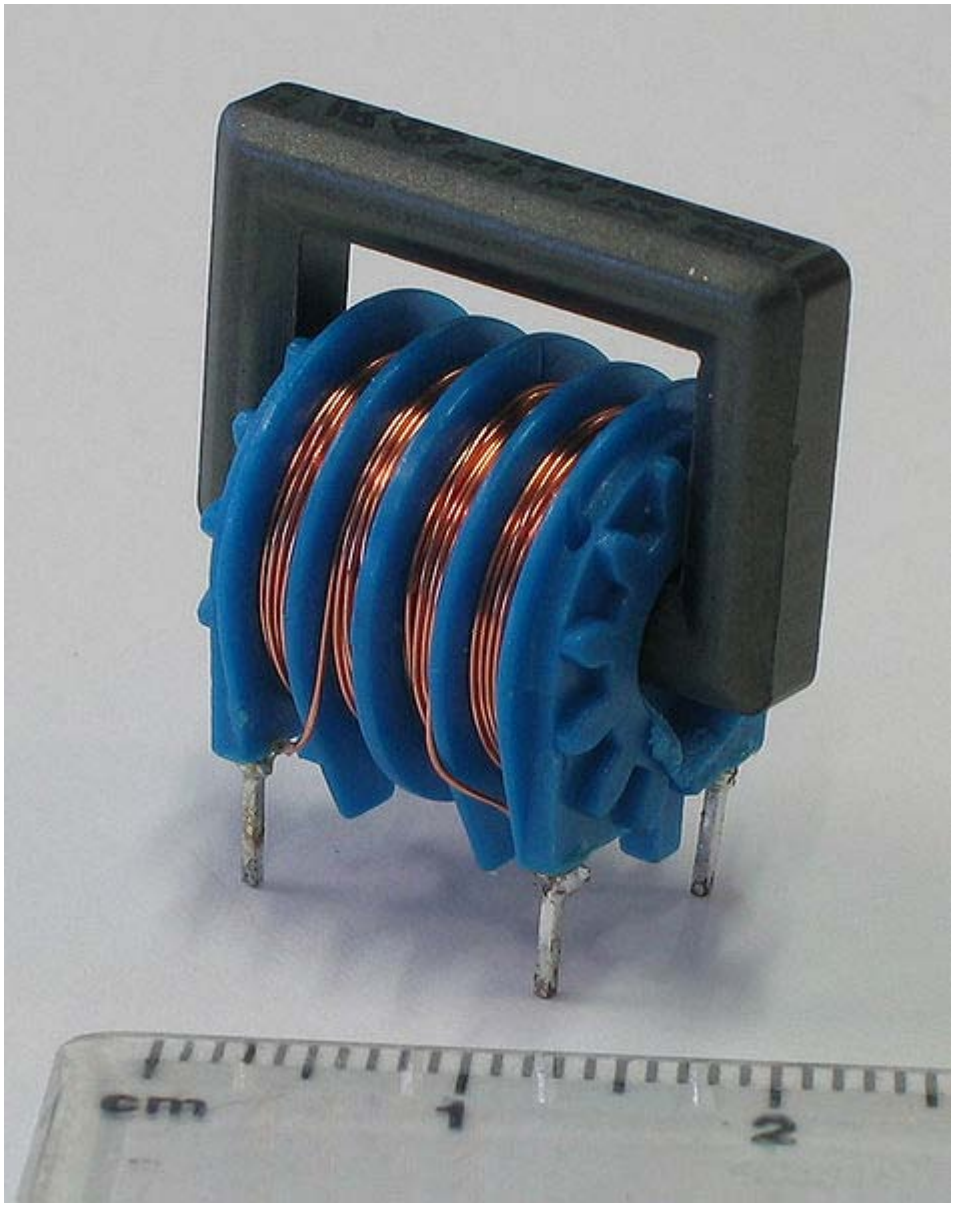
Overview

Inductance (L) results from the magnetic field forming around a current-carrying conductor which tends to resist changes in the current. Electric current through the conductor creates a magnetic flux proportional to the current, and a change in this current creates a corresponding change in magnetic flux which, in turn, by Faraday's Law generates an electromotive force (EMF) that opposes this change in current. Inductance is a measure of the amount of EMF generated per unit change in current. For example, an inductor with an inductance of 1 henry produces an EMF of 1 volt when the current through the inductor changes at the rate of 1 ampere per second. The number of loops, the size of each loop, and the material it is wrapped around all affect the inductance. For example, the magnetic flux linking these turns can be increased by coiling the conductor around a material with a high permeability such as iron. This can increase the inductance by 2000 times.

Ideal and real inductors

An "ideal inductor" has inductance, but no resistance or capacitance, and does not dissipate or radiate energy. A real inductor may be partially modeled by a combination of inductance, resistance (due to the resistance of the wire and losses in core material), and capacitance. At some frequency, some real inductors behave as resonant circuits (due to their self capacitance). At some frequency the capacitive component of impedance becomes dominant. Energy is dissipated by the resistance of the wire, and by any losses in the magnetic core due to hysteresis. Practical iron-core inductors at high currents show gradual departure from ideal behavior due to nonlinearity caused by magnetic saturation. At higher frequencies, resistance and resistive losses in inductors grow due to skin effect in the inductor's winding wires. Core losses also contribute to inductor losses at higher frequencies. Practical inductors work as antennas, radiating a part of energy processed into surrounding space and circuits, and accepting electromagnetic emissions from other circuits, taking part in electromagnetic interference. Circuits and materials close to the inductor will have near-field coupling to the inductor's magnetic field, which may cause additional energy loss. Real-world inductor applications may consider the parasitic parameters as important as the inductance.

Applications



An inductor with two 47mH windings, as may be found in a power supply.

Inductors are used extensively in analog circuits and signal processing. Inductors in conjunction with capacitors and other components form tuned circuits which can emphasize or filter out specific signal frequencies. Applications range from the use of large inductors in power supplies, which in conjunction with filter capacitors remove residual hums known as the mains hum or other fluctuations from the direct current output, to the small inductance of the ferrite bead or torus installed around a cable to prevent radio frequency interference from being transmitted down the wire. Smaller inductor/capacitor combinations provide tuned circuits used in radio reception and broadcasting, for instance.

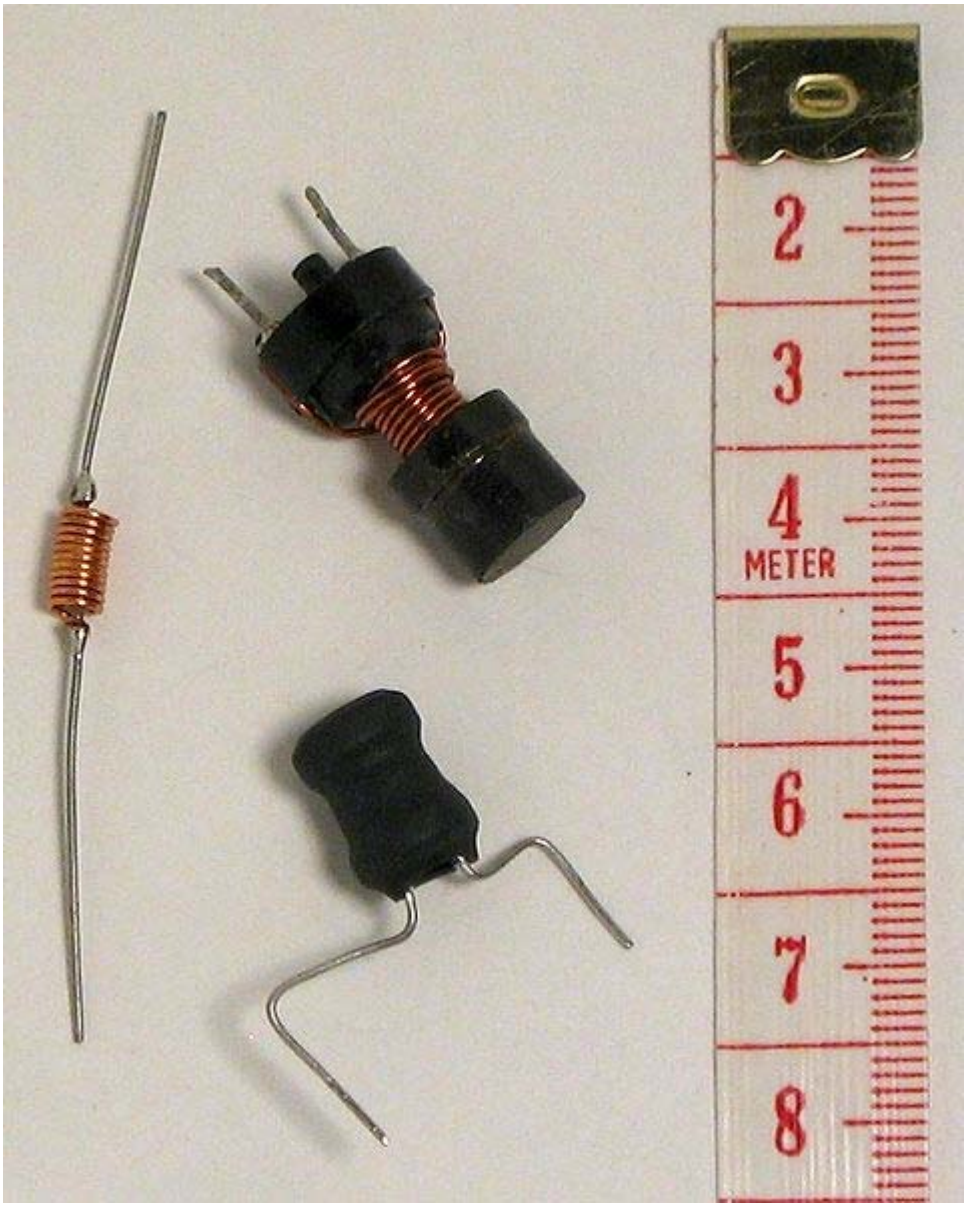
Two (or more) inductors that have coupled magnetic flux form a transformer, which is a fundamental component of every electric utility power grid. The efficiency of a transformer may decrease as the frequency increases due to eddy currents in the core material and skin effect on the windings. Size of the core can be decreased at higher frequencies and, for this reason, aircraft use 400 hertz alternating current rather than the usual 50 or 60 hertz, allowing a great saving in weight from the use of smaller transformers.

An inductor is used as the energy storage device in some switched-mode power supplies. The inductor is energized for a specific fraction of the regulator's switching frequency, and de-energized for the remainder of the cycle. This energy transfer ratio determines the input-voltage to output-voltage ratio. This X_L is used in complement with an active semiconductor device to maintain very accurate voltage control.

Inductors are also employed in electrical transmission systems, where they are used to depress voltages from lightning strikes and to limit switching currents and fault current. In this field, they are more commonly referred to as reactors.

Larger value inductors may be simulated by use of gyrator circuits.

Inductor construction



Inductors. Major scale in centimetres.

An inductor is usually constructed as a coil of conducting material, typically copper wire, wrapped around a core either of air or of ferromagnetic or ferrimagnetic material. Core materials with a higher permeability than air increase the magnetic field and confine it closely to the inductor, thereby increasing the inductance. Low frequency inductors are constructed like transformers, with cores of electrical steel laminated to prevent eddy currents. 'Soft' ferrites are widely used for cores above audio frequencies, since they do not cause the large energy losses at high frequencies that ordinary iron alloys do. Inductors come in many shapes. Most are constructed as enamel coated wire (magnet wire) wrapped around a ferrite bobbin with wire exposed on the outside, while some enclose the wire completely in ferrite and are referred to as "shielded". Some inductors

have an adjustable core, which enables changing of the inductance. Inductors used to block very high frequencies are sometimes made by stringing a ferrite cylinder or bead on a wire.

Small inductors can be etched directly onto a printed circuit board by laying out the trace in a spiral pattern. Some such planar inductors use a planar core.

Small value inductors can also be built on integrated circuits using the same processes that are used to make transistors. Aluminium interconnect is typically used, laid out in a spiral coil pattern. However, the small dimensions limit the inductance, and it is far more common to use a circuit called a "gyrator" that uses a capacitor and active components to behave similarly to an inductor.

Types of inductors

Air core coil

The term *air core coil* describes an inductor that does not use a magnetic core made of a ferromagnetic material. The term refers to coils wound on plastic, ceramic, or other nonmagnetic forms, as well as those that actually have air inside the windings. Air core coils have lower inductance than ferromagnetic core coils, but are often used at high frequencies because they are free from energy losses called core losses that occur in ferromagnetic cores, which increase with frequency. A side effect that can occur in air core coils in which the winding is not rigidly supported on a form is 'microphony': mechanical vibration of the windings can cause variations in the inductance.

Radio frequency inductors

At high frequencies, particularly radio frequencies (RF), inductors have higher resistance and other losses. In addition to causing power loss, in resonant circuits this can reduce the Q factor of the circuit, broadening the bandwidth. In RF inductors, which are mostly air core types, specialized construction techniques are used to minimize these losses. The losses are due to these effects:

- **Skin effect**: The resistance of a wire to high frequency current is higher than its resistance to direct current because of skin effect. Radio frequency alternating current does not penetrate far into the body of a conductor but travels along its surface. Therefore, in a solid wire, most of the cross sectional area of the wire is not used to conduct the current, which is in a narrow annulus on the surface. This effect increases the resistance of the wire in the coil, which may already have a relatively high resistance due to its length and small diameter.
- **Proximity effect**: Another similar effect that also increases the resistance of the wire at high frequencies is proximity effect, which occurs in parallel wires that lie close to each other. The individual magnetic field of adjacent turns induces eddy currents in the wire of the coil, which causes the current in the conductor to be concentrated in a thin strip on the side near the adjacent wire. Like skin effect,

this reduces the effective cross-sectional area of the wire conducting current, increasing its resistance.
- **Parasitic capacitance**: The capacitance between individual wire turns of the coil, called parasitic capacitance, does not cause energy losses but can change the behavior of the coil. Each turn of the coil is at a slightly different potential, so the electric field between neighboring turns stores charge on the wire. So the coil acts as if it has a capacitor in parallel with it. At a high enough frequency this capacitance can resonate with the inductance of the coil forming a tuned circuit, causing the coil to become self-resonant.

To reduce parasitic capacitance and proximity effect, RF coils are constructed to avoid having many turns lying close together, parallel to one another. The windings of RF coils are often limited to a single layer, and the turns are spaced apart. To reduce resistance due to skin effect, in high-power inductors such as those used in transmitters the windings are sometimes made of a metal strip or tubing which has a larger surface area, and the surface is silver-plated.

- **Honeycomb coils**: To reduce proximity effect and parasitic capacitance, multilayer RF coils are wound in patterns in which successive turns are not parallel but crisscrossed at an angle; these are often called *honeycomb* or *basket-weave* coils.
- **Spiderweb coils**: Another construction technique with similar advantages is flat spiral coils. These are often wound on a flat insulating support with radial spokes or slots, with the wire weaving in and out through the slots; these are called *spiderweb* coils. The form has an odd number of slots, so successive turns of the spiral lie on opposite sides of the form, increasing separation.
- **Litz wire**: To reduce skin effect losses, some coils are wound with a special type of radio frequency wire called litz wire. Instead of a single solid conductor, litz wire consists of several smaller wire strands that carry the current. Unlike ordinary stranded wire, the strands are insulated from each other, to prevent skin effect from forcing the current to the surface, and are braided together. The braid pattern ensures that each wire strand spends the same amount of its length on the outside of the braid, so skin effect distributes the current equally between the strands, resulting in a larger cross-sectional conduction area than an equivalent single wire.

Ferromagnetic core coil

Ferromagnetic-core or iron-core inductors use a magnetic core made of a ferromagnetic or ferrimagnetic material such as iron or ferrite to increase the inductance. A magnetic core can increase the inductance of a coil by a factor of several thousand, by increasing the magnetic field due to its higher magnetic permeability. However the magnetic properties of the core material cause several side effects which alter the behavior of the inductor and require special construction:

- Core losses: A time-varying current in a ferromagnetic inductor, which causes a time-varying magnetic field in its core, causes energy losses in the core material that are dissipated as heat, due to two processes:
    - Eddy currents: From Faraday's law of induction, the changing magnetic field can induce circulating loops of electric current in the conductive metal core. The energy in these currents is dissipated as heat in the resistance of the core material. The amount of energy lost increases with the area inside the loop of current.
    - Hysteresis: Changing or reversing the magnetic field in the core also causes losses due to the motion of the tiny magnetic domains it is composed of. The energy loss is proportional to the area of the hysteresis loop in the BH graph of the core material. Materials with low coercivity have narrow hysteresis loops and so low hysteresis losses.

  For both of these processes, the energy loss per cycle of alternating current is constant, so core losses increase linearly with frequency.

- Nonlinearity: If the current through a ferromagnetic core coil is high enough that the magnetic core saturates, the inductance will not remain constant but will change with the current through the device. This is called nonlinearity and results in distortion of the signal. For example, audio signals can suffer intermodulation distortion in saturated inductors. To prevent this, in linear circuits the current through iron core inductors must be limited below the saturation level. Using a powdered iron core with a distributed air gap allows higher levels of magnetic flux which in turn allows a higher level of direct current through the inductor before it saturates.

Laminated core inductor

Low-frequency inductors are often made with laminated cores to prevent eddy currents, using construction similar to transformers. The core is made of stacks of thin steel sheets or laminations oriented parallel to the field, with an insulating coating on the surface. The insulation prevents eddy currents between the sheets, so any remaining currents must be within the cross sectional area of the individual laminations, reducing the area of the loop and thus the energy loss greatly. The laminations are made of low-coercivity silicon steel, to reduce hysteresis losses.

Ferrite-core inductor

For higher frequencies, inductors are made with cores of ferrite. Ferrite is a ceramic ferrimagnetic material that is nonconductive, so eddy currents cannot flow within it. The formulation of ferrite is $xxFe_2O_4$ where xx represents various metals. For inductor cores soft ferrites are used, which have low coercivity and thus low hysteresis losses. Another similar material is powdered iron cemented with a binder.

Toroidal core coils

In an inductor wound on a straight rod-shaped core, the magnetic field lines emerging from one end of the core must pass through the air to reenter the core at the other end. This reduces the field, because much of the magnetic field path is in air rather than the higher permeability core material. A higher magnetic field and inductance can be achieved by forming the core in a closed magnetic circuit. The magnetic field lines form closed loops within the core without leaving the core material. The shape often used is a toroidal or doughnut-shaped ferrite core. Because of their symmetry, toroidal cores allow a minimum of the magnetic flux to escape outside the core (called *leakage flux*), so they radiate less electromagnetic interference than other shapes. Toroidal core coils are manufactured of various materials, primarily ferrite, Kool Mu MPP, powdered iron and laminated cores.

Variable inductor

A variable inductor can be constructed by making one of the terminals of the device a sliding spring contact that can move along the surface of the coil, increasing or decreasing the number of turns of the coil included in the circuit. An alternate construction method is to use a moveable magnetic core, which can be slid in or out of the coil. Moving the core farther into the coil increases the permeability, increasing the inductance. Many inductors used in radio applications (usually less than 100 MHz) use adjustable cores in order to tune such inductors to their desired value, since manufacturing processes have certain tolerances (inaccuracy).

Core loss

Core loss calculators can be used to determine the type of inductor required. Using inputs such as input voltage, output voltage, output current, frequency, ambient temperature, and inductance these calculators can predict the losses of the inductors core and AC/DC based on the operating condition of the circuit being used.

In electric circuits

The effect of an inductor in a circuit is to oppose changes in current through it by developing a voltage across it proportional to the rate of change of the current. An ideal inductor would offer no resistance to a constant direct current; however, only superconducting inductors have truly zero electrical resistance.

The relationship between the time-varying voltage $v(t)$ across an inductor with inductance L and the time-varying current $i(t)$ passing through it is described by the differential equation:

$$v(t) = L\frac{di(t)}{dt}$$

When there is a sinusoidal alternating current (AC) through an inductor, a sinusoidal voltage is induced. The amplitude of the voltage is proportional to the product of the amplitude (I_P) of the current and the frequency (f) of the current.

$$i(t) = I_P \sin(2\pi f t)$$
$$\frac{di(t)}{dt} = 2\pi f I_P \cos(2\pi f t)$$
$$v(t) = 2\pi f L I_P \cos(2\pi f t)$$

In this situation, the phase of the current lags that of the voltage by π/2.

If an inductor is connected to a direct current source with value I via a resistance R, and then the current source is short-circuited, the differential relationship above shows that the current through the inductor will discharge with an exponential decay:

$$i(t) = Ie^{-(R/L)t}$$

Laplace circuit analysis (s-domain)

When using the Laplace transform in circuit analysis, the impedance of an ideal inductor with no initial current is represented in the s domain by:

$$Z(s) = Ls$$

where
L is the inductance, and
s is the complex frequency.

If the inductor does have initial current, it can be represented by:

- adding a voltage source in series with the inductor, having the value:

$$LI_0$$

(*Note that the source should have a polarity that is aligned with the initial current*)

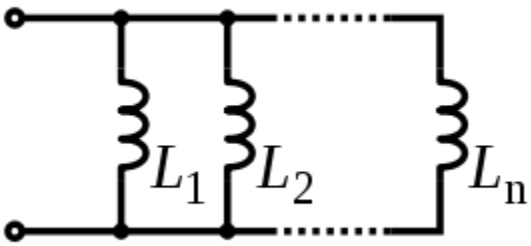
- or by adding a current source in parallel with the inductor, having the value:

$$\frac{I_0}{s}$$

where
L is the inductance, and
I_0 *is the initial current in the inductor.*

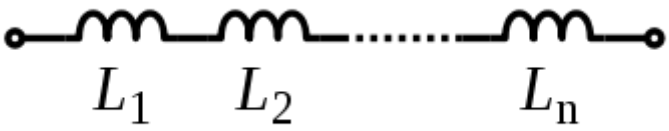
Inductor networks

Inductors in a parallel configuration each have the same potential difference (voltage). To find their total equivalent inductance (L_{eq}):



$$\frac{1}{L_{\text{eq}}} = \frac{1}{L_1} + \frac{1}{L_2} + \cdots + \frac{1}{L_n}$$

The current through inductors in series stays the same, but the voltage across each inductor can be different. The sum of the potential differences (voltage) is equal to the total voltage. To find their total inductance:



$$L_{\text{eq}} = L_1 + L_2 + \cdots + L_n$$

These simple relationships hold true only when there is no mutual coupling of magnetic fields between individual inductors.

Stored energy

The energy (measured in joules, in SI) stored by an inductor is equal to the amount of work required to establish the current through the inductor, and therefore the magnetic field. This is given by:

$$E_{\text{stored}} = \frac{1}{2}LI^2$$

where L is inductance and I is the current through the inductor.

This relationship is only valid for linear (non-saturated) regions of the magnetic flux linkage and current relationship.

Q *factor*

An ideal inductor will be lossless irrespective of the amount of current through the winding. However, typically inductors have winding resistance from the metal wire forming the coils. Since the winding resistance appears as a resistance in series with the inductor, it is often called the *series resistance*. The inductor's series resistance converts electric current through the coils into heat, thus causing a loss of inductive quality. The quality factor (or Q) of an inductor is the ratio of its inductive reactance to its resistance at a given frequency, and is a measure of its efficiency. The higher the Q factor of the inductor, the closer it approaches the behavior of an ideal, lossless, inductor.

The Q factor of an inductor can be found through the following formula, where R is its internal electrical resistance and ωL is capacitive or inductive reactance at resonance:

$$Q = \frac{\omega L}{R}$$

By using a ferromagnetic core, the inductance is greatly increased for the same amount of copper, multiplying up the Q. Cores however also introduce losses that increase with frequency. A grade of core material is chosen for best results for the frequency band. At VHF or higher frequencies an air core is likely to be used.

Inductors wound around a ferromagnetic core may saturate at high currents, causing a dramatic decrease in inductance (and Q). This phenomenon can be avoided by using a (physically larger) air core inductor. A well designed air core inductor may have a Q of several hundred.

An almost ideal inductor (Q approaching infinity) can be created by immersing a coil made from a superconducting alloy in liquid helium or liquid nitrogen. This supercools the wire, causing its winding resistance to disappear. Because a superconducting inductor is virtually lossless, it can store a large amount of electrical energy within the surrounding magnetic field. Bear in mind that for inductors with cores, core losses still exist.

Inductance formulae

The table below lists some common simplified formulas for calculating the approximate inductance of several inductor constructions.

Construction	Formula	Dimensions	Notes
Cylindrical air-core coil	$L = \frac{\mu_0 K N^2 A}{l}$	L = inductance in henries (H) μ_0 = permeability of free space = $4\pi \times 10^{-7}$ H/m K = Nagaoka coefficient N = number of turns A = area of cross-section of the coil in square metres (m^2) l = length of coil in metres (m)	
Straight wire conductor	$L = \frac{\mu_0}{2\pi}\left[l \ln \frac{l + \sqrt{l^2 + c^2}}{c} - \sqrt{l^2 + c^2} + c\right] + \frac{\mu}{2\pi} o\left(\frac{l}{4 + c\sqrt{\frac{2\omega\mu}{\rho}}}\right)$	L = inductance l = cylinder length c = cylinder radius μ_0 = vacuum permeability = 4π nH/cm μ =	exact if ω = 0 or ω = ∞

			conductor permeability	
			p = resistivity	
			ω = phase rate	
	$L = 0.2l\left(\ln\frac{4l}{d} - 1\right)_{\text{-0+3\%}}$		L = inductance (μH)	Cu or Al
			l = length of conductor (mm)	$l >$ 100 d
			d = diameter of conductor (mm)	$d^2 f >$ 1 mm^2 MHz
			f = frequency	
	$L = 0.2l\left(\ln\frac{4l}{d} - \frac{3}{4}\right)_{\text{+0-3\%}}$		L = inductance (μH)	Cu or Al
			l = length of conductor (mm)	$l >$ 100 d
			d = diameter of conductor (mm)	$d^2 f <$ 1 mm^2 MHz
			f = frequency	
Short air-core cylindrical coil	$L = \frac{r^2 N^2}{9r + 10l}$		L = inductance (μH)	
			r = outer radius of coil (in)	
			l = length of coil (in)	

Multilayer air-core coil	$L = \frac{0.8r^2N^2}{6r + 9l + 10d}$	N = number of turns L = inductance (μH) r = mean radius of coil (in) l = physical length of coil winding (in)
Flat spiral air-core coil	$L = \frac{r^2N^2}{(20r + 28d)}$	N = number of turns d = depth of coil (outer radius minus inner radius) (cm) L = inductance (μH) r = mean radius of coil (cm)
	$L = \frac{r^2N^2}{8r + 11d}$	N = number of turns d = depth of coil (outer radius minus inner radius) (cm) L = inductance (μH) r = mean radius of coil

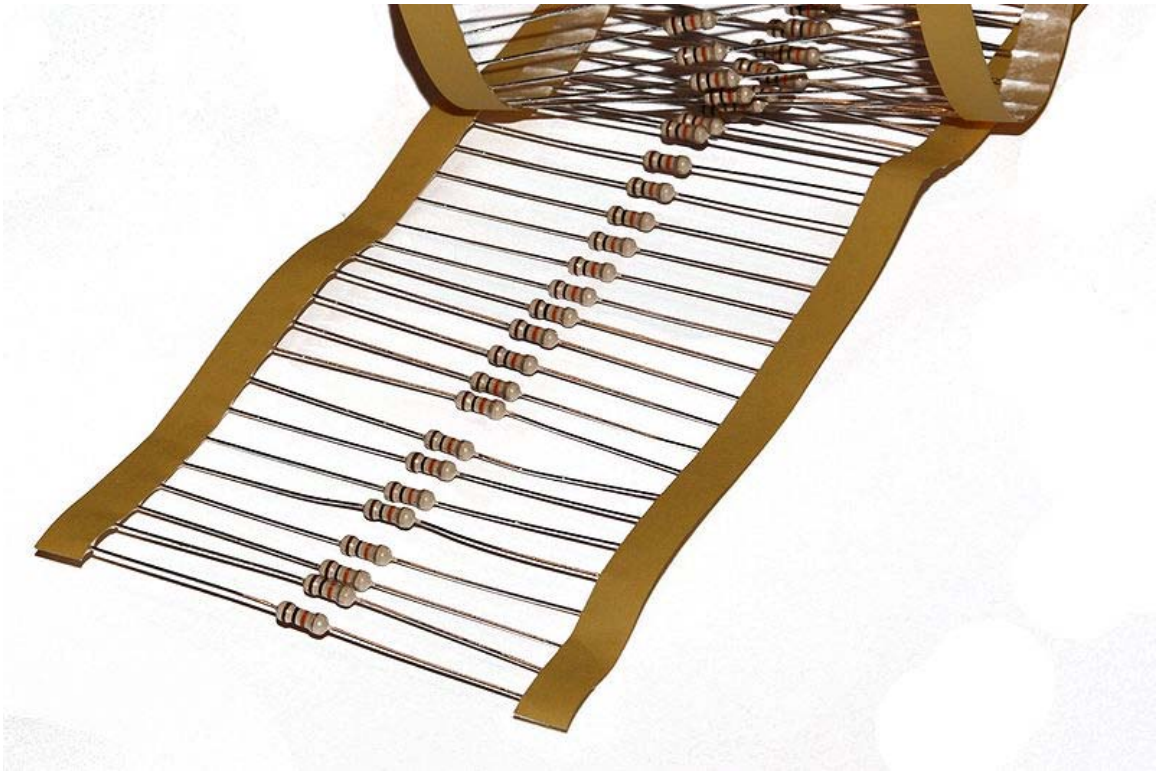
		(in) N = number of turns d = depth of coil (outer radius minus inner radius) (in) L = inductance (H) μ_0 = permeability of free space = $4\pi \times 10^{-7}$ H/m μ_r = relative permeability of core material r = radius of coil winding (m) N = number of turns D = overall diameter of toroid (m)
Toroidal core (circular cross-section)	$L = \mu_0 \mu_r \frac{r^2 N^2}{D}$	

Chapter-6

Resistor

A typical axial-lead resistor

Partially exposed Tesla TR-212 1 kΩ carbon film resistor



Axial-lead resistors on tape. The tape is removed during assembly before the leads are formed and the part is inserted into the board.

Three carbon composition resistors in a 1960s valve (vacuum tube) radio

A **resistor** is a two-terminal passive electronic component which implements electrical resistance as a circuit element. When a voltage V is applied across the terminals of a resistor, a current I will flow through the resistor in direct proportion to that voltage. The reciprocal of the constant of proportionality is known as the resistance R, since, with a given voltage V, a larger value of R further "resists" the flow of current I as given by Ohm's law:

$$I = \frac{V}{R}$$

Resistors are common elements of electrical networks and electronic circuits and are ubiquitous in most electronic equipment. Practical resistors can be made of various compounds and films, as well as resistance wire (wire made of a high-resistivity alloy, such as nickel-chrome). Resistors are also implemented within integrated circuits, particularly analog devices, and can also be integrated into hybrid and printed circuits.

The electrical functionality of a resistor is specified by its resistance: common commercial resistors are manufactured over a range of more than 9 orders of magnitude. When specifying that resistance in an electronic design, the required precision of the resistance may require attention to the manufacturing tolerance of the chosen resistor, according to its specific application. The temperature coefficient of the resistance may also be of concern in some precision applications. Practical resistors are also specified as having a maximum power rating which must exceed the anticipated power dissipation of that resistor in a particular circuit: this is mainly of concern in power electronics

applications. Resistors with higher power ratings are physically larger and may require heat sinking. In a high voltage circuit, attention must sometimes be paid to the rated maximum working voltage of the resistor.

The series inductance of a practical resistor causes its behavior to depart from ohms law; this specification can be important in some high-frequency applications for smaller values of resistance. In a low-noise amplifier or pre-amp the noise characteristics of a resistor may be an issue. The unwanted inductance, excess noise, and temperature coefficient are mainly dependent on the technology used in manufacturing the resistor. They are not normally specified individually for a particular family of resistors manufactured using a particular technology. A family of discrete resistors is also characterized according to its form factor, that is, the size of the device and position of its leads (or terminals) which is relevant in the practical manufacturing of circuits using them.

Units

The ohm (symbol: Ω) is the SI unit of electrical resistance, named after Georg Simon Ohm. An ohm is equivalent to a volt per ampere. Since resistors are specified and manufactured over a very large range of values, the derived units of milliohm (1 mΩ = 10^{-3} Ω), kilohm (1 kΩ = 10^{3} Ω), and megohm (1 MΩ = 10^{6} Ω) are also in common usage.

The reciprocal of resistance R is called conductance G = 1/R and is measured in Siemens (SI unit), sometimes referred to as a mho. Thus a Siemens is the reciprocal of an ohm: $S = \Omega^{-1}$. Although the concept of conductance is often used in circuit analysis, practical resistors are always specified in terms of their resistance (ohms) rather than conductance.

Theory of operation

Ohm's law

The behavior of an ideal resistor is dictated by the relationship specified in Ohm's law:

$$V = I \cdot R$$

Ohm's law states that the voltage (V) across a resistor is proportional to the current (I) passing through it, where the constant of proportionality is the resistance (R).

Equivalently, Ohm's law can be stated:

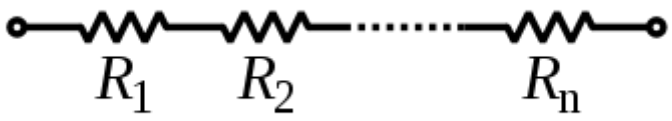
$$I = \frac{V}{R}$$

This formulation of Ohm's law states that, when a voltage (V) is present across a resistance (R), a current (I) will flow through the resistance. This is directly used in practical computations. For example, if a 300 ohm resistor is attached across the

terminals of a 12 volt battery, then a current of 12 / 300 = 0.04 amperes (or 40 milliamperes) will flow through that resistor.

Series and parallel resistors

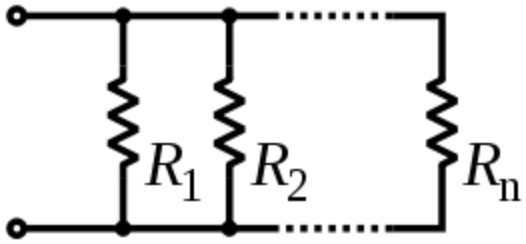
In a series configuration, the current through all of the resistors is the same, but the voltage across each resistor will be in proportion to its resistance. The potential difference (voltage) seen across the network is the sum of those voltages, thus the total resistance can be found as the sum of those resistances:



$$R_{\text{eq}} = R_1 + R_2 + \cdots + R_n$$

As a special case, the resistance of N resistors connected in series, each of the same resistance R, is given by NR.

Resistors in a parallel configuration are each subject to the same potential difference (voltage), however the currents through them add. The conductances of the resistors then add to determine the conductance of the network. Thus the equivalent resistance (R_{eq}) of the network can be computed:



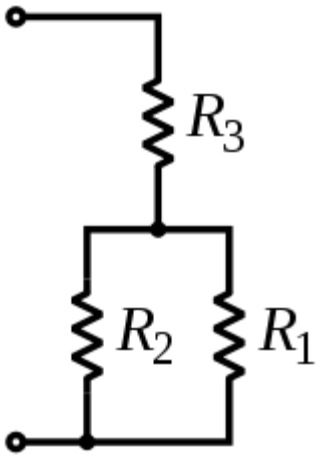
$$\frac{1}{R_{\text{eq}}} = \frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_n}$$

The parallel equivalent resistance can be represented in equations by two vertical lines "||" (as in geometry) as a simplified notation. For the case of two resistors in parallel, this can be calculated using:

$$R_{\text{eq}} = R_1 \| R_2 = \frac{R_1 R_2}{R_1 + R_2}$$

As a special case, the resistance of N resistors connected in parallel, each of the same resistance R, is given by R/N.

A resistor network that is a combination of parallel and series connections can be broken up into smaller parts that are either one or the other. For instance,



$$R_{\text{eq}} = (R_1 \| R_2) + R_3 = \frac{R_1 R_2}{R_1 + R_2} + R_3$$

However, some complex networks of resistors cannot be resolved in this manner, requiring more sophisticated circuit analysis. For instance, consider a cube, each edge of which has been replaced by a resistor. What then is the resistance that would be measured between two opposite vertices? In the case of 12 equivalent resistors, it can be shown that the corner-to-corner resistance is $^5/_6$ of the individual resistance. More generally, the Y-Δ transform, or matrix methods can be used to solve such a problem.

One practical application of these relationships is that a non-standard value of resistance can generally be synthesized by connecting a number of standard values in series and/or parallel. This can also be used to obtain a resistance with a higher power rating than that of the individual resistors used. In the special case of N identical resistors all connected in series or all connected in parallel, the power rating of the individual resistors is thereby multiplied by N.

Power dissipation

The power P dissipated by a resistor (or the equivalent resistance of a resistor network) is calculated as: $P = I^2R = IV = \frac{V^2}{R}$

The first form is a restatement of Joule's first law. Using Ohm's law, the two other forms can be derived.

The total amount of heat energy released over a period of time can be determined from the integral of the power over that period of time:

$$W = \int_{t_1}^{t_2} v(t)i(t)\,dt.$$

Practical resistors are rated according to their maximum power dissipation. The vast majority of resistors used in electronic circuits absorb much less than a watt of electrical power and require no attention to their power rating. Such resistors in their discrete form, including most of the packages detailed below, are typically rated as 1/10, 1/8, or 1/4 watt.

Resistors required to dissipate substantial amounts of power, particularly used in power supplies, power conversion circuits, and power amplifiers, are generally referred to as *power resistors*; this designation is loosely applied to resistors with power ratings of 1 watt or greater. Power resistors are physically larger and tend not to use the preferred values, color codes, and external packages described below.

If the average power dissipated by a resistor is more than its power rating, damage to the resistor may occur, permanently altering its resistance; this is distinct from the reversible change in resistance due to its temperature coefficient when it warms. Excessive power dissipation may raise the temperature of the resistor to a point where it can burn the circuit board or adjacent components, or even cause a fire. There are flameproof resistors that fail (open circuit) before they overheat dangerously.

Note that the nominal power rating of a resistor is not the same as the power that it can safely dissipate in practical use. Air circulation and proximity to a circuit board, ambient temperature, and other factors can reduce acceptable dissipation significantly. Rated power dissipation may be given for an ambient temperature of 25 °C in free air. Inside an equipment case at 60 °C, rated dissipation will be significantly less; a resistor dissipating a bit less than the maximum figure given by the manufacturer may still be outside the safe operating area and may prematurely fail.

Construction

A single in line (SIL) resistor package with 8 individual, 47 ohm resistors. One end of each resistor is connected to a separate pin and the other ends are all connected together to the remaining (common) pin - pin 1, at the end identified by the white dot.

Lead arrangements

Resistors with wire leads for through-hole mounting

Through-hole components typically have leads leaving the body axially. Others have leads coming off their body radially instead of parallel to the resistor axis. Other components may be SMT (surface mount technology) while high power resistors may have one of their leads designed into the heat sink.

Carbon composition

Carbon composition resistors consist of a solid cylindrical resistive element with embedded wire leads or metal end caps to which the lead wires are attached. The body of the resistor is protected with paint or plastic. Early 20th-century carbon composition resistors had uninsulated bodies; the lead wires were wrapped around the ends of the resistance element rod and soldered. The completed resistor was painted for color coding of its value.

The resistive element is made from a mixture of finely ground (powdered) carbon and an insulating material (usually ceramic). A resin holds the mixture together. The resistance is determined by the ratio of the fill material (the powdered ceramic) to the carbon. Higher concentrations of carbon, a weak conductor, result in lower resistance. Carbon composition resistors were commonly used in the 1960s and earlier, but are not so popular for general use now as other types have better specifications, such as tolerance, voltage dependence, and stress (carbon composition resistors will change value when stressed with over-voltages). Moreover, if internal moisture content (from exposure for some length of time to a humid environment) is significant, soldering heat will create a non-reversible change in resistance value. Carbon composition resistors have poor stability with time and were consequently factory sorted to, at best, only 5% tolerance. These resistors, however, if never subjected to overvoltage nor overheating were remarkably reliable considering the component's size

They are still available, but comparatively quite costly. Values ranged from fractions of an ohm to 22 megohms. Because of the high price, these resistors are no longer used in most applications. However, carbon resistors are used in power supplies and welding controls.

Carbon film

A carbon film is deposited on an insulating substrate, and a helix cut in it to create a long, narrow resistive path. Varying shapes, coupled with the resistivity of carbon, (ranging from 90 to 400 nΩ m) can provide a variety of resistances. Carbon film resistors feature a power rating range of 0.125 W to 5 W at 70 °C. Resistances available range from 1 ohm to 10 megohm. The carbon film resistor has an operating temperature range of -55 °C to 155 °C. It has 200 to 600 volts maximum working voltage range. Special carbon film resistors are used in applications requiring high pulse stability.

Thick and thin film

Thick film resistors became popular during the 1970s, and most SMD (surface mount device) resistors today are of this type. The principal difference between thin film and thick film resistors is not the actual thickness of the film, but rather how the film is applied to the cylinder (axial resistors) or the surface (SMD resistors).

Thin film resistors are made by sputtering (a method of vacuum deposition) the resistive material onto an insulating substrate. The film is then etched in a similar manner to the old (subtractive) process for making printed circuit boards; that is, the surface is coated with a photo-sensitive material, then covered by a pattern film, irradiated with ultraviolet light, and then the exposed photo-sensitive coating is developed, and underlying thin film is etched away.

Thick film resistors are manufactured using screen and stencil printing processes.

Because the time during which the sputtering is performed can be controlled, the thickness of the thin film can be accurately controlled. The type of material is also usually different consisting of one or more ceramic (cermet) conductors such as tantalum nitride (TaN), ruthenium dioxide (RuO_2), lead oxide (PbO), bismuth ruthenate ($Bi_2Ru_2O_7$), nickel chromium (NiCr), and/or bismuth iridate ($Bi_2Ir_2O_7$).

The resistance of both thin and thick film resistors after manufacture is not highly accurate; they are usually trimmed to an accurate value by abrasive or laser trimming. Thin film resistors are usually specified with tolerances of 0.1, 0.2, 0.5, or 1%, and with temperature coefficients of 5 to 25 ppm/K.

Thick film resistors may use the same conductive ceramics, but they are mixed with sintered (powdered) glass and some kind of liquid so that the composite can be screen-printed. This composite of glass and conductive ceramic (cermet) material is then fused (baked) in an oven at about 850 °C.

Thick film resistors, when first manufactured, had tolerances of 5%, but standard tolerances have improved to 2% or 1% in the last few decades. Temperature coefficients of thick film resistors are high, typically ±200 or ±250 ppm/K; a 40 kelvin (70 °F) temperature change can change the resistance by 1%.

Thin film resistors are usually far more expensive than thick film resistors. For example, SMD thin film resistors, with 0.5% tolerances, and with 25 ppm/K temperature coefficients, when bought in full size reel quantities, are about twice the cost of 1%, 250 ppm/K thick film resistors.

Metal film

A common type of axial resistor today is referred to as a metal-film resistor. Metal electrode leadless face (MELF) resistors often use the same technology, but are a cylindrically shaped resistor designed for surface mounting. Note that other types of resistors (e.g., carbon composition) are also available in MELF packages.

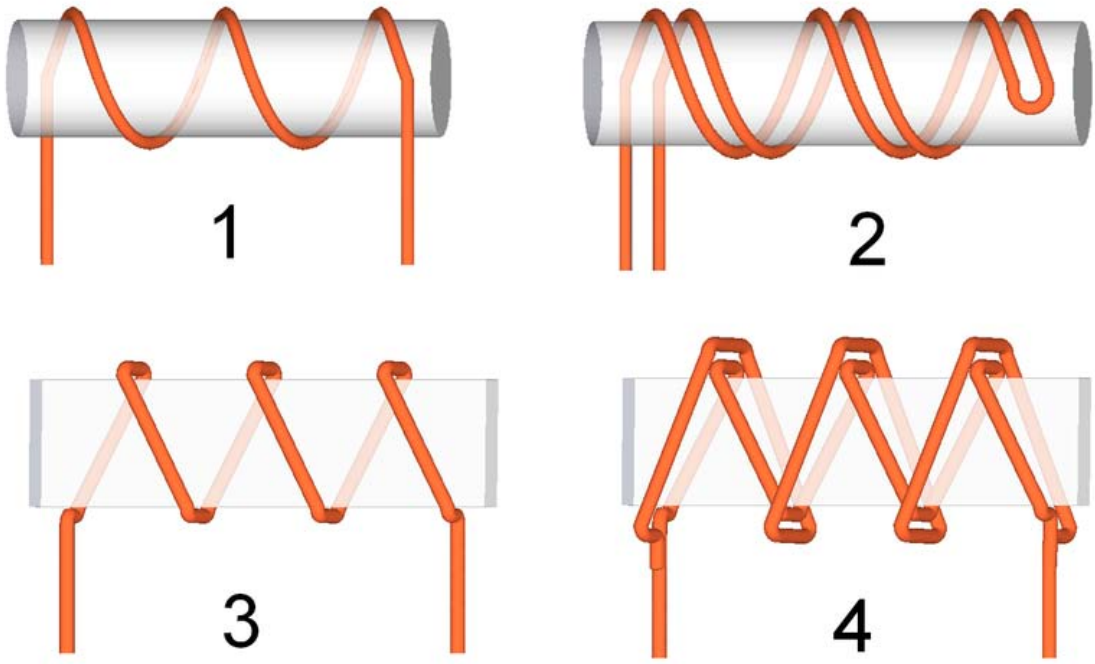
Metal film resistors are usually coated with nickel chromium (NiCr), but might be coated with any of the cermet materials listed above for thin film resistors. Unlike thin film resistors, the material may be applied using different techniques than sputtering (though that is one such technique). Also, unlike thin-film resistors, the resistance value is

determined by cutting a helix through the coating rather than by etching. (This is similar to the way carbon resistors are made.) The result is a reasonable tolerance (0.5, 1, or 2%) and a temperature coefficient that is generally between 50 and 100 ppm/K. Metal film resistors possess good noise characteristics and low non-linearity due to a low voltage coefficient. Also beneficial are the components efficient tolerance, temperature coefficient and stability.

Metal Oxide film

Metal-Oxide film resistors resemble Metal film types, but are made of metal oxides such as tin oxide. This results in a higher operating temperature and greater stability/reliability than Metal film. They are used in applications with high endurance demands.

Wirewound



Types of windings in wire resistors:
1 - common
2 - bifilar
3 - common on a thin former
4 - Ayrton-Perry

Wirewound resistors are commonly made by winding a metal wire, usually nichrome, around a ceramic, plastic, or fiberglass core. The ends of the wire are soldered or welded to two caps or rings, attached to the ends of the core. The assembly is protected with a layer of paint, molded plastic, or an enamel coating baked at high temperature. Because of the very high surface temperature these resistors can withstand temperatures of up to

+450 °C. Wire leads in low power wirewound resistors are usually between 0.6 and 0.8 mm in diameter and tinned for ease of soldering. For higher power wirewound resistors, either a ceramic outer case or an aluminum outer case on top of an insulating layer is used. The aluminum-cased types are designed to be attached to a heat sink to dissipate the heat; the rated power is dependent on being used with a suitable heat sink, e.g., a 50 W power rated resistor will overheat at a fraction of the power dissipation if not used with a heat sink. Large wirewound resistors may be rated for 1,000 watts or more.

Because wirewound resistors are coils they have more undesirable inductance than other types of resistor, although winding the wire in sections with alternately reversed direction can minimize inductance. Other techniques employ bifilar winding, or a flat thin former (to reduce cross-section area of the coil). For most demanding circuits resistors with Ayrton-Perry winding are used.

Applications of wirewound resistors are similar to those of composition resistors with the exception of the high frequency. The high frequency of wirewound resistors is substantially worse than that of a composition resistor.

Foil resistor

The primary resistance element of a foil resistor is a special alloy foil several micrometres thick. Since their introduction in the 1960s, foil resistors have had the best precision and stability of any resistor available. One of the important parameters influencing stability is the temperature coefficient of resistance (TCR). The TCR of foil resistors is extremely low, and has been further improved over the years. One range of ultra-precision foil resistors offers a TCR of 0.14 ppm/°C, tolerance ±0.005%, long-term stability (1 year) 25 ppm, (3 year) 50 ppm (further improved 5-fold by hermetic sealing), stability under load (2000 hours) 0.03%, thermal EMF 0.1 μV/°C, noise -42 dB, voltage coefficient 0.1 ppm/V, inductance 0.08 μH, capacitance 0.5 pF.

Ammeter shunts

An ammeter shunt is a special type of current-sensing resistor, having four terminals and a value in milliohms or even micro-ohms. Current-measuring instruments, by themselves, can usually accept only limited currents. To measure high currents, the current passes through the shunt, where the voltage drop is measured and interpreted as current. A typical shunt consists of two solid metal blocks, sometimes brass, mounted on to an insulating base. Between the blocks, and soldered or brazed to them, are one or more strips of low temperature coefficient of resistance (TCR) manganin alloy. Large bolts threaded into the blocks make the current connections, while much-smaller screws provide voltage connections. Shunts are rated by full-scale current, and often have a voltage drop of 50 mV at rated current. Such meters are adapted to the shunt full current rating by using an appropriately marked dial face; no change need be made to the other parts of the meter.

Grid resistor

In heavy-duty industrial high-current applications, a grid resistor is a large convection-cooled lattice of stamped metal alloy strips connected in rows between two electrodes. Such industrial grade resistors can be as large as a refrigerator; some designs can handle over 500 amperes of current, with a range of resistances extending lower than 0.04 ohms. They are used in applications such as dynamic braking and load banking for locomotives and trams, neutral grounding for industrial AC distribution, control loads for cranes and heavy equipment, load testing of generators and harmonic filtering for electric substations.

The term *grid resistor* is sometimes used to describe a resistor of any type connected to the control grid of a vacuum tube. This is not a resistor technology; it is an electronic circuit topology.

Special varieties

- Metal oxide varistor
- Cermet
- Phenolic
- Tantalum
- Water resistor

Variable resistors

Adjustable resistors

A resistor may have one or more fixed tapping points so that the resistance can be changed by moving the connecting wires to different terminals. Some wirewound power resistors have a tapping point that can slide along the resistance element, allowing a larger or smaller part of the resistance to be used.

Where continuous adjustment of the resistance value during operation of equipment is required, the sliding resistance tap can be connected to a knob accessible to an operator. Such a device is called a rheostat and has two terminals.

Potentiometers

A common element in electronic devices is a three-terminal resistor with a continuously adjustable tapping point controlled by rotation of a shaft or knob. These variable resistors are known as potentiometers when all three terminals are present, since they act as a continuously adjustable voltage divider. A common example is a volume control for a radio receiver.

Accurate, high-resolution panel-mounted potentiometers (or "pots") have resistance elements typically wirewound on a helical mandrel, although some include a conductive-

plastic resistance coating over the wire to improve resolution. These typically offer ten turns of their shafts to cover their full range. They are usually set with dials that include a simple turns counter and a graduated dial. Electronic analog computers used them in quantity for setting coefficients, and delayed-sweep oscilloscopes of recent decades included one on their panels.

Resistance decade boxes

A resistance decade box or resistor substitution box is a unit containing resistors of many values, with one or more mechanical switches which allow any one of various discrete resistances offered by the box to be dialed in. Usually the resistance is accurate to high precision, ranging from laboratory/calibration grade accuracy of 20 parts per million, to field grade at 1%. Inexpensive boxes with lesser accuracy are also available. All types offer a convenient way of selecting and quickly changing a resistance in laboratory, experimental and development work without needing to attach resistors one by one, or even stock each value. The range of resistance provided, the maximum resolution, and the accuracy characterize the box. For example, one box offers resistances from 0 to 24 megohms, maximum resolution 0.1 ohm, accuracy 0.1%.

Special devices

There are various devices whose resistance changes with various quantities. The resistance of thermistors exhibit a strong negative temperature coefficient, making them useful for measuring temperatures. Since their resistance can be large until they are allowed to heat up due to the passage of current, they are also commonly used to prevent excessive current surges when equipment is powered on. Similarly, the resistance of a humistor varies with humidity. Metal oxide varistors drop to a very low resistance when a high voltage is applied, making them useful for protecting electronic equipment by absorbing dangerous voltage surges. One sort of photodetector, the photoresistor, has a resistance which varies with illumination.

The strain gauge, invented by Edward E. Simmons and Arthur C. Ruge in 1938, is a type of resistor that changes value with applied strain. A single resistor may be used, or a pair (half bridge), or four resistors connected in a Wheatstone bridge configuration. The strain resistor is bonded with adhesive to an object that will be subjected to mechanical strain. With the strain gauge and a filter, amplifier, and analog/digital converter, the strain on an object can be measured.

A related but more recent invention uses a Quantum Tunnelling Composite to sense mechanical stress. It passes a current whose magnitude can vary by a factor of 10^{12} in response to changes in applied pressure.

Measurement

The value of a resistor can be measured with an ohmmeter, which may be one function of a multimeter. Usually, probes on the ends of test leads connect to the resistor. A simple

ohmmeter may apply a voltage from a battery across the unknown resistor (with an internal resistor of a known value in series) producing a current which drives a meter movement. The current flow, in accordance with Ohm's Law, is inversely proportional to the sum of the internal resistance and the resistor being tested, resulting in an analog meter scale which is very non-linear, calibrated from infinity to 0 ohms. A digital multimeter, using active electronics, may instead pass a specified current through the test resistance. The voltage generated across the test resistance in that case is linearly proportional to its resistance, which is measured and displayed. In either case the low-resistance ranges of the meter pass much more current through the test leads than do high-resistance ranges, in order for the voltages present to be at reasonable levels (generally below 10 volts) but still measurable.

Measuring low-value resistors, such as fractional-ohm resistors, with acceptable accuracy requires four-terminal connections. One pair of terminals applies a known, calibrated current to the resistor, while the other pair senses the voltage drop across the resistor. Some laboratory quality ohmmeters, especially milliohmmeters, and even some of the better digital multimeters sense using four input terminals for this purpose, which may be used with special test leads. Each of the two so-called Kelvin clips has a pair of jaws insulated from each other. One side of each clip applies the measuring current, while the other connections are only to sense the voltage drop. The resistance is again calculated using Ohm's Law as the measured voltage divided by the applied current.

Standards

Production resistors

Resistor characteristics are quantified and reported using various national standards. In the US, MIL-STD-202 contains the relevant test methods to which other standards refer.

There are various standards specifying properties of resistors for use in equipment:

- BS 1852
- EIA-RS-279
- MIL-PRF-26
- MIL-PRF-39007 (Fixed Power, established reliability)
- MIL-PRF-55342 (Surface-mount thick and thin film)
- MIL-PRF-914
- MIL-R-11
- MIL-R-39017 (Fixed, General Purpose, Established Reliability)
- MIL-PRF-32159 (zero ohm jumpers)

There are other United States military procurement MIL-R- standards.

Resistance standards

The primary standard for resistance, the "mercury ohm" was initially defined in 1884 in as a column of mercury 106 mm long and 1 square millimeter in cross-section, at 0 degrees Celsius. Difficulties in precisely measuring the physical constants to replicate this standard result in variations of as much as 30 ppm. From 1900 the mercury ohm was replaced with a precision machined plate of manganin. Since 1990 the international resistance standard has been based on the quantized Hall effect discovered by Klaus von Klitzing, for which he won the Nobel Prize in Physics in 1985.

Resistors of extremely high precision are manufactured for calibration and laboratory use. They may have four terminals, using one pair to carry an operating current and the other pair to measure the voltage drop; this eliminates errors caused by voltage drops across the lead resistances, because no current flows through voltage sensing leads. It is important in small value resistors (100–0.0001 ohm) where lead resistance is significant or even comparable with respect to resistance standard value.

Resistor marking

Most axial resistors use a pattern of colored stripes to indicate resistance. Surface-mount resistors are marked numerically, if they are big enough to permit marking; more-recent small sizes are impractical to mark. Cases are usually tan, brown, blue, or green, though other colors are occasionally found such as dark red or dark gray.

Early 20th century resistors, essentially uninsulated, were dipped in paint to cover their entire body for color coding. A second color of paint was applied to one end of the element, and a color dot (or band) in the middle provided the third digit. The rule was "body, tip, dot", providing two significant digits for value and the decimal multiplier, in that sequence. Default tolerance was ±20%. Closer-tolerance resistors had silver (±10%) or gold-colored (±5%) paint on the other end.

Four-band resistors

Four-band identification is the most commonly used color-coding scheme on resistors. It consists of four colored bands that are painted around the body of the resistor. The first two bands encode the first two significant digits of the resistance value, the third is a power-of-ten multiplier or number-of-zeroes, and the fourth is the tolerance accuracy, or acceptable error, of the value. The first three bands are equally spaced along the resistor; the spacing to the fourth band is wider. Sometimes a fifth band identifies the thermal coefficient, but this must be distinguished from the true 5-color system, with 3 significant digits.

For example, green-blue-yellow-red is $56\times10^4\ \Omega = 560\ \text{k}\Omega \pm 2\%$. An easier description can be as followed: the first band, green, has a value of 5 and the second band, blue, has a value of 6, and is counted as 56. The third band, yellow, has a value of 10^4, which adds

four 0's to the end, creating 560,000 Ω at ±2% tolerance accuracy. 560,000 Ω changes to 560 kΩ ±2% (as a kilo- is 10^3).

Each color corresponds to a certain digit, progressing from darker to lighter colors, as shown in the chart below.

Color	1st band	2nd band	3rd band (multiplier)	4th band (tolerance)	Temp. Coefficient
Black	0	0	$\times 10^0$		
Brown	1	1	$\times 10^1$	±1% (F)	100 ppm
Red	2	2	$\times 10^2$	±2% (G)	50 ppm
Orange	3	3	$\times 10^3$		15 ppm
Yellow	4	4	$\times 10^4$		25 ppm
Green	5	5	$\times 10^5$	±0.5% (D)	
Blue	6	6	$\times 10^6$	±0.25% (C)	
Violet	7	7	$\times 10^7$	±0.1% (B)	
Gray	8	8	$\times 10^8$	±0.05% (A)	
White	9	9	$\times 10^9$		
Gold			$\times 10^{-1}$	±5% (J)	
Silver			$\times 10^{-2}$	±10% (K)	
None				±20% (M)	

There are many mnemonics for remembering these colors.

Preferred values

Early resistors were made in more or less arbitrary round numbers; a series might have 100, 125, 150, 200, 300, etc. Resistors as manufactured are subject to a certain percentage tolerance, and it makes sense to manufacture values that correlate with the tolerance, so that the actual value of a resistor overlaps slightly with its neighbors. Wider spacing leaves gaps; narrower spacing increases manufacturing and inventory costs to provide resistors that are more or less interchangeable.

A logical scheme is to produce resistors in a range of values which increase in a geometrical progression, so that each value is greater than its predecessor by a fixed multiplier or percentage, chosen to match the tolerance of the range. For example, for a tolerance of ±20% it makes sense to have each resistor about 1.5 times its predecessor, covering a decade in 6 values. In practice the factor used is 1.4678, giving values of 1.47, 2.15, 3.16, 4.64, 6.81, 10 for the 1-10 decade (a decade is a range increasing by a factor of 10; 0.1-1 and 10-100 are other examples); these are rounded in practice to 1.5, 2.2, 3.3, 4.7, 6.8, 10; followed, of course by 15, 22, 33, … and preceded by … 0.47, 0.68, 1. This scheme has been adopted as the **E6** range of the IEC 60063 preferred number series. There are also **E12**, **E24**, **E48**, **E96** and **E192** ranges for components of ever tighter

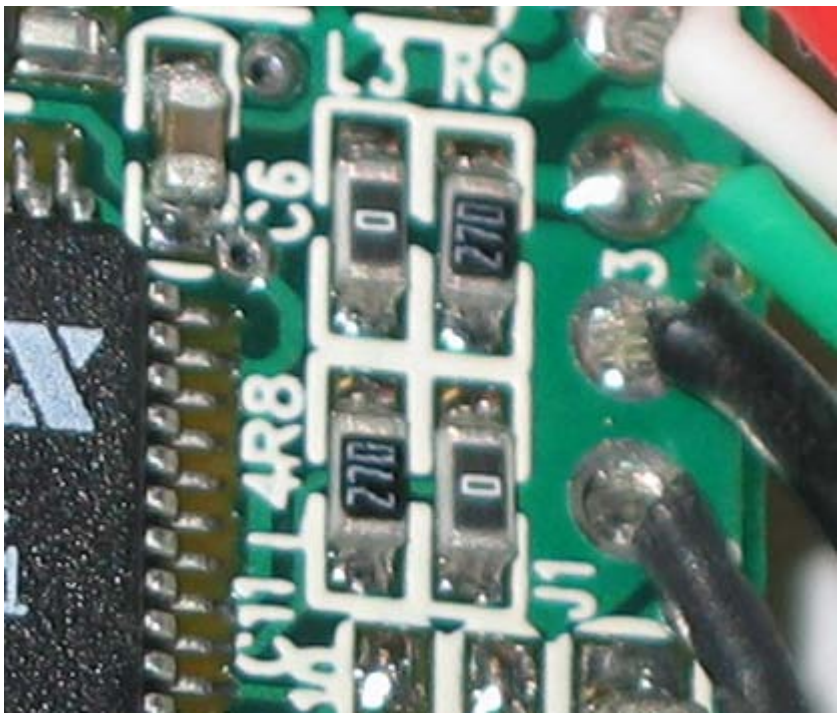
tolerance, with 12, 24, 96, and 192 different values within each decade. The actual values used are in the IEC 60063 lists of preferred numbers.

A resistor of 100 ohms ±20% would be expected to have a value between 80 and 120 ohms; its E6 neighbors are 68 (54-82) and 150 (120-180) ohms. A sensible spacing, E6 is used for ±20% components; E12 for ±10%; E24 for ±5%; E48 for ±2%, E96 for ±1%; E192 for ±0.5% or better. Resistors are manufactured in values from a few milliohms to about a gigaohm in IEC60063 ranges appropriate for their tolerance.

5-band axial resistors

5-band identification is used for higher precision (lower tolerance) resistors (1%, 0.5%, 0.25%, 0.1%), to specify a third significant digit. The first three bands represent the significant digits, the fourth is the multiplier, and the fifth is the tolerance. Five-band resistors with a gold or silver 4th band are sometimes encountered, generally on older or specialized resistors. The 4th band is the tolerance and the 5th the temperature coefficient.

SMD resistors



This image shows four surface-mount resistors (the component at the upper left is a capacitor) including two zero-ohm resistors. Zero-ohm links are often used instead of wire links, so that they can be inserted by a resistor-inserting machine. Of course, their resistance is non-zero, although quite low. *Zero* is simply a brief description of their function.

Surface mounted resistors are printed with numerical values in a code related to that used on axial resistors. Standard-tolerance surface-mount technology (SMT) resistors are marked with a three-digit code, in which the first two digits are the first two significant digits of the value and the third digit is the power of ten (the number of zeroes). For example:

334 = 33 × 10^4 ohms = 330 kilohms
222 = 22 × 10^2 ohms = 2.2 kilohms
473 = 47 × 10^3 ohms = 47 kilohms
105 = 10 × 10^5 ohms = 1.0 megohm

Resistances less than 100 ohms are written: 100, 220, 470. The final zero represents ten to the power zero, which is 1. For example:

100 = 10 × 10^0 ohm = 10 ohms
220 = 22 × 10^0 ohm = 22 ohms

Sometimes these values are marked as *10* or *22* to prevent a mistake.

Resistances less than 10 ohms have 'R' to indicate the position of the decimal point (radix point). For example:

4R7 = 4.7 ohms
R300 = 0.30 ohms
0R22 = 0.22 ohms
0R01 = 0.01 ohms

Precision resistors are marked with a four-digit code, in which the first three digits are the significant figures and the fourth is the power of ten. For example:

1001 = 100 × 10^1 ohms = 1.00 kilohm
4992 = 499 × 10^2 ohms = 49.9 kilohm
1000 = 100 × 10^0 ohm = 100 ohms

000 and *0000* sometimes appear as values on surface-mount zero-ohm links, since these have (approximately) zero resistance.

More recent surface-mount resistors are too small, physically, to permit practical markings to be applied.

Industrial type designation

Format: *[two letters]<space>[resistance value (three digit)]<nospace>[tolerance code(numerical - one digit)]*

Power Rating at 70 °C

Type No.	Power rating (watts)	MIL-R-11 Style	MIL-R-39008 Style
BB	⅛	RC05	RCR05
CB	¼	RC07	RCR07
EB	½	RC20	RCR20
GB	1	RC32	RCR32
HB	2	RC42	RCR42
GM	3	-	-
HM	4	-	-

Tolerance Code

Industrial type designation	Tolerance	MIL Designation
5	±5%	J
2	±20%	M
1	±10%	K
-	±2%	G
-	±1%	F
-	±0.5%	D
-	±0.25%	C
-	±0.1%	B

The operational temperature range distinguishes commercial grade, industrial grade and military grade components.

- Commercial grade: 0 °C to 70 °C
- Industrial grade: −40 °C to 85 °C (sometimes −25 °C to 85 °C)
- Military grade: −55 °C to 125 °C (sometimes -65 °C to 275 °C)
- Standard Grade -5 °C to 60 °C

Electrical and thermal noise

In amplifying faint signals, it is often necessary to minimize electronic noise, particularly in the first stage of amplification. As dissipative elements, even an ideal resistor will naturally produce a randomly fluctuating voltage or "noise" across its terminals. This Johnson–Nyquist noise is a fundamental noise source which depends only upon the temperature and resistance of the resistor, and is predicted by the fluctuation–dissipation theorem. Using a larger resistor produces a larger voltage noise, whereas with a smaller value of resistance there will be more current noise, assuming a given temperature. The thermal noise of a practical resistor may also be somewhat larger than the theoretical prediction and that increase is typically frequency-dependent.

However the "excess noise" of a practical resistor is an additional source of noise observed only when a current flows through it. This is specified in unit of μV/V/decade - μV of noise per volt applied across the resistor per decade of frequency. The μV/V/decade value is frequently given in dB so that a resistor with a noise index of 0 dB will exhibit 1 μV (rms) of excess noise for each volt across the resistor in each frequency decade. Excess noise is thus an example of 1/*f* noise. Thick-film and carbon composition resistors generate more excess noise than other types at low frequencies; wire-wound and thin-film resistors, though much more expensive, are often utilized for their better noise characteristics. Carbon composition resistors can exhibit a noise index of 0 dB while bulk metal foil resistors may have a noise index of -40 dB, usually making the excess noise of metal foil resistors insignificant. Thin film surface mount resistors typically have lower noise and better thermal stability than thick film surface mount resistors. However, the design engineer must read the data sheets for the family of devices to weigh the various device tradeoffs.

While not an example of "noise" per se, a resistor may act as a thermocouple, producing a small DC voltage differential across it due to the thermoelectric effect if its ends are at somewhat different temperatures. This induced DC voltage can degrade the precision of instrumentation amplifiers in particular. Such voltages appear in the junctions of the resistor leads with the circuit board and with the resistor body. Common metal film resistors show such an effect at a magnitude of about 20 μV/°C. Some carbon composition resistors can exhibit thermoelectric offsets as high as 400 μV/°C, whereas specially constructed resistors can reduce this number to 0.05 μV/°C. In applications where the thermoelectric effect may become important, care has to be taken (for example) to mount the resistors horizontally to avoid temperature gradients and to mind the air flow over the board.

Failure modes

The failure rate of resistors in a properly designed circuit is low compared to other electronic components such as semiconductors and electrolytic capacitors. Damage to resistors most often occurs due to overheating when the average power delivered to it (as computed above) greatly exceeds its ability to dissipate heat (specified by the resistor's *power rating*). This may be due to a fault external to the circuit, but is frequently caused

by the failure of another component (such as a transistor that shorts out) in the circuit connected to the resistor. Operating a resistor too close to its power rating can limit the resistor's lifespan or cause a change in its resistance over time which may or may not be noticeable. A safe design generally uses overrated resistors in power applications to avoid this danger.

When overheated, carbon-film resistors may decrease or increase in resistance. Carbon film and composition resistors can fail (open circuit) if running close to their maximum dissipation. This is also possible but less likely with metal film and wirewound resistors.

There can also be failure of resistors due to mechanical stress and adverse environmental factors including humidity. If not enclosed, wirewound resistors can corrode.

Variable resistors degrade in a different manner, typically involving poor contact between the wiper and the body of the resistance. This may be due to dirt or corrosion and is typically perceived as "crackling" as the contact resistance fluctuates; this is especially noticed as the device is adjusted. This is similar to crackling caused by poor contact in switches, and like switches, potentiometers are to some extent self-cleaning: running the wiper across the resistance may improve the contact. Potentiometers which are seldom adjusted, especially in dirty or harsh environments, are most likely to develop this problem. When self-cleaning of the contact is insufficient, improvement can usually be obtained through the use of contact cleaner (also known as "tuner cleaner") spray. The crackling noise associated with turning the shaft of a dirty potentiometer in an audio circuit (such as the volume control) is greatly accentuated when an undesired DC voltage is present, often implicating the failure of a DC blocking capacitor in the circuit.

Chapter-7

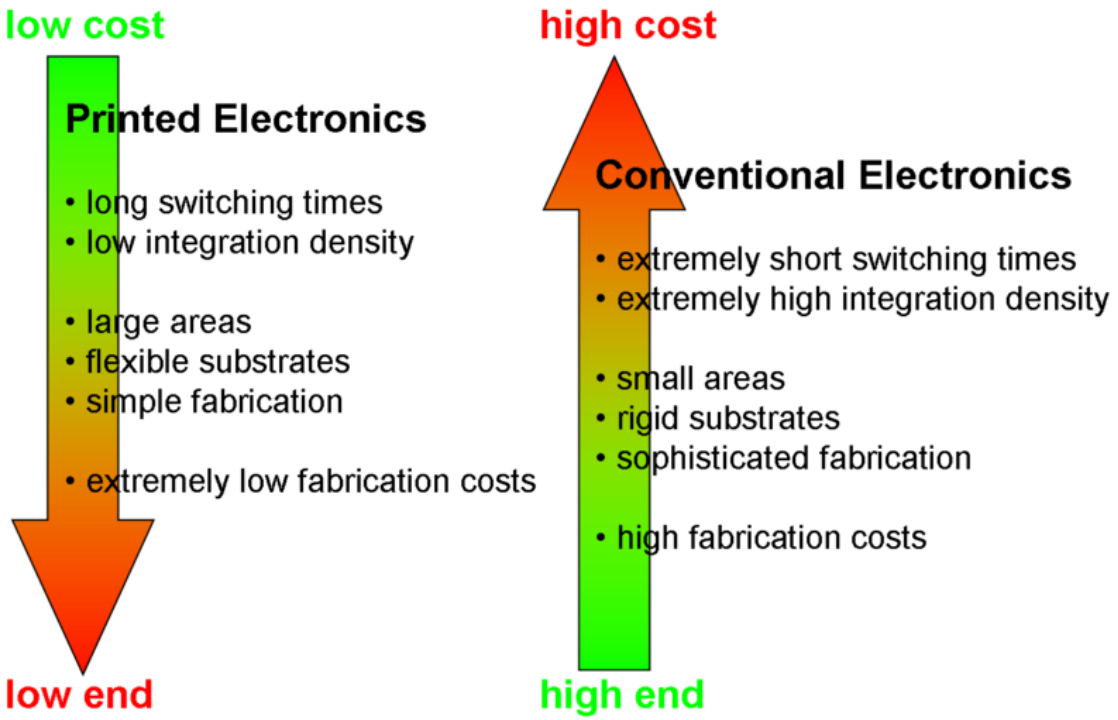
Printed Electronics

Printed electronics is a set of printing methods used to create electrical devices. Paper's rough surface and high water absorption rate has focused attention on materials such as plastic, ceramics and silicon. Printing typically uses common printing equipment, such as screen printing, flexography, gravure, offset lithography and inkjet. Electrically functional electronic or optical inks are deposited on the substrate, creating active or passive devices, such as thin film transistors or resistors. Printed electronics is expected to facilitate widespread, very low-cost, low-performance electronics for applications such as flexible displays, smart labels, decorative and animated posters, and active clothing that do not require high performance.

The term *printed electronics* is related to organic electronics or plastic electronics, in which one or more inks are composed of carbon-based compounds. These other terms refer to the ink material, which can be deposited by solution-based, vacuum-based or some other method. Printed electronics, in contrast, specifies the process, and can utilize any solution-based material, including organic semiconductors, inorganic semiconductors, metallic conductors, nanoparticles, nanotubes, etc.

For the preparation of printed electronics nearly all industrial printing methods are employed. Similar to conventional printing, printed electronics applies ink layers one atop another. so that the coherent development of printing methods and ink materials are the field's essential tasks.

The most important benefit of printing is low-cost volume fabrication. The lower cost enables use in more applications. An example is RFID-systems, which enable contactless identification in trade and transport. In some domains, such as light-emitting diodes printing does not impact performance. Printing on flexible substrates allows electronics to be placed on curved surfaces, for example, putting solar cells on vehicle roofs. More typically, conventional semiconductors justify their much higher costs by providing much higher performance.



Printed and conventional electronics as complementary technologies.

Resolution, registration, thickness, holes, materials

The maximum required resolution of structures in conventional printing is determined by the human eye. Feature sizes smaller than approximately 20 µm cannot be distinguished by the human eye and consequently exceed the capabilities of conventional printing processes. In contrast, higher resolution and smaller structures are necessary in electronics printing, because they directly affect circuit density and functionality (especially transistors). A similar requirement holds for the precision with which layers are printed on top of each other (layer to layer registration).

Control of thickness, holes, and material compatibility (wetting, adhesion, solvation) are essential, but matter in conventional printing only if the eye can detect them. Conversely, the visual impression is irrelevant.

Printing Technologies

The attraction of printing technology for the fabrication of electronics mainly results from the possibility to prepare stacks of micro-structured layers (and thereby thin-film devices) in a much more simple and cost-effective way compared to conventional electronics. Beside this, also the possibility to implement new or improved functionalities (e.g. mechanical flexibility) plays a role. The selection of used printing methods is determined by requirements concerning printed layers, by properties of printed materials as well as economic and technical considerations in terms of printed products.

Printing technologies divide between sheet-based and roll-to-roll-based approaches. Sheet-based techniques, such as inkjet and screen printing are best for low-volume, high-precision work. Gravure, offset and flexographic printing are more common for high-volume production, such as solar cells, reaching 10.000 square meters per hour (m^2/h). While offset and flexographic printing are mainly used for inorganic and organic conductors (the latter also for dielectrics), gravure printing is especially suitable for quality-sensitive layers like organic semiconductors and semiconductor/dielectric-interfaces in transistors, due to high layer quality. In connection with high resolution, is also suitable for inorganic and organic conductors. Organic field-effect transistors and integrated circuits can be prepared completely by means of mass-printing methods.

Inkjets are flexible and versatile, and can be set up with relatively low effort. Inkjets are probably the most commonly used method. However, inkjets offer lower througput of around 100 m^2/h and lower resolution (ca. 50 µm). It is well suited for low-viscosity, soluble materials like organic semiconductors. With high-viscosity materials, like organic dielectrics, and dispersed particles, like inorganic metal inks, difficulties due to nozzle clogging occur. Because ink is deposited via droplets, thickness and dispersion homogeneity is reduced. Simultaneously using many nozzles and pre-structuring the substrate allows improvements in productivity and resolution, respectively. However, in the latter case non-printing methods must be employed for the actual patterning step. Inkjet printing is preferable for organic semiconductors in organic field-effect transistors (OFETs) and organic light-emitting diodes (OLEDs), but also OFETs completely prepared by this method have been demonstrated. Frontplanes and backplanes of OLED-displays, integrated circuits, organic photovoltaic cells (OPVCs) and other devices can be prepared with inkjets.

Screen printing is appropriate for fabricating electrics and electronics on industrial scales due to its ability to produce thick layers from paste-like materials. This method can produce conducting lines from inorganic materials (e.g. for circuit boards and antennas), but also insulating and passivating layers, whereby layer thickness is more important than high resolution. Its 50 m^2/h throughput and 100 µm resolution are similar to inkjets. This versatile and comparatively simple method is used mainly for conductive and dielectric layers, but also organic semiconductors, e.g. for OPVCs, and even complete OFETs can be printed.

Other methods with similarities to printing, among them micro-contact printing and nano-imprint lithography are of interest. Here, µm- and nm-sized layers, respectively, are prepared by methods similar to stamping with soft and hard forms, respectively. Often the actual structures are prepared subtractively, e.g. by deposition of etch masks or by lift-off processes. For example electrodes for OFETs can be prepared Sporadically pad printing is used in a similar manner. Occasionally so-called transfer methods, where solid layers are transferred from a carrier to the substrate, are considered printed electronics. Electrophotography is currently not used in printed electronics.

Materials

Both organic and inorganic materials are used for printed electronics. Ink materials must be available in liquid form, for solution, dispersion or suspension. They must function as conductors, semiconductors, dielectrics, or insulators. Material costs must be fit the application.

Electronic functionality and printability can interfere with each other, mandating careful optimization. For example, a higher molecular weight in polymers enhances conductivity, but diminishes solubility. For printing, viscosity, surface tension and solid content must be tightly controlled. Cross-layer interactions such as wetting, adhesion, and solubility as well as post-deposition drying procedures affect the outcome. Additives often used in conventional printing inks are unavailable, because they often defeat electronic functionality.

Material properties largely determine the differences between printed and conventional electronics. Printable materials provide decisive advantages beside printability, such as mechanical flexibility and functional adjustment by chemical modification (e.g. light color in OLEDs).

Printed conductors offer lower conductivity and charge carrier mobility.

With a few exceptions, inorganic ink materials are dispersions of metallic micro- and nano-particles.

PMOS but not CMOS is possible in printed electronics.

Organic materials

Organic printed electronics integrates knowledge and developments from printing, electronics, chemistry, and materials science, especially from organic and polymer chemistry. Organic materials in part differ from conventional electronics in terms of structure, operation and functionality, which influences device and circuit design and optimization as well as fabrication method.

The discovery of conjugated polymers and their development into soluble materials provided the first organic ink materials. Materials from this class of polymers variously possess conducting, semiconducting, electroluminescent, photovoltaic and other properties. Other polymers are used mostly as insulators and dielectrics.

In most organic materials, hole transport is favored over electron transport. Recent studies indicate that this is a specific feature of organic semiconductor/dielectric-interfaces, which play a major role in OFETs. Therefore p-type devices should dominate over n-type devices. Durability (resistance to dispersion) and lifetime is less than conventional materials.

Organic semiconductors include the conductive polymers poly(3,4-ethylene dioxitiophene), doped with poly(styrene sulfonate), (PEDOT:PSS) and poly(aniline) (PANI). Both polymers are commercially available in different formulations and have been printed using inkjet, screen and offset printing or screen, flexo and gravure printing, respectively.

Polymer semiconductors are processed using inkjet printing, such as poly(thiopene)s like poly(3-hexylthiophene) (P3HT) and poly(9,9-dioctylfluorene co-bithiophen) (F8T2). The latter material has also been gravure printed. Different electroluminescent polymers are used with inkjet printing, as well as active materials for photovoltaics (e.g. blends of P3HT with fullerene derivatives), which in part also can be deposited using screen printing (e.g. blends of poly(phenylene vinylene) with fullerene derivatives).

Printable organic and inorganic insulators and dielectrics exist, which can be processed with different printing methods.

Inorganic materials

Inorganic electronics provides highly ordered layers and interfaces that organic and polymer materials cannot provide.

Silver nanoparticles are used with flexo, offset and inkjet. Gold particles are used with inkjet.

A.C. electroluminescent (EL) multi-color displays can cover many tens of square meters, or be incorporated in watch faces and instrument displays. They involve six to eight printed inorganic layers, including a copper doped phosphor, on a plastic film substrate.

CIGS cells can be printed directly onto molybdenum coated glass sheets.

A printed gallium arsenide germanium solar cell demonstrated 40.7% conversion efficiency, eight times that of the best organic cells, approaching the best performance of heavy silicon.

Substrates

Printed electronics allows the use of flexible substrates, which lowers production costs and allows fabrication of mechanically flexible circuits. While inkjet and screen printing typically imprint rigid substrates like glass and silicon, mass-printing methods nearly exclusively use flexible foil and paper. Poly(ethylene terephthalate)-foil (PET) is a common choice, due to its low cost and higher temperature stability. Poly(ethylene naphthalate)- (PEN) and poly(imide)-foil (PI) are alternatives. Paper's low costs and manifold applications make it an attractive substrate, however, its high roughness and large absorbency make it problematic for electronics.

Other important substrate criteria are low roughness and suitable wettability, which can be tuned pre-treatment (coating, corona). In contrast to conventional printing, high absorbency is usually disadvantageous.

Applications

Printed electronics are in use or under consideration for:

- Radio frequency identification (RFID) tags
- Monitoring
- Data storage
- Display and visual effects
- Toys
- Thin film solar cell

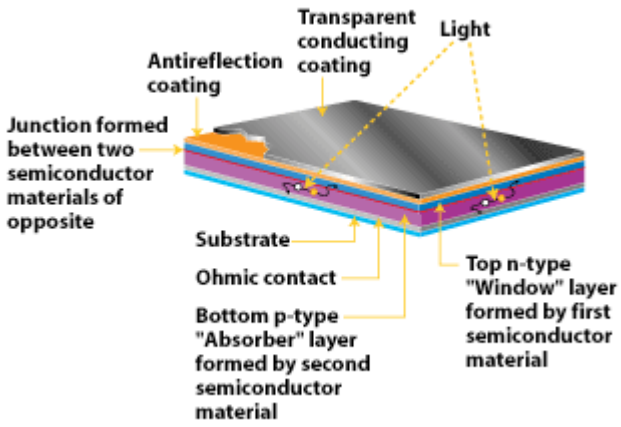
Standards development and activities

Technical standards and roadmapping initiatives are intended to facilitate value chain development (for sharing of product specifications, characterization standards, etc.) This strategy of standards development mirrors the approach used by silicon-based electronics over the past 50 years. Initiatives include:

- The IEEE Standards Association has published IEEE 1620-2004 and IEEE 1620.1-2006.
- Similar to the well-established International Technology Roadmap for Semiconductors (ITRS), the International Electronics Manufacturing Initiative (iNEMI) has published a roadmap for printed and other organic electronics.

Chapter-8

Thin Film Solar Cell



Cross-section of thin film polycrystalline solar cell.

A **thin-film solar cell** (TFSC), also called a **thin-film photovoltaic cell** (TFPV), is a solar cell that is made by depositing one or more thin layers (thin film) of photovoltaic material on a substrate. The thickness range of such a layer is wide and varies from a few nanometers to tens of micrometers.

Many different photovoltaic materials are deposited with various deposition methods on a variety of substrates. Thin-film solar cells are usually categorized according to the photovoltaic material used:

- Amorphous silicon (a-Si) and other thin-film silicon (TF-Si)
- Cadmium Telluride (CdTe)
- Copper indium gallium selenide (CIS or CIGS)
- Dye-sensitized solar cell (DSC) and other organic solar cells

$$\begin{pmatrix} \mathrm{Cu} \\ \mathrm{Ag} \\ \mathrm{Au} \end{pmatrix} \begin{pmatrix} \mathrm{Al} \\ \mathrm{Ga} \\ \mathrm{In} \end{pmatrix} \begin{pmatrix} \mathrm{S} \\ \mathrm{Se} \\ \mathrm{Te} \end{pmatrix}_2$$

Possible combinations of Group-(I, III, VI) elements in the periodic table that yield a compound showing photovoltaic effect (*Cu, Ag, Au* | *Al, Ga, In* | *S, Se, Te*).

History

Initially appearing as small strips powering hand-held calculators, thin-film PV is now available in very large modules used in sophisticated building-integrated installations and vehicle charging systems. GBI Research projects thin film production to grow 24% from 2009 levels and to reach 22,214 MW in 2020. "Expectations are that in the long-term, thin-film solar PV technology would surpass dominating conventional solar PV technology, thus enabling the long sought-after grid parity objective."

Thin-film silicon

A silicon thin-film cell uses amorphous (a-Si or a-Si:H), protocrystalline, nanocrystalline (nc-Si or nc-Si:H) or black silicon. Thin-film silicon is opposed to *wafer* (or *bulk*) silicon (monocrystalline or polycrystalline).

Design and fabrication

The silicon is mainly deposited by chemical vapor deposition, typically plasma-enhanced (PE-CVD), from silane gas and hydrogen gas. Other deposition techniques being investigated include sputtering and hot wire techniques.

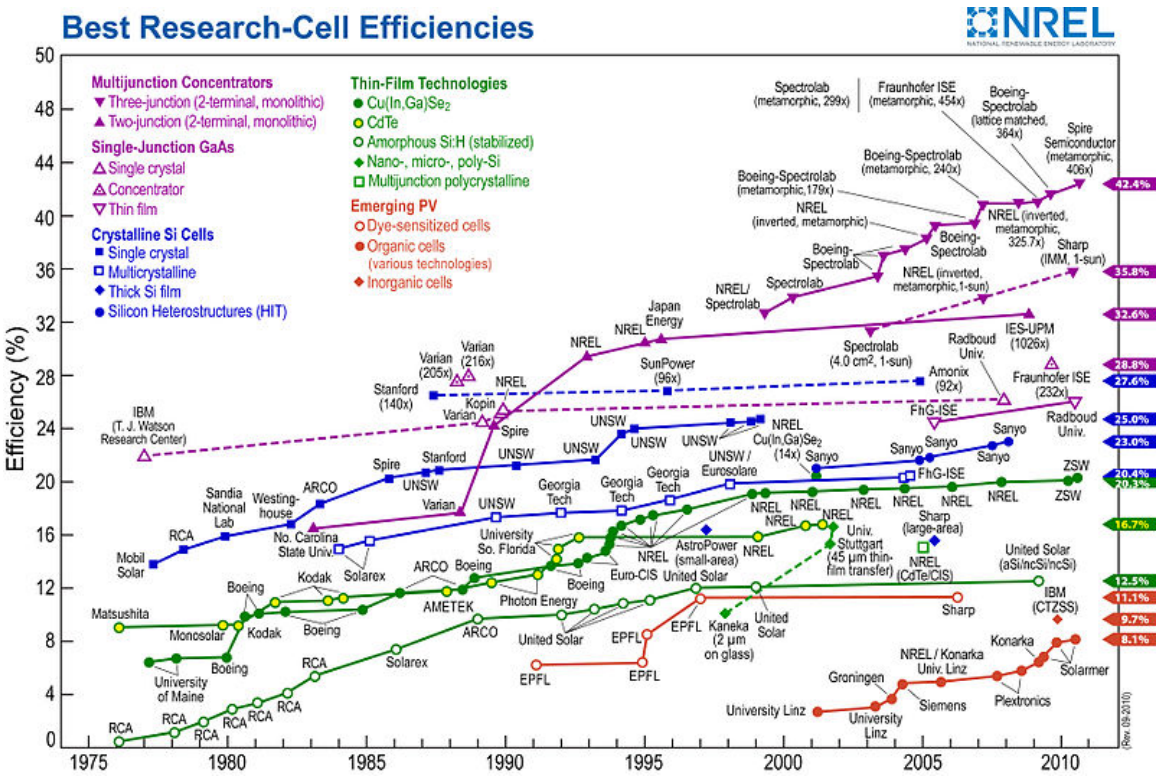
The silicon is deposited on glass, plastic or metal which has been coated with a layer of transparent conducting oxide (TCO).

A p-i-n structure is usually used, as opposed to an n-i-p structure. This is because the mobility of electrons in a-Si:H is roughly 1 or 2 orders of magnitude larger than that of holes, and thus the collection rate of electrons moving from the p- to n-type contact is better than holes moving from p- to n-type contact. Therefore, the p-type layer should be placed at the top where the light intensity is stronger, so that the majority of the charge carriers crossing the junction would be electrons.

Micromorphous silicon

Micromorphous silicon module technology combines two different types of silicon, amorphous and microcrystalline, in a top and a bottom photovoltaic cell. Use of protocrystalline silicon for the intrinsic layer has shown to optimize the open circuit voltage of an a-Si photovoltaic cell.

Efficiency



Solar Cell Efficiencies

These types of silicon present dangling and twisted bonds, which results in deep defects (energy levels in the bandgap) as well as deformation of the valence and conduction bands (band tails). The solar cells made from these materials tend to have lower energy conversion efficiency than bulk silicon (also called crystalline or wafer silicon), but are also less expensive to produce. The quantum efficiency of thin-film solar cells is also lower due to reduced number of collected charge carriers per incident photon.

Amorphous silicon has a higher bandgap (1.7 eV) than crystalline silicon (c-Si) (1.1 eV), which means it absorbs the visible part of the solar spectrum more strongly than the infrared portion of the spectrum. As nc-Si has about the same bandgap as c-Si, the nc-Si and a-Si can advantageously be combined in thin layers, creating a layered cell called a *tandem cell*. The top cell in a-Si absorbs the visible light and leaves the infrared part of the spectrum for the bottom cell in nc-Si.

Recently, solutions to overcome the limitations of thin-film silicon have been developed. Light trapping schemes where the incoming light is obliquely coupled into the silicon and the light traverses the film several times enhance the absorption of sunlight in the films. Thermal processing techniques enhance the crystallinity of the silicon and pacify electronic defects.

Building integrated photovoltaics

Thin film photovoltaic panels being installed onto a roof

Thin film solar panels are commercially available for installation onto the roofs of buildings, either applied onto the finished roof, or integrated into the roof covering. The advantage over tradition PV panels is that they are very low in weight, are not subject to wind lifting, and can be walked on (with care). The comparable disadvantages are increased cost and reduced efficiency.

A silicon thin film technology is being developed for building integrated photovoltaics (BIPV) in the form of semitransparent solar cells which can be applied as window glazing. These cells function as window tinting while generating electricity.

Organic solar cells

The Organic solar cell is another alternative to the more conventional materials used to make photovoltaics. Although a very novel technology it is promising since it offers a very low cost solution.

Efficiencies, volumes and prices

Since the invention of the first modern silicon solar cell in 1954, incremental improvements have resulted in modules capable of converting 12 to 18 percent of solar

radiation into electricity. The performance and potential of thin-film materials are high, reaching cell efficiencies of 12%-20%; prototype module efficiencies of 7%-13%; and production modules in the range of 9%. Future module efficiencies are expected to climb close to the state-of-the-art of today's best cells, or to about 10-16%.

Annual manufacturing volume in the United States has grown from about 12 megawatts (MW) per year in 2003 to more than 20 MW/yr in 2004; 40-50 MW/yr production levels are expected in 2005 with continued rapid growth in the years after that.

Costs are expected to drop to below $\$100/m^2$ in volume production, and could reach even lower levels—well under $\$50/m^2$, the DOE/NREL goal for thin films—when fully optimized. At these levels, thin-film modules will cost less than fifty cents per watt to manufacture, opening new markets such as cost-effective distributed power and utility production to thin-film electricity generation.

As crystalline silicon price rose, the production cost of silicon-based solar cell module in 2008 was at some point 4-5 times higher than that of thin film modules. Thin-film producers still enjoy in 2009 price advantage as its production cost is 20% less than that of silicon modules.It is expected that the production cost of thin-film will continue dropping (40% less than silicon), as Chinese producers are now putting more resources into R&D and partnering with manufacturing equipment suppliers

Production, cost and market

In recent years, the manufacturers of thin-film solar modules are bringing costs down and gaining in competitive strength through advanced thin film technology. However, the traditional crystalline silicon technologies will not give up their market positions for a few years because they still hold considerable development potential in terms of the cost. Efficiency of thin film solar is considerably lower and thin film solar manufacturing equipment suppliers intend to score costs of below USD 1/W, and Anwell Technologies Limited claimed that they intend to bring it down further to USD 0.5/W. Those equipment suppliers have been doing R&D for micro-morphous silicon modules since 2008. This technology represents a development based on the thin-film panels made of ordinary amorphous silicon marketed at present that brings higher cell efficiency by depositing an additional absorber layer made of micro crystalline silicon on the amorphous layer. Some equipment suppliers even claim that there will be machinery in market to manufacture these new modules at $0.70. With such potential of further development of thin film solar technology, the European Photovoltaic Industry Association (EPIA) expects that manufacturing capacities for these technologies will double to over 4GW by 2010 representing a market share of around 20%.

Installations

First Solar, the CdTe thin-film manufacturer stated that "at the end of 2007, over 300 MW of First Solar PV modules had been installed worldwide." Below is a list of several recent installations:

- Since 16 October 2008, Germany's largest thin-film pitched roof system, constructed by Riedel Recycling, has been in operation and producing solar power in Moers near Duisburg. Over eleven thousand cadmium telluride modules, from First Solar, deliver a total of 837 kW .
- First Solar recently completed a 2.4 MW rooftop installation as part of Southern California Edison program to install 250 MW of rooftop solar panels throughout Southern California over by 2013.
- First Solar announced a 7.5 MW system to be installed in Blythe, CA, where the California Public Utilities Commission has accepted a 12 ¢/kWh power purchase agreement with First Solar (after the application of all incentives).
- Construction of a 10 MW plant in the Nevada desert began in July 2008. First Solar is partnering with Sempra Generation, which will own and operate the PV power-plant, being built next to their natural gas plant.
- Stadtwerke Trier (SWT) in Trier, Germany is expected to produce over 9 GWh annually
- A 40 MW system is being installed by Juwi in Waldpolenz Solar Park, Germany. At the time of its announcement, it was both the largest planned and lowest cost PV system in the world. The price of 3.25 euros translated then (when the euro was equal to US$1.3) to $4.2 per installed watt.
- 4.8KW of thin film flexible solar panels manufactured by Uni-Solar Ovonic installed on a South Beach hurricane-prone residence in 2008..

Denver-based Conergy Americas and officials at California's South San Joaquin Irrigation District (SSJID) have installed what is believed to be the world's first single-axis solar tracking system featuring thin-film photovoltaic cells.

Chapter-9

Light-emitting Diode

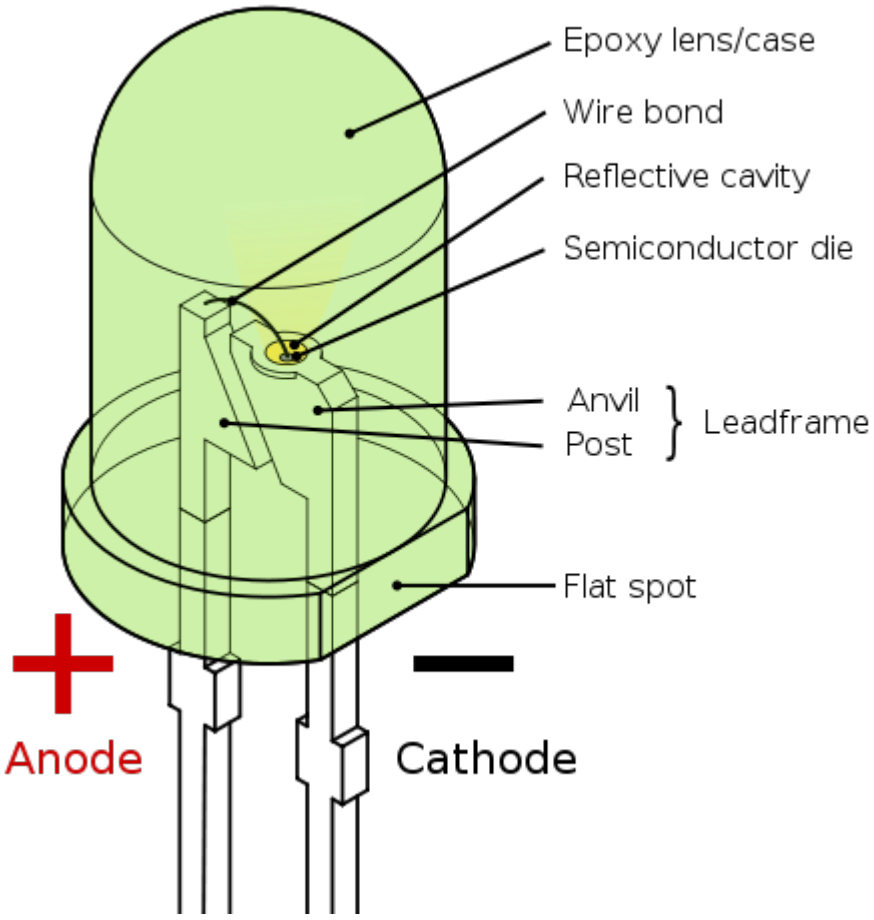
Light-emitting diode

Red, green and blue LEDs of the 5mm diffused type

Type	Passive, optoelectronic
Working principle	Electroluminescence
Invented	Nick Holonyak Jr. (1962)

Electronic symbol

Pin configuration	Anode and Cathode



Parts of an LED

LED spotlight using 38 individual diodes for mains voltage power

A **light-emitting diode** is a semiconductor light source. LEDs are used as indicator lamps in many devices and are increasingly used for other lighting. Introduced as a practical electronic component in 1962, early LEDs emitted low-intensity red light, but modern versions are available across the visible, ultraviolet and infrared wavelengths, with very high brightness.

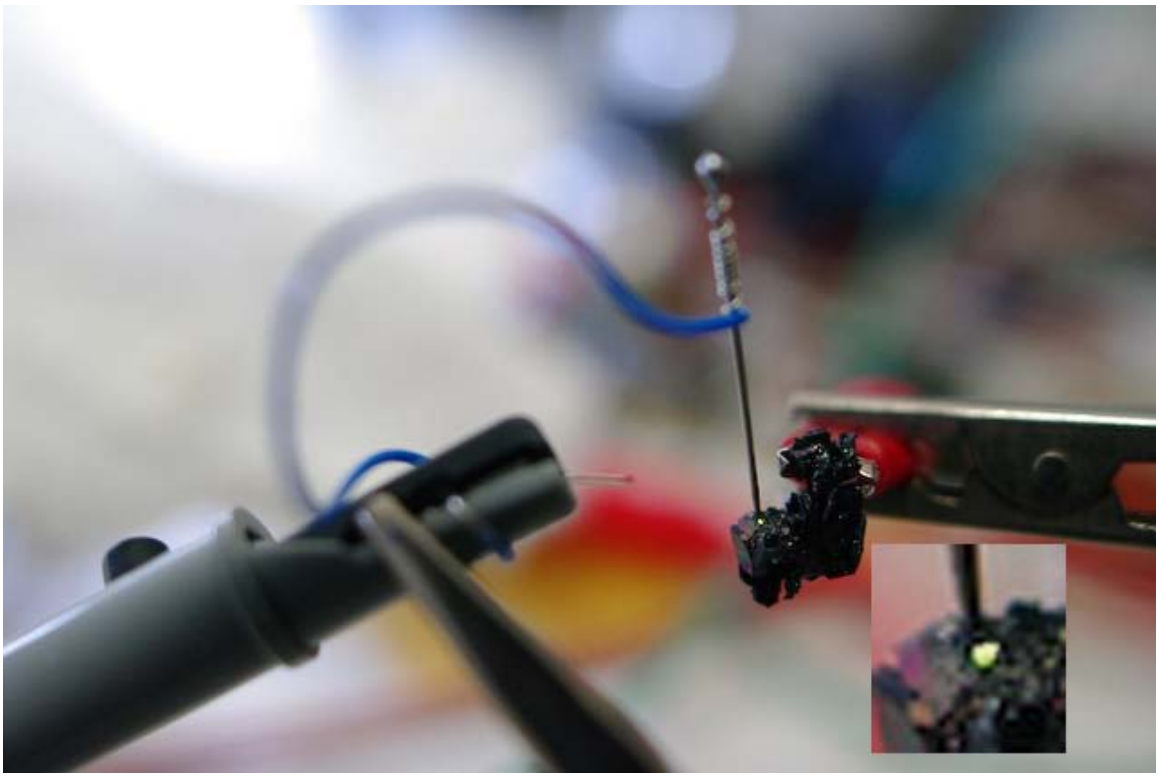
When a light-emitting diode is forward biased (switched on), electrons are able to recombine with electron holes within the device, releasing energy in the form of photons. This effect is called electroluminescence and the color of the light (corresponding to the energy of the photon) is determined by the energy gap of the semiconductor. An LED is often small in area (less than 1 mm^2), and integrated optical components may be used to shape its radiation pattern. LEDs present many advantages over incandescent light sources including lower energy consumption, longer lifetime, improved robustness, smaller size, faster switching, and greater durability and reliability. LEDs powerful enough for room lighting are relatively expensive and require more precise current and heat management than compact fluorescent lamp sources of comparable output.

Light-emitting diodes are used in applications as diverse as replacements for aviation lighting, automotive lighting (particularly brake lamps, turn signals and indicators) as well as in traffic signals. The compact size, the possibility of narrow bandwidth,

switching speed, and extreme reliability of LEDs has allowed new text and video displays and sensors to be developed, while their high switching rates are also useful in advanced communications technology. Infrared LEDs are also used in the remote control units of many commercial products including televisions, DVD players, and other domestic appliances.

History

Discoveries and early devices



Green electroluminescence from a point contact on a crystal of SiC recreates H. J. Round's original experiment from 1907.

Electroluminescence as a phenomenon was discovered in 1907 by the British experimenter H. J. Round of Marconi Labs, using a crystal of silicon carbide and a cat's-whisker detector. Russian Oleg Vladimirovich Losev reported on the creation of a first LED in 1927. His research was distributed in Russian, German and British scientific journals, but no practical use was made of the discovery for several decades. Rubin Braunstein of the Radio Corporation of America reported on infrared emission from gallium arsenide (GaAs) and other semiconductor alloys in 1955. Braunstein observed infrared emission generated by simple diode structures using gallium antimonide (GaSb), GaAs, indium phosphide (InP), and silicon-germanium (SiGe) alloys at room temperature and at 77 kelvin.

In 1961, American experimenters Robert Biard and Gary Pittman working at Texas Instruments, found that GaAs emitted infrared radiation when electric current was applied and received the patent for the infrared LED.

The first practical visible-spectrum (red) LED was developed in 1962 by Nick Holonyak Jr., while working at General Electric Company. Holonyak is seen as the "father of the light-emitting diode". M. George Craford, a former graduate student of Holonyak, invented the first yellow LED and improved the brightness of red and red-orange LEDs by a factor of ten in 1972. In 1976, T.P. Pearsall created the first high-brightness, high efficiency LEDs for optical fiber telecommunications by inventing new semiconductor materials specifically adapted to optical fiber transmission wavelengths.

Until 1968, visible and infrared LEDs were extremely costly, on the order of US $200 per unit, and so had little practical use. The Monsanto Company was the first organization to mass-produce visible LEDs, using gallium arsenide phosphide in 1968 to produce red LEDs suitable for indicators. Hewlett Packard (HP) introduced LEDs in 1968, initially using GaAsP supplied by Monsanto. The technology proved to have major uses for alphanumeric displays and was integrated into HP's early handheld calculators. In the 1970s commercially successful LED devices at under five cents each were produced by Fairchild Optoelectronics. These devices employed compound semiconductor chips fabricated with the planar process invented by Dr. Jean Hoerni at Fairchild Semiconductor. The combination of planar processing for chip fabrication and innovative packaging methods enabled the team at Fairchild led by optoelectronics pioneer Thomas Brandt to achieve the needed cost reductions. These methods continue to be used by LED producers.

Practical use

The first commercial LEDs were commonly used as replacements for incandescent and neon indicator lamps, and in seven-segment displays, first in expensive equipment such as laboratory and electronics test equipment, then later in such appliances as TVs, radios, telephones, calculators, and even watches. These red LEDs were bright enough only for use as indicators, as the light output was not enough to illuminate an area. Readouts in calculators were so small that plastic lenses were built over each digit to make them legible. Later, other colors grew widely available and also appeared in appliances and equipment. As LED materials technology grew more advanced, light output rose, while maintaining efficiency and reliability at acceptable levels. The invention and development of the high power white light LED led to use for illumination, which is fast replacing incandescent and fluorescent lighting. Most LEDs were made in the very common 5 mm T1¾ and 3 mm T1 packages, but with rising power output, it has grown increasingly necessary to shed excess heat to maintain reliability, so more complex packages have been adapted for efficient heat dissipation. Packages for state-of-the-art high power LEDs bear little resemblance to early LEDs.

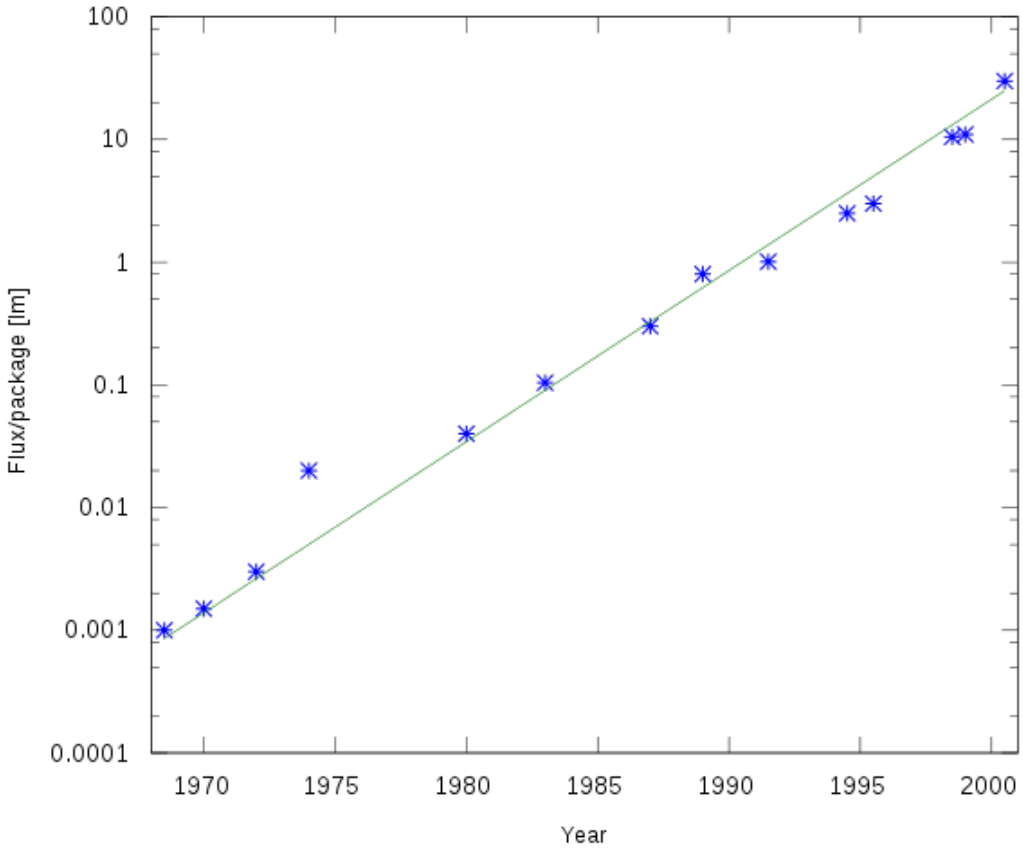


Illustration of Haitz's Law. Light output per LED as a function of production year, note the logarithmic scale on the vertical axis.

Continuing development

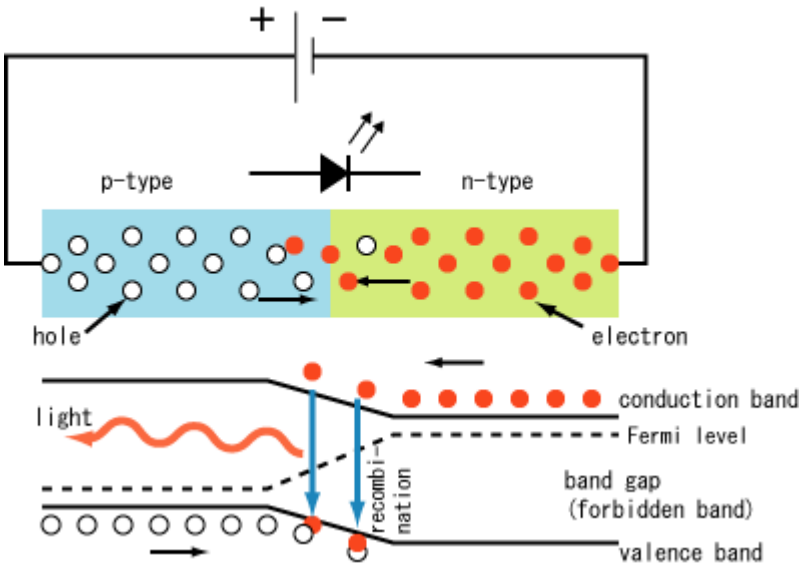
The first high-brightness blue LED was demonstrated by Shuji Nakamura of Nichia Corporation and was based on InGaN borrowing on critical developments in GaN nucleation on sapphire substrates and the demonstration of p-type doping of GaN which were developed by Isamu Akasaki and H. Amano in Nagoya. In 1995, Alberto Barbieri at the Cardiff University Laboratory (GB) investigated the efficiency and reliability of high-brightness LEDs and demonstrated a very impressive result by using a transparent contact made of indium tin oxide (ITO) on (AlGaInP/GaAs) LED. The existence of blue LEDs and high efficiency LEDs quickly led to the development of the first white LED, which employed a $Y_3Al_5O_{12}$:Ce, or "YAG", phosphor coating to mix yellow (down-converted) light with blue to produce light that appears white. Nakamura was awarded the 2006 Millennium Technology Prize for his invention.

The development of LED technology has caused their efficiency and light output to rise exponentially, with a doubling occurring about every 36 months since the 1960s, in a way similar to Moore's law. The advances are generally attributed to the parallel development of other semiconductor technologies and advances in optics and material science. This trend is normally called Haitz's Law after Dr. Roland Haitz.

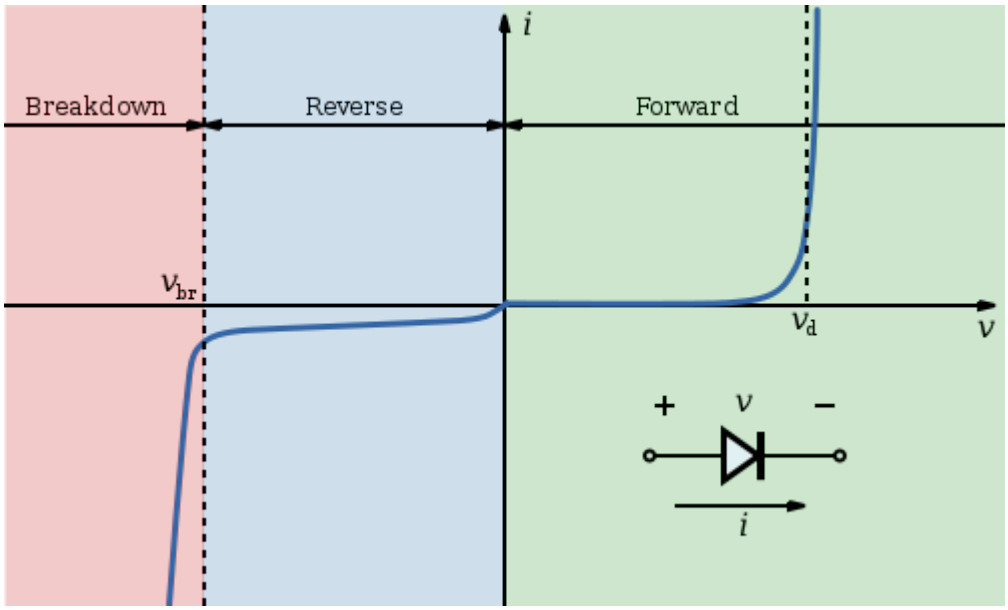
In February 2008, 300 lumens of visible light per watt luminous efficacy (not per electrical watt) and warm-light emission was achieved by using nanocrystals.

In 2009, a process for growing gallium nitride (GaN) LEDs on silicon has been reported. Epitaxy costs could be reduced by up to 90% using six-inch silicon wafers instead of two-inch sapphire wafers.

Technology



The inner workings of an LED



I-V diagram for a diode. An LED will begin to emit light when the on-voltage is exceeded. Typical on voltages are 2–3 volts

Physics

The LED consists of a chip of semiconducting material doped with impurities to create a *p-n junction*. As in other diodes, current flows easily from the p-side, or anode, to the n-side, or cathode, but not in the reverse direction. Charge-carriers—electrons and holes—flow into the junction from electrodes with different voltages. When an electron meets a hole, it falls into a lower energy level, and releases energy in the form of a photon.

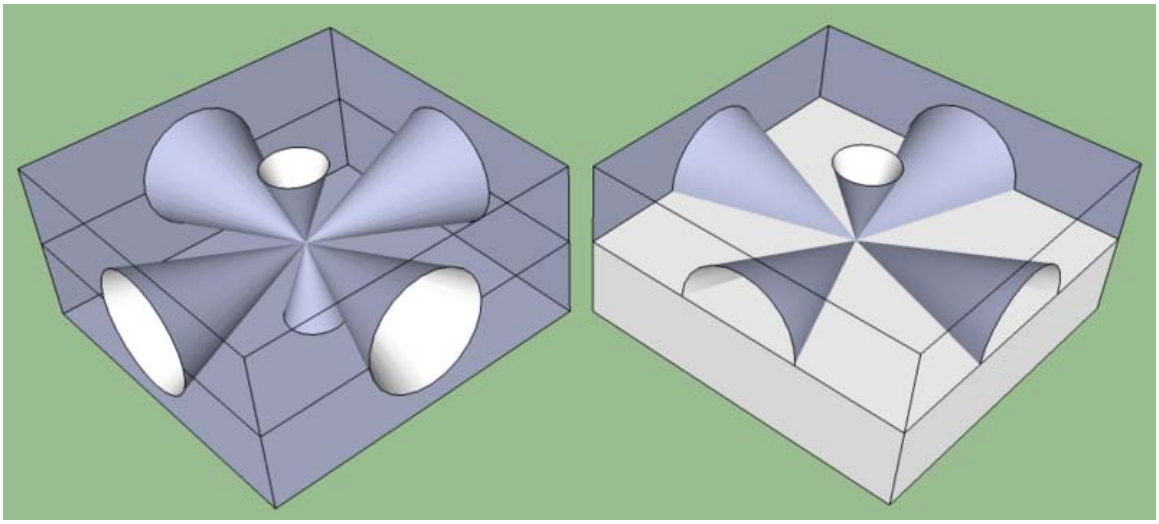
The wavelength of the light emitted, and thus its color depends on the band gap energy of the materials forming the *p-n junction*. In silicon or germanium diodes, the electrons and holes recombine by a *non-radiative transition* which produces no optical emission, because these are indirect band gap materials. The materials used for the LED have a direct band gap with energies corresponding to near-infrared, visible or near-ultraviolet light.

LED development began with infrared and red devices made with gallium arsenide. Advances in materials science have enabled making devices with ever-shorter wavelengths, emitting light in a variety of colors.

LEDs are usually built on an n-type substrate, with an electrode attached to the p-type layer deposited on its surface. P-type substrates, while less common, occur as well. Many commercial LEDs, especially GaN/InGaN, also use sapphire substrate.

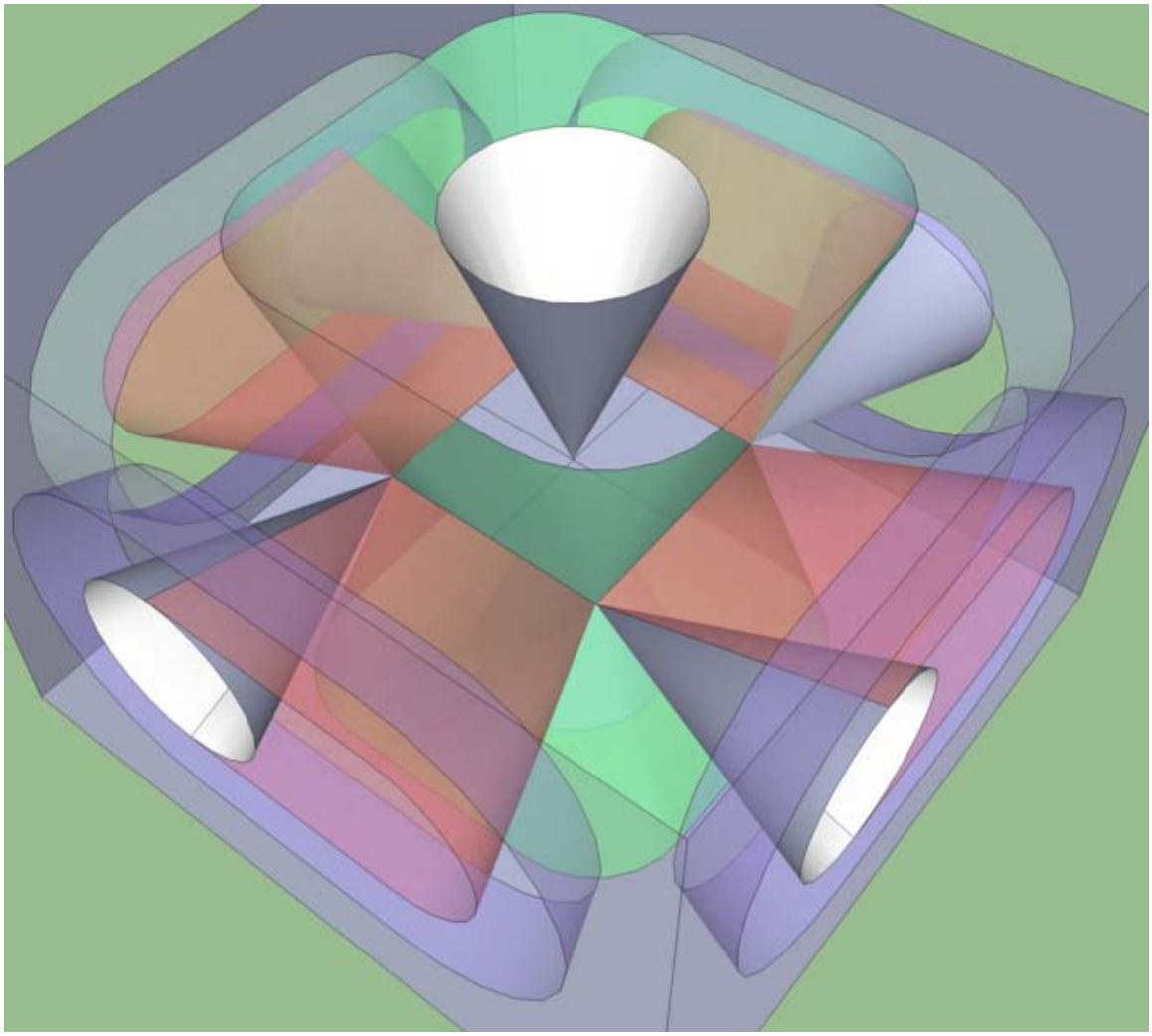
Most materials used for LED production have very high refractive indices. This means that much light will be reflected back into the material at the material/air surface interface. Thus, light extraction in LEDs is an important aspect of LED production, subject to much research and development.

Refractive Index



Idealized example of light emission cones in a semiconductor, for a single point-source emission zone. The left illustration is for a fully translucent wafer, while the right

illustration shows the half-cones formed when the bottom layer is fully opaque. The light is actually emitted equally in all directions from the point-source, so the areas between the cones shows the large amount of trapped light energy that is wasted as heat.



The light emission cones of a real LED wafer are far more complex than a single point-source light emission. Typically the light emission zone is a 2D plane between the wafers. Across this 2D plane, there is effectively a separate set of emission cones for every atom.

Drawing the billions of overlapping cones is impossible, so this is a simplified diagram showing the extents of all the emission cones combined. The larger side cones are clipped to show the interior features and reduce image complexity; they would extend to the opposite edges of the 2D emission plane.

Bare uncoated semiconductors such as silicon exhibit a very high refractive index relative to open air, which prevents passage of photons at sharp angles relative to the air-contacting surface of the semiconductor. This property affects both the light-emission

efficiency of LEDs as well as the light-absorption efficiency of photovoltaic cells. The refractive index of silicon is 4.24, while air is 1.00002926

Generally a flat-surfaced uncoated LED semiconductor chip will only emit light perpendicular to the semiconductor's surface, and a few degrees to the side, in a cone shape referred to as the *light cone*, *cone of light*, or the *escape cone*. The maximum angle of incidence is referred to as the critical angle. When this angle is exceeded photons no longer penetrate the semiconductor, but are instead reflected both internally inside the semiconductor crystal, and externally off the surface of the crystal as if it were a mirror.

Internal reflections can escape through other crystalline faces, if the incidence angle is low enough and the crystal is sufficiently transparent to not re-absorb the photon emission. But for a simple square LED with 90-degree angled surfaces on all sides, the faces all act as equal angle mirrors. In this case the light can not escape and is lost as waste heat in the crystal.

A convoluted chip surface with angled facets similar to a jewel or fresnel lens can increase light output by allowing light to be emitted perpendicular to the chip surface while far to the sides of the photon emission point.

The ideal shape of a semiconductor with maximum light output would be a microsphere with the photon emission occurring at the exact center, with electrodes penetrating to the center to contact at the emission point. All light rays emanating from the center would be perpendicular to the entire surface of the sphere, resulting in no internal reflections. A hemispherical semiconductor would also work, with the flat back-surface serving as a mirror to back-scattered photons.

Transition coatings

Many LED semiconductor chips are potted in clear or colored molded plastic shells. The plastic shell has three purposes:

1. Mounting the semiconductor chip in devices is easier to accomplish.
2. The tiny fragile electrical wiring is physically supported and protected from damage
3. The plastic acts as a refractive intermediary between the relatively high-index semiconductor and low-index open air.

The third feature helps to boost the light emission from the semiconductor by acting as a diffusing lens, allowing light to be emitted at a much higher angle of incidence from the light cone, than the bare chip is able to emit alone.

Efficiency and operational parameters

Typical indicator LEDs are designed to operate with no more than 30–60 milliwatts [mW] of electrical power. Around 1999, Philips Lumileds introduced power LEDs

capable of continuous use at one watt [W]. These LEDs used much larger semiconductor die sizes to handle the large power inputs. Also, the semiconductor dies were mounted onto metal slugs to allow for heat removal from the LED die.

One of the key advantages of LED-based lighting is its high efficiency, as measured by its light output per unit power input. White LEDs quickly matched and overtook the efficiency of standard incandescent lighting systems. In 2002, Lumileds made five-watt LEDs available with a luminous efficacy of 18–22 lumens per watt [lm/W]. For comparison, a conventional 60–100 W incandescent light bulb emits around 15 lm/W, and standard fluorescent lights emit up to 100 lm/W. A recurring problem is that efficiency falls sharply with rising current. This effect is known as droop and effectively limits the light output of a given LED, raising heating more than light output for higher current.

In September 2003, a new type of blue LED was demonstrated by the company Cree Inc. to provide 24 mW at 20 milliamperes [mA]. This produced a commercially packaged white light giving 65 lm/W at 20 mA, becoming the brightest white LED commercially available at the time, and more than four times as efficient as standard incandescents. In 2006, they demonstrated a prototype with a record white LED luminous efficacy of 131 lm/W at 20 mA. Also, Seoul Semiconductor plans for 135 lm/W by 2007 and 145 lm/W by 2008, which would be nearing an order of magnitude improvement over standard incandescents and better than even standard fluorescents. Nichia Corporation has developed a white LED with luminous efficacy of 150 lm/W at a forward current of 20 mA.

Practical general lighting needs high-power LEDs, of one watt or more. Typical operating currents for such devices begin at 350 mA.

Note that these efficiencies are for the LED chip only, held at low temperature in a lab. Lighting works at higher temperature and with drive circuit losses, so efficiencies are much lower. United States Department of Energy (DOE) testing of commercial LED lamps designed to replace incandescent lamps or CFLs showed that average efficacy was still about 46 lm/W in 2009 (tested performance ranged from 17 lm/W to 79 lm/W).

Cree issued a press release on February 3, 2010 about a laboratory prototype LED achieving 208 lumens per watt at room temperature. The correlated color temperature was reported to be 4579 K.

Lifetime and failure

Solid state devices such as LEDs are subject to very limited wear and tear if operated at low currents and at low temperatures. Many of the LEDs made in the 1970s and 1980s are still in service today. Typical lifetimes quoted are 25,000 to 100,000 hours but heat and current settings can extend or shorten this time significantly.

The most common symptom of LED (and diode laser) failure is the gradual lowering of light output and loss of efficiency. Sudden failures, although rare, can occur as well. Early red LEDs were notable for their short lifetime. With the development of high-power LEDs the devices are subjected to higher junction temperatures and higher current densities than traditional devices. This causes stress on the material and may cause early light-output degradation. To quantitatively classify lifetime in a standardized manner it has been suggested to use the terms L75 and L50 which is the time it will take a given LED to reach 75% and 50% light output respectively.

Like other lighting devices, LED performance is temperature dependent. Most manufacturers' published ratings of LEDs are for an operating temperature of 25 °C. LEDs used outdoors, such as traffic signals or in-pavement signal lights, and that are utilized in climates where the temperature within the luminaire gets very hot, could result in low signal intensities or even failure.

LED light output actually rises at colder temperatures (leveling off depending on type at around −30C). Consequently, LED technology may be a good replacement in uses such as supermarket freezer lighting and will last longer than other technologies. Because LEDs emit less heat than incandescent bulbs, they are an energy-efficient technology for uses such as freezers. However, because they emit little heat, ice and snow may build up on the LED luminaire in colder climates. This lack of waste heat generation has been observed to cause sometimes significant problems with street traffic signals and airport runway lighting in snow-prone areas, although some research has been done to try to develop heat sink technologies to transfer heat to other areas of the luminaire.

Colors and materials

Conventional LEDs are made from a variety of inorganic semiconductor materials, the following table shows the available colors with wavelength range, voltage drop and material:

Color	Wavelength (nm)	Voltage (V)	Semiconductor material
Infrared	$\lambda > 760$	$\Delta V < 1.9$	Gallium arsenide (GaAs) Aluminium gallium arsenide (AlGaAs)
Red	$610 < \lambda < 760$	$1.63 < \Delta V < 2.03$	Aluminium gallium arsenide (AlGaAs) Gallium arsenide phosphide (GaAsP) Aluminium gallium indium phosphide (AlGaInP) Gallium(III) phosphide (GaP)
Orange	$590 < \lambda < 610$	$2.03 < \Delta V < 2.10$	Gallium arsenide phosphide (GaAsP) Aluminium gallium indium phosphide (AlGaInP) Gallium(III) phosphide (GaP)
Yellow	$570 < \lambda < 590$	$2.10 < \Delta V <$	Gallium arsenide phosphide (GaAsP)

		2.18	Aluminium gallium indium phosphide (AlGaInP) Gallium(III) phosphide (GaP)
Green	$500 < \lambda < 570$	$1.9 < \Delta V < 4.0$	Indium gallium nitride (InGaN) / Gallium(III) nitride (GaN) Gallium(III) phosphide (GaP) Aluminium gallium indium phosphide (AlGaInP) Aluminium gallium phosphide (AlGaP)
Blue	$450 < \lambda < 500$	$2.48 < \Delta V < 3.7$	Zinc selenide (ZnSe) Indium gallium nitride (InGaN) Silicon carbide (SiC) as substrate Silicon (Si) as substrate — (under development)
Violet	$400 < \lambda < 450$	$2.76 < \Delta V < 4.0$	Indium gallium nitride (InGaN)
Purple	multiple types	$2.48 < \Delta V < 3.7$	Dual blue/red LEDs, blue with red phosphor, or white with purple plastic
Ultraviolet	$\lambda < 400$	$3.1 < \Delta V < 4.4$	Diamond (235 nm) Boron nitride (215 nm) Aluminium nitride (AlN) (210 nm) Aluminium gallium nitride (AlGaN) Aluminium gallium indium nitride (AlGaInN) — (down to 210 nm)
White	Broad spectrum	$\Delta V = 3.5$	Blue/UV diode with yellow phosphor

Ultraviolet and blue LEDs

Blue LEDs.

Blue LEDs are based on the wide band gap semiconductors GaN (gallium nitride) and InGaN (indium gallium nitride). They can be added to existing red and green LEDs to produce the impression of white light, though white LEDs today rarely use this principle.

The first blue LEDs were made in 1971 by Jacques Pankove (inventor of the gallium nitride LED) at RCA Laboratories. These devices had too little light output to be of much practical use. In the late 1980s, key breakthroughs in GaN epitaxial growth and p-type doping ushered in the modern era of GaN-based optoelectronic devices. Building upon this foundation, in 1993 high brightness blue LEDs were demonstrated.

By the late 1990s, blue LEDs had become widely available. They have an active region consisting of one or more InGaN quantum wells sandwiched between thicker layers of GaN, called cladding layers. By varying the relative InN-GaN fraction in the InGaN quantum wells, the light emission can be varied from violet to amber. AlGaN aluminium gallium nitride of varying AlN fraction can be used to manufacture the cladding and quantum well layers for ultraviolet LEDs, but these devices have not yet reached the level of efficiency and technological maturity of the InGaN-GaN blue/green devices. If the active quantum well layers are GaN, instead of alloyed InGaN or AlGaN, the device will emit near-ultraviolet light with wavelengths around 350–370 nm. Green LEDs

manufactured from the InGaN-GaN system are far more efficient and brighter than green LEDs produced with non-nitride material systems.

With nitrides containing aluminium, most often AlGaN and AlGaInN, even shorter wavelengths are achievable. Ultraviolet LEDs in a range of wavelengths are becoming available on the market. Near-UV emitters at wavelengths around 375–395 nm are already cheap and often encountered, for example, as black light lamp replacements for inspection of anti-counterfeiting UV watermarks in some documents and paper currencies. Shorter wavelength diodes, while substantially more expensive, are commercially available for wavelengths down to 247 nm. As the photosensitivity of microorganisms approximately matches the absorption spectrum of DNA, with a peak at about 260 nm, UV LED emitting at 250–270 nm are to be expected in prospective disinfection and sterilization devices. Recent research has shown that commercially available UVA LEDs (365 nm) are already effective disinfection and sterilization devices.

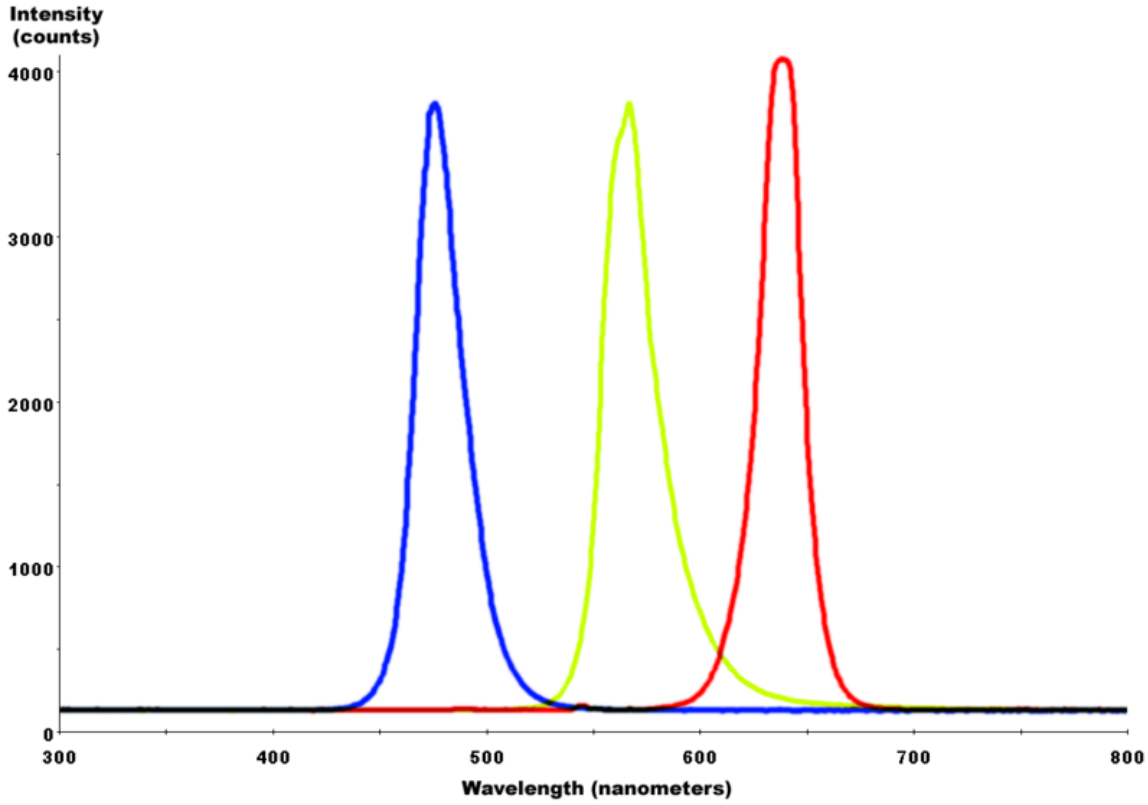
Deep-UV wavelengths were obtained in laboratories using aluminium nitride (210 nm), boron nitride (215 nm) and diamond (235 nm).

White light

There are two primary ways of producing high intensity white-light using LEDs. One is to use individual LEDs that emit three primary colors—red, green, and blue—and then mix all the colors to form white light. The other is to use a phosphor material to convert monochromatic light from a blue or UV LED to broad-spectrum white light, much in the same way a fluorescent light bulb works.

Due to metamerism, it is possible to have quite different spectra that appear white.

RGB systems



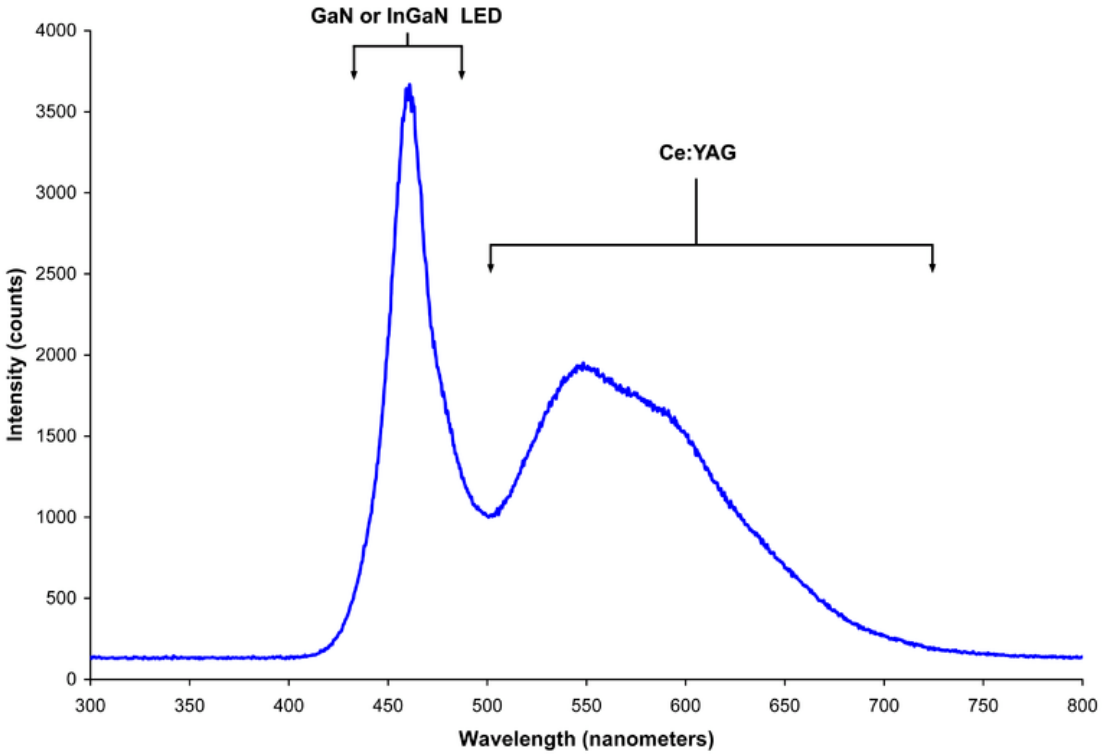
Combined spectral curves for blue, yellow-green, and high brightness red solid-state semiconductor LEDs. FWHM spectral bandwidth is approximately 24–27 nm for all three colors.

White light can be formed by mixing differently colored lights, the most common method is to use red, green and blue (RGB). Hence the method is called multi-colored white LEDs (sometimes referred to as RGB LEDs). Because these need electronic circuits to control the blending and diffusion of different colors, these are seldom used to produce white lighting. Nevertheless, this method is particularly interesting in many uses because of the flexibility of mixing different colors, and, in principle, this mechanism also has higher quantum efficiency in producing white light.

There are several types of multi-colored white LEDs: di-, tri-, and tetrachromatic white LEDs. Several key factors that play among these different methods, include color stability, color rendering capability, and luminous efficacy. Often higher efficiency will mean lower color rendering, presenting a trade off between the luminous efficiency and color rendering. For example, the dichromatic white LEDs have the best luminous efficacy (120 lm/W), but the lowest color rendering capability. Conversely, although tetrachromatic white LEDs have excellent color rendering capability, they often have poor luminous efficiency. Trichromatic white LEDs are in between, having both good luminous efficacy (>70 lm/W) and fair color rendering capability.

Multi-color LEDs offer not merely another means to form white light, but a new means to form light of different colors. Most perceivable colors can be formed by mixing different amounts of three primary colors. This allows precise dynamic color control. As more effort is devoted to investigating this method, multi-color LEDs should have profound influence on the fundamental method which we use to produce and control light color. However, before this type of LED can play a role on the market, several technical problems need solving. These include that this type of LED's emission power decays exponentially with rising temperature, resulting in a substantial change in color stability. Such problems inhibit and may preclude industrial use. Thus, many new package designs aimed at solving this problem have been proposed and their results are now being reproduced by researchers and scientists.

Phosphor-based LEDs



Spectrum of a “white” LED clearly showing blue light which is directly emitted by the GaN-based LED (peak at about 465 nm) and the more broadband Stokes-shifted light emitted by the Ce^{3+}:YAG phosphor which emits at roughly 500–700 nm.

This method involves coating an LED of one color (mostly blue LED made of InGaN) with phosphor of different colors to form white light; the resultant LEDs are called **phosphor-based white LEDs**. A fraction of the blue light undergoes the Stokes shift being transformed from shorter wavelengths to longer. Depending on the color of the original LED, phosphors of different colors can be employed. If several phosphor layers

of distinct colors are applied, the emitted spectrum is broadened, effectively raising the color rendering index (CRI) value of a given LED.

Phosphor based LEDs have a lower efficiency than normal LEDs due to the heat loss from the Stokes shift and also other phosphor-related degradation issues. However, the phosphor method is still the most popular method for making high intensity white LEDs. The design and production of a light source or light fixture using a monochrome emitter with phosphor conversion is simpler and cheaper than a complex RGB system, and the majority of high intensity white LEDs presently on the market are manufactured using phosphor light conversion.

The greatest barrier to high efficiency is the seemingly unavoidable Stokes energy loss. However, much effort is being spent on optimizing these devices to higher light output and higher operation temperatures. For instance, the efficiency can be raised by adapting better package design or by using a more suitable type of phosphor. Philips Lumileds' patented conformal coating process addresses the issue of varying phosphor thickness, giving the white LEDs a more homogeneous white light. With development ongoing, the efficiency of phosphor based LEDs generally rises with each new product announcement.

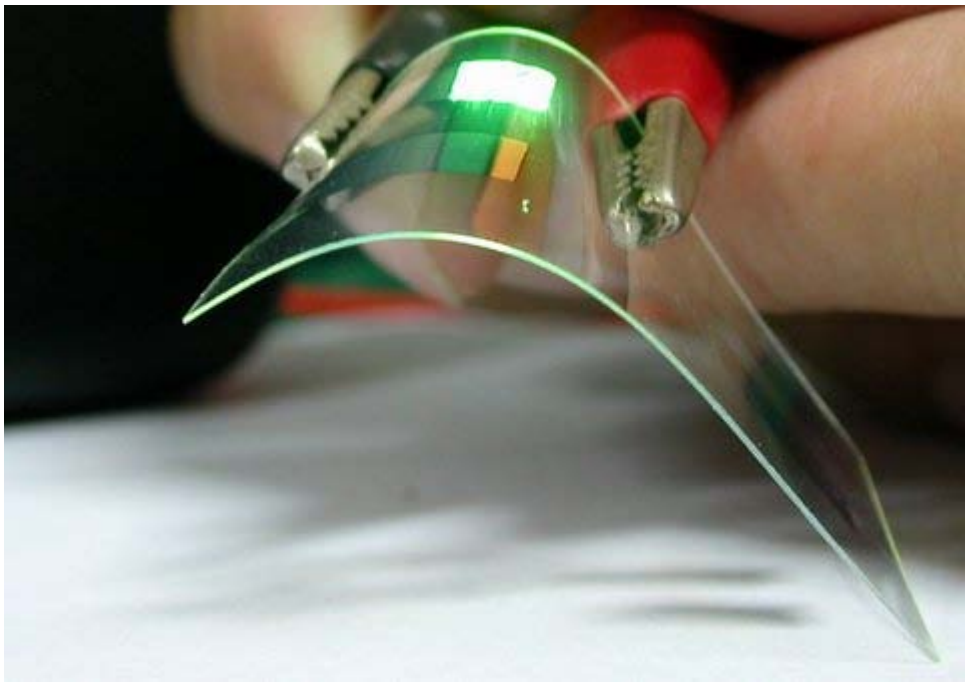
The phosphor based white LEDs encapsulate InGaN blue LEDs inside phosphor coated epoxy. A common yellow phosphor material is cerium-doped yttrium aluminium garnet (Ce^{3+}:YAG).

White LEDs can also be made by coating near ultraviolet (NUV) emitting LEDs with a mixture of high efficiency europium-based red and blue emitting phosphors plus green emitting copper and aluminium doped zinc sulfide (ZnS:Cu, Al). This is a method analogous to the way fluorescent lamps work. This method is less efficient than the blue LED with YAG:Ce phosphor, as the Stokes shift is larger, so more energy is converted to heat, but yields light with better spectral characteristics, which render color better. Due to the higher radiative output of the ultraviolet LEDs than of the blue ones, both methods offer comparable brightness. A concern is that UV light may leak from a malfunctioning light source and cause harm to human eyes or skin.

Other white LEDs

Another method used to produce experimental white light LEDs used no phosphors at all and was based on homoepitaxially grown zinc selenide (ZnSe) on a ZnSe substrate which simultaneously emitted blue light from its active region and yellow light from the substrate.

Organic light-emitting diodes (OLEDs)



Demonstration of a flexible OLED device

In an organic light emitting diode (OLED), the electroluminescent material comprising the emissive layer of the diode is an organic compound. The organic material is electrically conductive due to the delocalization of pi electrons caused by conjugation over all or part of the molecule, and the material therefore functions as an organic semiconductor. The organic materials can be small organic molecules in a crystalline phase, or polymers.

The potential advantages of OLEDs include thin, low cost displays with a low driving voltage, wide viewing angle and high contrast and colour gamut. Polymer LEDs have the added benefit of printable and flexible displays. OLEDs have been used to make visual displays for portable electronic devices such as cellphones, digital cameras, and MP3 players while possible future uses include lighting and televisions.

Quantum dot LEDs (experimental)

A new method developed by Michael Bowers, a graduate student at Vanderbilt University in Nashville, involves coating a blue LED with quantum dots that glow white in response to the blue light from the LED. This method emits a warm, yellowish-white light similar to that made by incandescent bulbs.

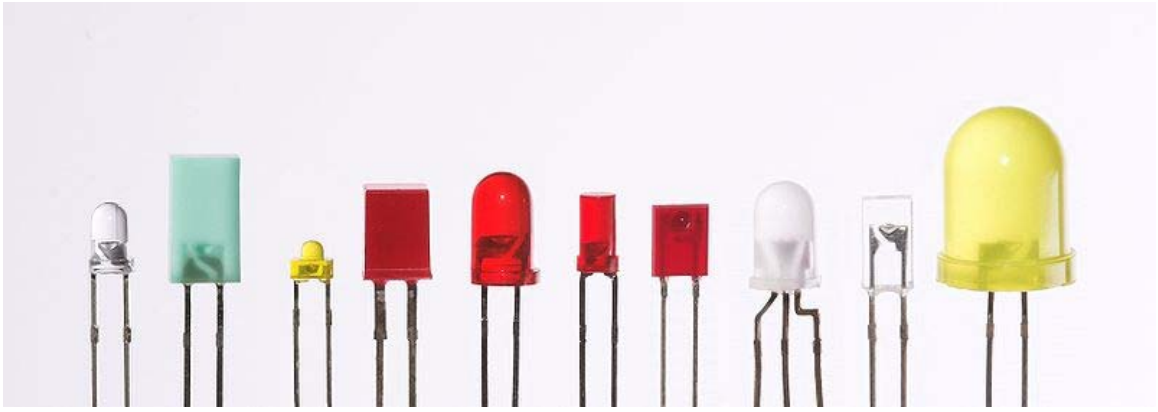
Quantum dots (QD) are semiconductor nanocrystals that possess unique optical properties. Their emission color can be tuned from the visible throughout the infrared spectrum. This allows quantum dot LEDs to create almost any color on the CIE diagram.

This provides more color options and better color rendering than white LEDs. Quantum dot LEDs are available in the same package types as traditional phosphor based LEDs.

In September 2009 Nanoco Group announced that it has signed a joint development agreement with a major Japanese electronics company under which it will design and develop quantum dots for use in light emitting diodes (LEDs) in liquid crystal display (LCD) televisions.

The major difficulty in using quantum dots based LEDs is the insufficient stability of QDs under prolonged irradiation. In February 2011 scientists at PlasmaChem GmbH could synthesize quantum dots for LED applications and build a light converter on their basis, which could efficiently convert light from blue to any other color for many hundred hours. Such QDs can be used to emit visible or near infrared light of any wavelength being excited by light with a shorter wavelength.

Types



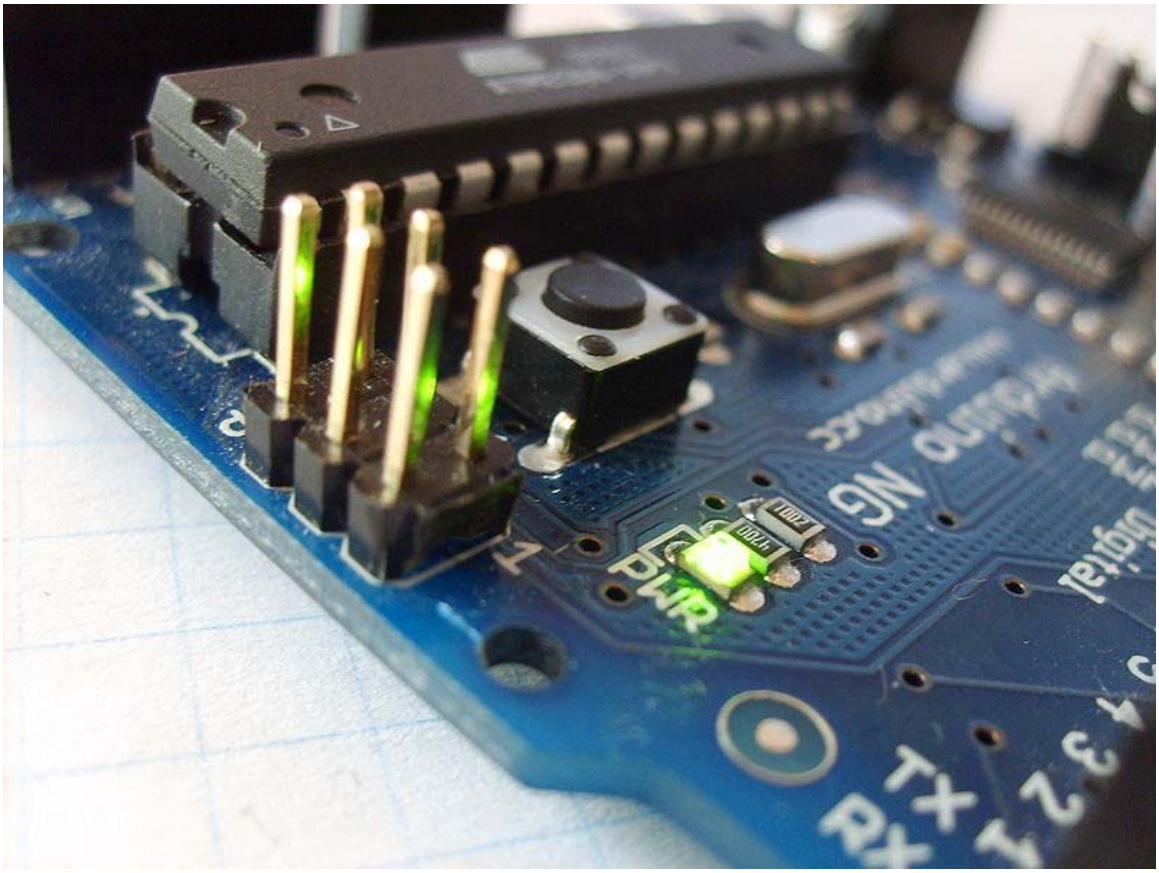
LEDs are produced in a variety of shapes and sizes. The 5 mm cylindrical package (red, fifth from the left) is the most common, estimated at 80% of world production. The color of the plastic lens is often the same as the actual color of light emitted, but not always. For instance, purple plastic is often used for infrared LEDs, and most blue devices have clear housings. There are also LEDs in SMT packages, such as those found on blinkies and on cell phone keypads (not shown).

The main types of LEDs are miniature, high power devices and custom designs such as alphanumeric or multi-color.

Miniature

Different sized LEDs. 8 mm, 5 mm and 3 mm, with a wooden match-stick for scale.

These are mostly single-die LEDs used as indicators, and they come in various-sizes from 2 mm to 8 mm, through-hole and surface mount packages. They are usually simple in design, not requiring any separate cooling body. Typical current ratings ranges from around 1 mA to above 20 mA. The small scale sets a natural upper boundary on power consumption due to heat caused by the high current density and need for heat sinking.

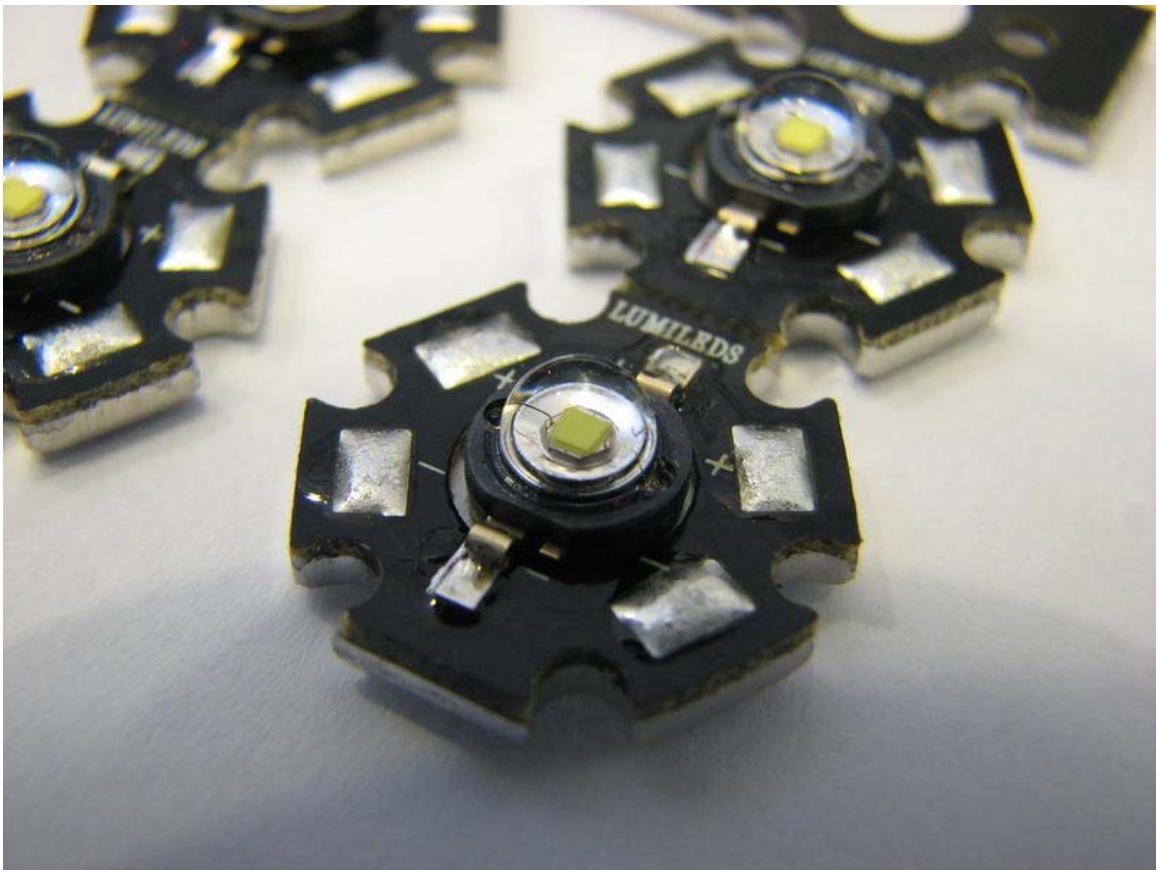


A green surface-mount LED mounted on a circuit board.

Mid-range

Medium power LEDs are often through-hole mounted and used when an output of a few lumen is needed. They sometimes have the diode mounted to four leads (two cathode leads, two anode leads) for better heat conduction and carry an integrated lens. An example of this is the Superflux package, from Philips Lumileds. These LEDs are most commonly used in light panels, emergency lighting and automotive tail-lights. Due to the larger amount of metal in the LED, they are able to handle higher currents (around 100 mA). The higher current allows for the higher light output required for tail-lights and emergency lighting.

High power



High-power light emiting diodes (Luxeon, Lumileds)

High power LEDs (HPLED) can be driven at currents from hundreds of mA to more than an ampere, compared with the tens of mA for other LEDs. Some can emit over a thousand lumens. Since overheating is destructive, the HPLEDs must be mounted on a heat sink to allow for heat dissipation. If the heat from a HPLED is not removed, the device will fail in seconds. One HPLED can often replace an incandescent bulb in a torch, or be set in an array to form a powerful LED lamp.

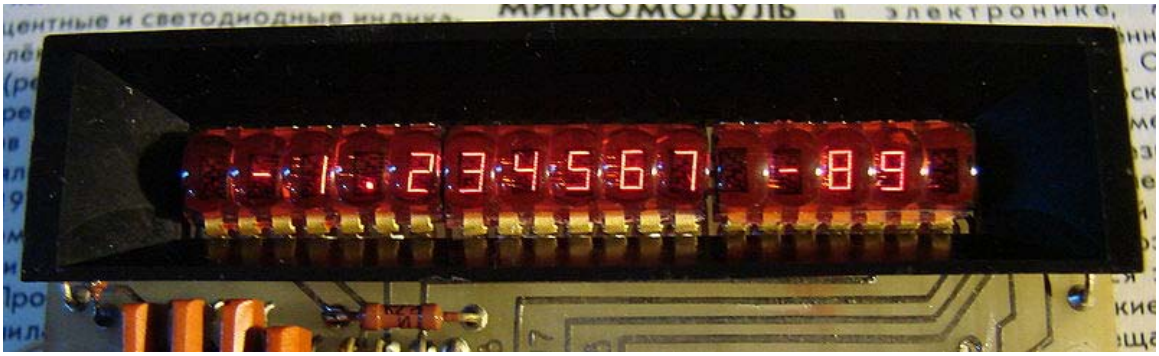
Some well-known HPLEDs in this category are the Lumileds Rebel Led, Osram Opto Semiconductors Golden Dragon and Cree X-lamp. As of September 2009 some HPLEDs manufactured by Cree Inc. now exceed 105 lm/W (e.g. the XLamp XP-G LED chip emitting Cool White light) and are being sold in lamps intended to replace incandescent, halogen, and even fluorescent lights, as LEDs grow more cost competitive.

LEDs have been developed by Seoul Semiconductor that can operate on AC power without the need for a DC converter. For each half cycle, part of the LED emits light and part is dark, and this is reversed during the next half cycle. The efficacy of this type of HPLED is typically 40 lm/W. A large number of LED elements in series may be able to operate directly from line voltage. In 2009 Seoul Semiconductor released a high DC voltage capable of being driven from AC power with a simple controlling circuit. The

low power dissipation of these LEDs affords them more flexibility than the original AC LED design.

Application-specific variations

- *Flashing LEDs* are used as attention seeking indicators without requiring external electronics. Flashing LEDs resemble standard LEDs but they contain an integrated multivibrator circuit which causes the LED to flash with a typical period of one second. In diffused lens LEDs this is visible as a small black dot. Most flashing LEDs emit light of one color, but more sophisticated devices can flash between multiple colors and even fade through a color sequence using RGB color mixing.



Calculator LED display, 1970s.

- *Bi-color LEDs* are actually two different LEDs in one case. They consist of two dies connected to the same two leads antiparallel to each other. Current flow in one direction emits one color, and current in the opposite direction emits the other color. Alternating the two colors with sufficient frequency causes the appearance of a blended third color. For example, a red/green LED operated in this fashion will color blend to emit a yellow appearance.

- *Tri-color LEDs* are two LEDs in one case, but the two LEDs are connected to separate leads so that the two LEDs can be controlled independently and lit simultaneously. A three-lead arrangement is typical with one common lead (anode or cathode).

- *RGB LEDs* contain red, green and blue emitters, generally using a four-wire connection with one common lead (anode or cathode). These LEDs can have either common positive or common negative leads. Others however, have only two leads (positive and negative) and have a built in tiny electronic control unit.

- *Alphanumeric LED displays* are available in seven-segment and starburst format. Seven-segment displays handle all numbers and a limited set of letters. Starburst displays can display all letters. Seven-segment LED displays were in widespread use in the 1970s and 1980s, but rising use of liquid crystal displays, with their

lower power needs and greater display flexibility, has reduced the popularity of numeric and alphanumeric LED displays.

Considerations for use

Power sources

The current/voltage characteristic of an LED is similar to other diodes, in that the current is dependent exponentially on the voltage. This means that a small change in voltage can cause a large change in current. If the maximum voltage rating is exceeded by a small amount, the current rating may be exceeded by a large amount, potentially damaging or destroying the LED. The typical solution is to use constant current power supplies, or driving the LED at a voltage much below the maximum rating. Since most common power sources (batteries, mains) are not constant current sources, most LED fixtures must include a power converter. However, the I/V curve of nitride-based LEDs is quite steep above the knee and gives an I_f of a few milliamperes at a V_f of 3 V, making it possible to power a nitride-based LED from a 3 V battery such as a coin cell without the need for a current limiting resistor.

Electrical polarity

As with all diodes, current flows easily from p-type to n-type material. However, no current flows and no light is emitted if a small voltage is applied in the reverse direction. If the reverse voltage grows large enough to exceed the breakdown voltage, a large current flows and the LED may be damaged. If the reverse current is sufficiently limited to avoid damage, the reverse-conducting LED is a useful noise diode.

Safety and health

The vast majority of devices containing LEDs are "safe under all conditions of normal use", and so are classified as "Class 1 LED product"/"LED Klasse 1". At present, only a few LEDs—extremely bright LEDs that also have a tightly focused viewing angle of 8° or less—could, in theory, cause temporary blindness, and so are classified as "Class 2". In general, laser safety regulations—and the "Class 1", "Class 2", etc. system—also apply to LEDs.

While LEDs have the advantage over fluorescent lamps that they do not contain mercury, they may contain other hazardous metals such as lead and arsenic. A study published in 2011 states: "According to federal standards, LEDs are not hazardous except for low-intensity red LEDs, which leached Pb [lead] at levels exceeding regulatory limits (186 mg/L; regulatory limit: 5). However, according to California regulations, excessive levels of copper (up to 3892 mg/kg; limit: 2500), Pb (up to 8103 mg/kg; limit: 1000), nickel (up to 4797 mg/kg; limit: 2000), or silver (up to 721 mg/kg; limit: 500) render all except low-intensity yellow LEDs hazardous.".

Advantages

- **Efficiency:** LEDs emit more light per watt than incandescent light bulbs. Their efficiency is not affected by shape and size, unlike fluorescent light bulbs or tubes.
- **Color:** LEDs can emit light of an intended color without using any color filters as traditional lighting methods need. This is more efficient and can lower initial costs.
- **Size:** LEDs can be very small (smaller than 2 mm^2) and are easily populated onto printed circuit boards.
- **On/Off time:** LEDs light up very quickly. A typical red indicator LED will achieve full brightness in under a microsecond. LEDs used in communications devices can have even faster response times.
- **Cycling:** LEDs are ideal for uses subject to frequent on-off cycling, unlike fluorescent lamps that fail faster when cycled often, or HID lamps that require a long time before restarting.
- **Dimming:** LEDs can very easily be dimmed either by pulse-width modulation or lowering the forward current.
- **Cool light:** In contrast to most light sources, LEDs radiate very little heat in the form of IR that can cause damage to sensitive objects or fabrics. Wasted energy is dispersed as heat through the base of the LED.
- **Slow failure:** LEDs mostly fail by dimming over time, rather than the abrupt failure of incandescent bulbs.
- **Lifetime:** LEDs can have a relatively long useful life. One report estimates 35,000 to 50,000 hours of useful life, though time to complete failure may be longer. Fluorescent tubes typically are rated at about 10,000 to 15,000 hours, depending partly on the conditions of use, and incandescent light bulbs at 1,000–2,000 hours.
- **Shock resistance:** LEDs, being solid state components, are difficult to damage with external shock, unlike fluorescent and incandescent bulbs which are fragile.
- **Focus:** The solid package of the LED can be designed to focus its light. Incandescent and fluorescent sources often require an external reflector to collect light and direct it in a usable manner.

Disadvantages

- **High initial price:** LEDs are currently more expensive, price per lumen, on an initial capital cost basis, than most conventional lighting technologies. The additional expense partially stems from the relatively low lumen output and the drive circuitry and power supplies needed.
- **Temperature dependence:** LED performance largely depends on the ambient temperature of the operating environment. Over-driving an LED in high ambient temperatures may result in overheating the LED package, eventually leading to device failure. Adequate heat sinking is needed to maintain long life. This is especially important in automotive, medical, and military uses where devices must operate over a wide range of temperatures, and need low failure rates.

- **Voltage sensitivity:** LEDs must be supplied with the voltage above the threshold and a current below the rating. This can involve series resistors or current-regulated power supplies.
- **Light quality:** Most cool-white LEDs have spectra that differ significantly from a black body radiator like the sun or an incandescent light. The spike at 460 nm and dip at 500 nm can cause the color of objects to be perceived differently under cool-white LED illumination than sunlight or incandescent sources, due to metamerism, red surfaces being rendered particularly badly by typical phosphor based cool-white LEDs. However, the color rendering properties of common fluorescent lamps are often inferior to what is now available in state-of-art white LEDs.
- **Area light source:** LEDs do not approximate a "point source" of light, but rather a lambertian distribution. So LEDs are difficult to apply to uses needing a spherical light field. LEDs cannot provide divergence below a few degrees. In contrast, lasers can emit beams with divergences of 0.2 degrees or less.
- **Blue hazard:** There is a concern that blue LEDs and cool-white LEDs are now capable of exceeding safe limits of the so-called blue-light hazard as defined in eye safety specifications such as ANSI/IESNA RP-27.1–05: Recommended Practice for Photobiological Safety for Lamp and Lamp Systems.

- **Electrical Polarity:** Unlike incandescent light bulbs, which illuminate regardless of the electrical polarity, LEDs will only light with correct electrical polarity.

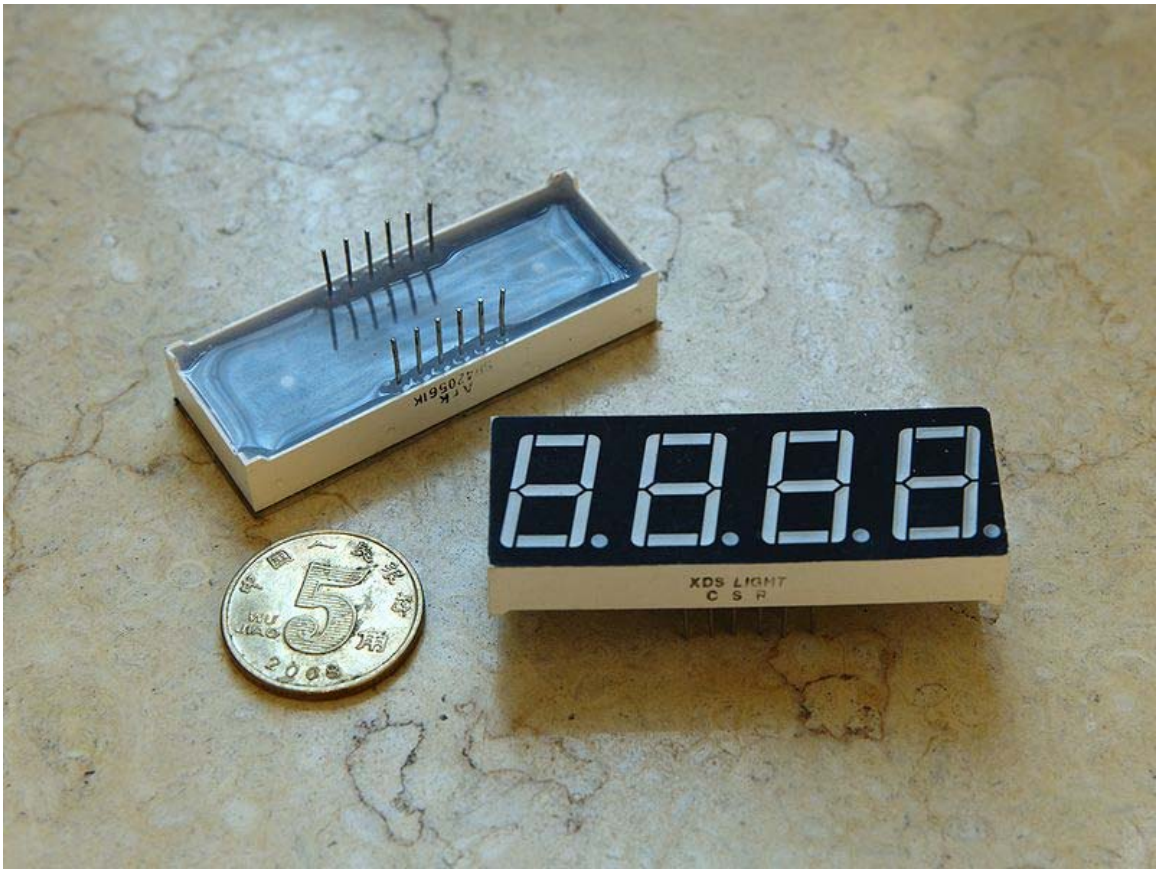
- **Blue pollution:** Because cool-white LEDs (i.e., LEDs with high color temperature) emit proportionally more blue light than conventional outdoor light sources such as high-pressure sodium vapor lamps, the strong wavelength dependence of Rayleigh scattering means that cool-white LEDs can cause more light pollution than other light sources. The International Dark-Sky Association discourages using white light sources with correlated color temperature above 3,000 K.
- **Droop**: The efficiency of LEDs tends to decrease as one increases current.

Applications

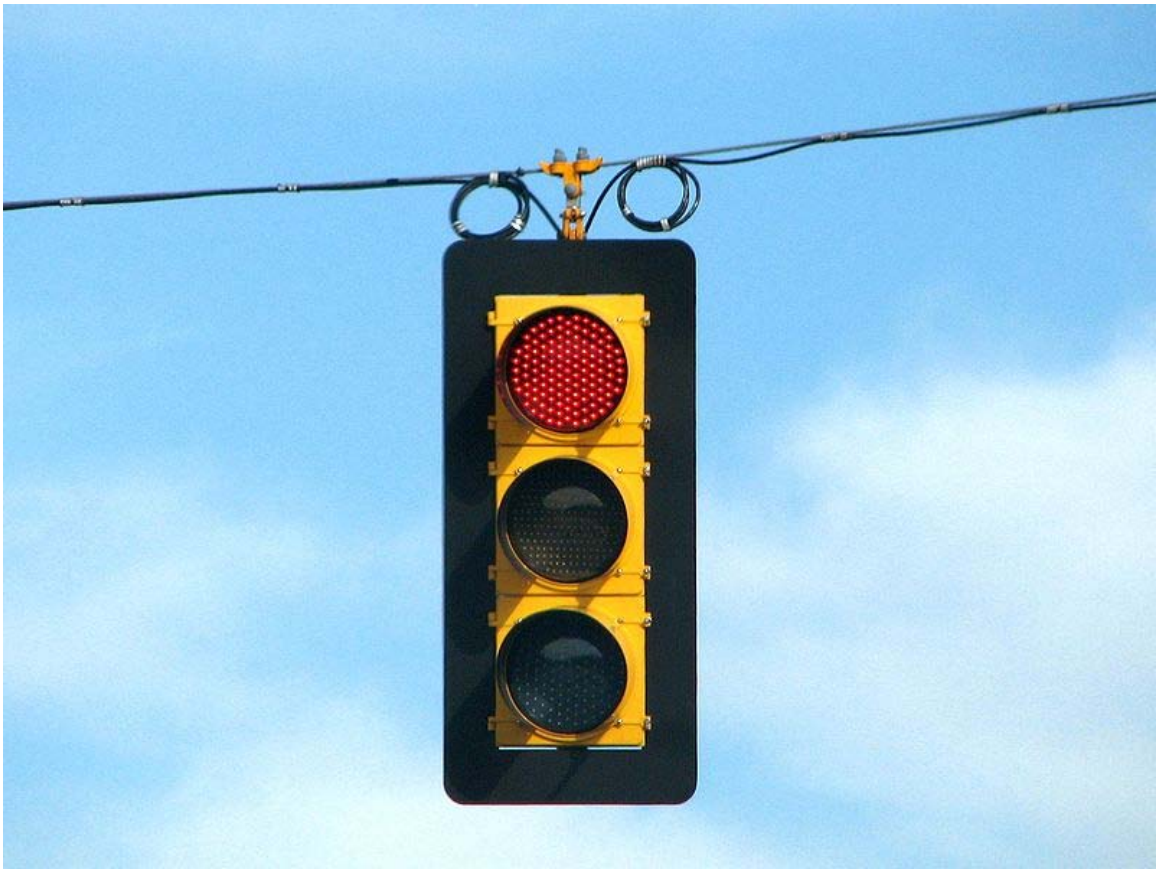
LED lighting in the aircraft cabin of an Airbus A320 Enhanced.

A large LED display behind a disc jockey.

LED destination signs on buses, one with a colored route number.



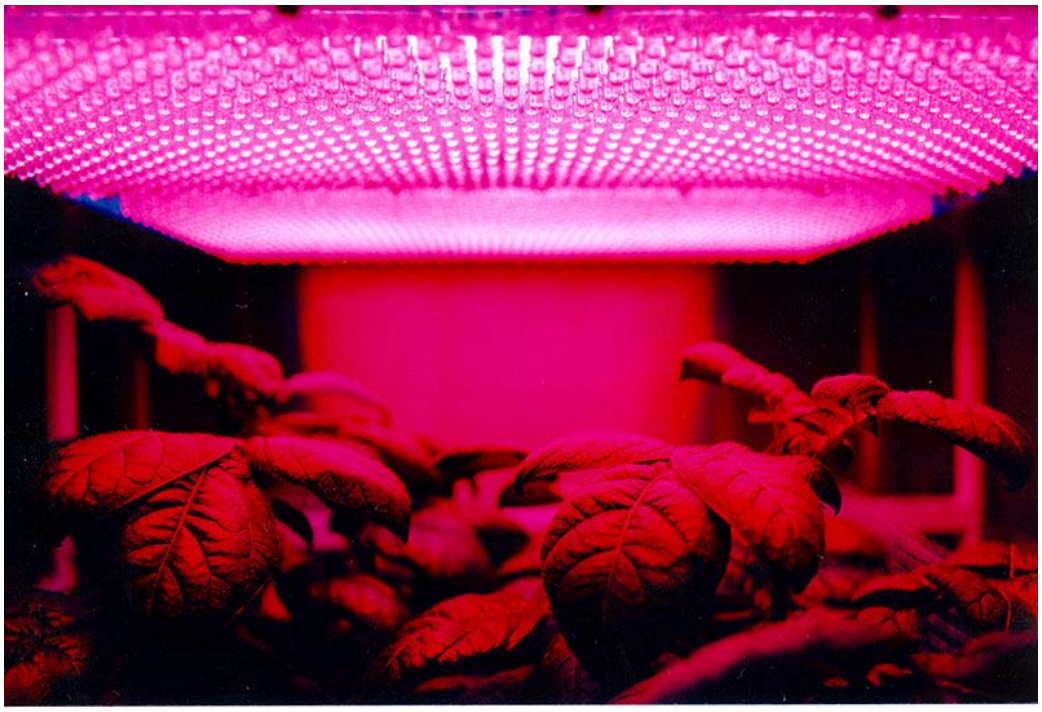
LED digital display that can display 4 digits along with points.



Traffic light using LED

Western Australia Police car using LED

LED daytime running lights of Audi A4



LED panel light source used in an experiment on plant growth. The findings of such experiments may be used to grow food in space on long duration missions.

LED Illumination.

LED uses fall into four major categories:

- Visual signals where light goes more or less directly from the source to the human eye, to convey a message or meaning.
- Illumination where light is reflected from objects to give visual response of these objects.
- Measuring and interacting with processes involving no human vision.
- Narrow band light sensors where LEDs operate in a reverse-bias mode and respond to incident light, instead of emitting light.

For more than 70 years, until the LED, practically all lighting was incandescent and fluorescent with the first fluorescent light only being commercially available after the 1939 World's Fair.

Indicators and signs

The low energy consumption, low maintenance and small size of modern LEDs has led to uses as status indicators and displays on a variety of equipment and installations. Large-area LED displays are used as stadium displays and as dynamic decorative displays. Thin, lightweight message displays are used at airports and railway stations, and as destination displays for trains, buses, trams, and ferries.

One-color light is well suited for traffic lights and signals, exit signs, emergency vehicle lighting, ships' navigation lights or lanterns (chromacity and luminance standards being set under the Convention on the International Regulations for Preventing Collisions at Sea 1972, Annex I and the CIE) and LED-based Christmas lights. In cold climates, LED traffic lights may remain snow covered. Red or yellow LEDs are used in indicator and alphanumeric displays in environments where night vision must be retained: aircraft cockpits, submarine and ship bridges, astronomy observatories, and in the field, e.g. night time animal watching and military field use.

Because of their long life and fast switching times, LEDs have been used in brake lights for cars high-mounted brake lights, trucks, and buses, and in turn signals for some time, but many vehicles now use LEDs for their rear light clusters. The use in brakes improves safety, due to a great reduction in the time needed to light fully, or faster rise time, up to 0.5 second faster than an incandescent bulb. This gives drivers behind more time to react. It is reported that at normal highway speeds, this equals one car length equivalent in increased time to react. In a dual intensity circuit (i.e., rear markers and brakes) if the LEDs are not pulsed at a fast enough frequency, they can create a phantom array, where ghost images of the LED will appear if the eyes quickly scan across the array. White LED headlamps are starting to be used. Using LEDs has styling advantages because LEDs can form much thinner lights than incandescent lamps with parabolic reflectors.

Due to the relative cheapness of low output LEDs, they are also used in many temporary uses such as glowsticks, throwies, and the photonic textile Lumalive. Artists have also used LEDs for LED art.

Weather/all-hazards radio receivers with Specific Area Message Encoding (SAME) have three LEDs: red for warnings, orange for watches, and yellow for advisories & statements whenever issued.

Lighting

With the development of high efficiency and high power LEDs it has grown possible to use LEDs in lighting and illumination. Replacement light bulbs have been made, as well as dedicated fixtures and LED lamps. LEDs are used as street lights and in other architectural lighting where color changing is used. The mechanical robustness and long lifetime is used in automotive lighting on cars, motorcycles and on bicycle lights.

LED street lights are employed on poles and in parking garages. In 2007, the Italian village Torraca was the first place to convert its entire illumination system to LEDs.

LEDs are used in aviation lighting. Airbus has used LED lighting in their Airbus A320 Enhanced since 2007, and Boeing plans its use in the 787. LEDs are also being used now in airport and heliport lighting. LED airport fixtures currently include medium-intensity runway lights, runway centerline lights, taxiway centerline & edge lights, guidance signs and obstruction lighting.

LEDs are also suitable for backlighting for LCD televisions and lightweight laptop displays and light source for DLP projectors. RGB LEDs raise the color gamut by as much as 45%. Screens for TV and computer displays can be made thinner using LEDs for backlighting.

LEDs are used increasingly commonly in aquarium lights. Particularly for reef aquariums, LED lights provide an efficient light source with less heat output to help maintain optimal aquarium temperatures. LED-based aquarium fixtures also have the advantage of being manually adjustable to emit a specific color-spectrum for ideal coloration of corals, fish, and invertebrates while optimizing photosynthetically active radiation (PAR) which raises growth and sustainability of photosynthetic life such as corals, anemones, clams, and macroalgae. These fixtures can be electronically programmed to simulate various lighting conditions throughout the day, reflecting phases of the sun and moon for a dynamic reef experience. LED fixtures typically cost up to five times as much as similarly rated fluorescent or high-intensity discharge lighting designed for reef aquariums and are not as high output to date.

The lack of IR/heat radiation makes LEDs ideal for stage lights using banks of RGB LEDs that can easily change color and decrease heating from traditional stage lighting, as well as medical lighting where IR-radiation can be harmful.

LEDs are small, durable and need little power, so they are used in hand held devices such as flashlights. LED strobe lights or camera flashes operate at a safe, low voltage, instead of the 250+ volts commonly found in xenon flashlamp-based lighting. This is especially useful in cameras on mobile phones, where space is at a premium and bulky voltage-raising circuitry is undesirable. LEDs are used for infrared illumination in night vision uses including security cameras. A ring of LEDs around a video camera, aimed forward into a retroreflective background, allows chroma keying in video productions.

LEDs are used for decorative lighting as well. Uses include but are not limited to indoor/outdoor decor, limousines, cargo trailers, conversion vans, cruise ships, RVs, boats, automobiles, and utility trucks. Decorative LED lighting can also come in the form of lighted company signage and step and aisle lighting in theaters and auditoriums.

Smart lighting

Light can be used to transmit broadband data, which is already implemented in IrDA standards using infrared LEDs. Because LEDs can cycle on and off millions of times per second, they can be wireless transmitters and access points for data transport. Lasers can also be modulated in this manner.

Sustainable lighting

Efficient lighting is needed for sustainable architecture. A 13 watt LED lamp emits 450 to 650 lumens. which is equivalent to a standard 40 watt incandescent bulb. A standard 40 W incandescent bulb has an expected lifespan of 1,000 hours while an LED can

continue to operate with reduced efficiency for more than 50,000 hours, 50 times longer than the incandescent bulb.

Energy consumption

One kilowatt-hour of electricity will cause 1.34 pounds (610 g) of CO_2 emission. Assuming the average light bulb is on for 10 hours a day, one 40-watt incandescent bulb will cause 196 pounds (89 kg) of CO_2 emission per year. The 13-watt LED equivalent will only cause 63 pounds (29 kg) of CO_2 over the same time span. A building's carbon footprint from lighting can be reduced by 68% by exchanging all incandescent bulbs for new LEDs in warm climates. In cold climates, the energy saving may be lower, since more heating is needed to compensate for the lower temperature.

Economically sustainable

LED light bulbs could be a cost-effective option for lighting a home or office space because of their very long lifetimes. Consumer use of LEDs as a replacement for conventional lighting system is currently hampered by the high cost and low efficiency of available products. 2009 DOE testing results showed an average efficacy of 35 lm/W, below that of typical CFLs, and as low as 9 lm/W, worse than standard incandescents. The high initial cost of the commercial LED bulb is due to the expensive sapphire substrate which is key to the production process. The sapphire apparatus must be coupled with a mirror-like collector to reflect light that would otherwise be wasted.

Non-visual applications

The light from LEDs can be modulated very quickly so they are used extensively in optical fiber and Free Space Optics communications. This include remote controls, such as for TVs and VCRs, where infrared LEDs are often used. Opto-isolators use an LED combined with a photodiode or phototransistor to provide a signal path with electrical isolation between two circuits. This is especially useful in medical equipment where the signals from a low-voltage sensor circuit (usually battery powered) in contact with a living organism must be electrically isolated from any possible electrical failure in a recording or monitoring device operating at potentially dangerous voltages. An optoisolator also allows information to be transferred between circuits not sharing a common ground potential.

Many sensor systems rely on light as the signal source. LEDs are often ideal as a light source due to the requirements of the sensors. LEDs are used as movement sensors, for example in optical computer mice. The Nintendo Wii's sensor bar uses infrared LEDs. In pulse oximeters for measuring oxygen saturation. Some flatbed scanners use arrays of RGB LEDs rather than the typical cold-cathode fluorescent lamp as the light source. Having independent control of three illuminated colors allows the scanner to calibrate itself for more accurate color balance, and there is no need for warm-up. Further, its sensors only need be monochromatic, since at any one time the page being scanned is only lit by one color of light. Touch sensing: Since LEDs can also be used as

photodiodes, they can be used for both photo emission and detection. This could be used in for example a touch-sensing screen that register reflected light from a finger or stylus.

Many materials and biological systems are sensitive to, or dependent on light. Grow lights use LEDs to increase photosynthesis in plants and bacteria and viruses can be removed from water and other substances using UV LEDs for sterilization. Other uses are as UV curing devices for some ink and coating methods, and in LED printers.

Plant growers are interested in LEDs because they are more energy efficient, emit less heat (can damage plants close to hot lamps), and can provide the optimum light frequency for plant growth and bloom periods compared to currently used grow lights: HPS (high pressure sodium), MH (metal halide) or CFL/low-energy. However, LEDs have not replaced these grow lights due to higher price. As mass production and LED kits develop, the LED products will become cheaper.

LEDs have also been used as a medium quality voltage reference in electronic circuits. The forward voltage drop (e.g., about 1.7 V for a normal red LED) can be used instead of a Zener diode in low-voltage regulators. Red LEDs have the flattest I/V curve above the knee. Nitride-based LEDs have a fairly steep I/V curve and are useless for this purpose. Although LED forward voltage is far more current-dependent than a good Zener, Zener diodes are not widely available below voltages of about 3 V.

Light sources for machine vision systems

Machine vision systems often require bright and homogeneous illumination, so features of interest are easier to process. LEDs are often used for this purpose, and this is likely to remain one of their major uses until price drops low enough to make signaling and illumination uses more widespread. Barcode scanners are the most common example of machine vision, and many low cost ones use red LEDs instead of lasers. Optical computer mice are also another example of LEDs in machine vision, as it is used to provide an even light source on the surface for the miniature camera within the mouse. LEDs constitute a nearly ideal light source for machine vision systems for several reasons:

The size of the illuminated field is usually comparatively small and machine vision systems are often quite expensive, so the cost of the light source is usually a minor concern. However, it might not be easy to replace a broken light source placed within complex machinery, and here the long service life of LEDs is a benefit.

LED elements tend to be small and can be placed with high density over flat or even-shaped substrates (PCBs etc.) so that bright and homogeneous sources can be designed which direct light from tightly controlled directions on inspected parts. This can often be obtained with small, low-cost lenses and diffusers, helping to achieve high light densities with control over lighting levels and homogeneity. LED sources can be shaped in several configurations (spot lights for reflective illumination; ring lights for coaxial illumination;

back lights for contour illumination; linear assemblies; flat, large format panels; dome sources for diffused, omnidirectional illumination).

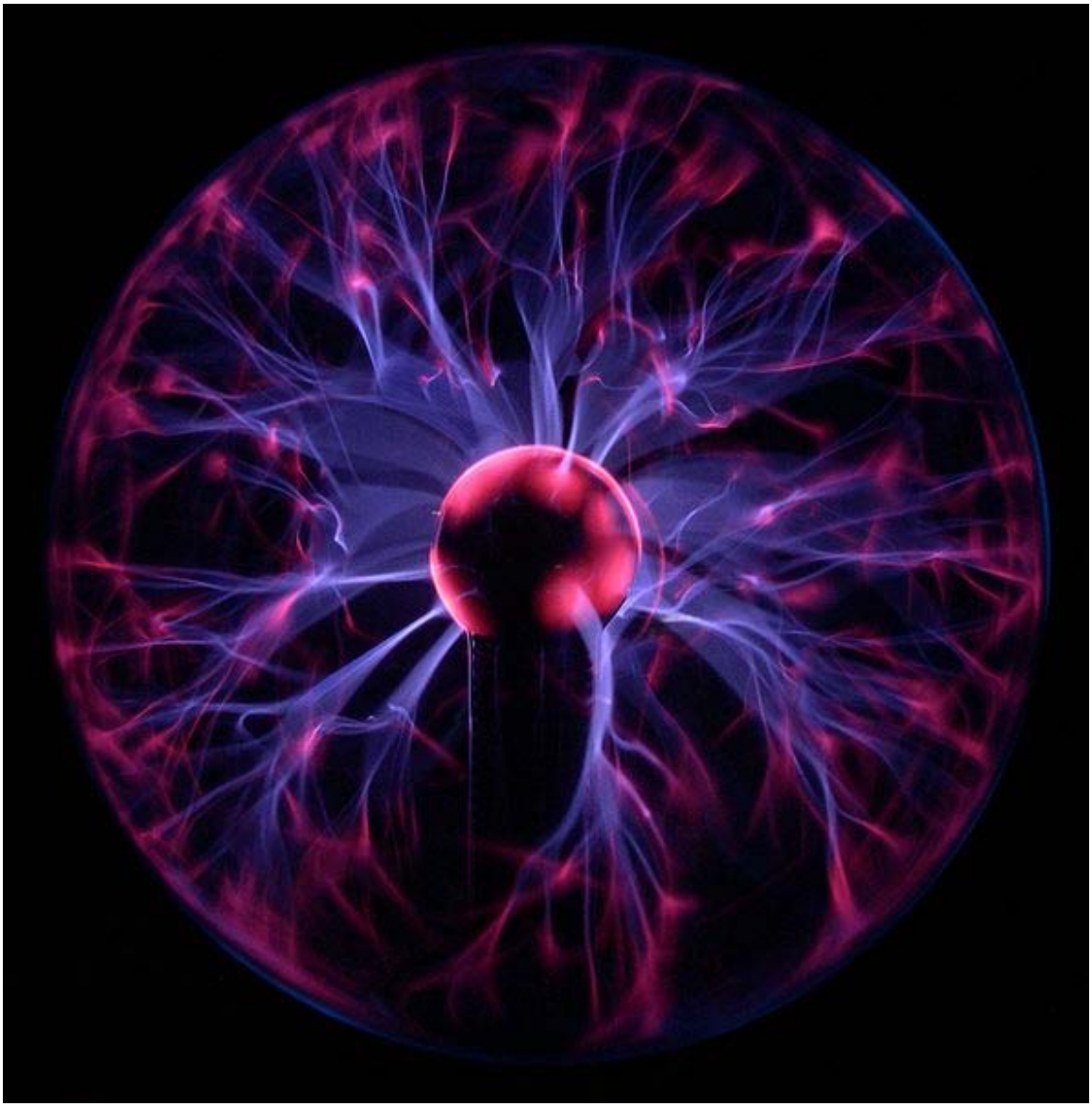
LEDs can be easily strobed (in the microsecond range and below) and synchronized with imaging. High-power LEDs are available allowing well lit images even with very short light pulses. This is often used to obtain crisp and sharp “still” images of quickly moving parts.

LEDs come in several different colors and wavelengths, allowing easy use of the best color for each need, where different color may provide better visibility of features of interest. Having a precisely known spectrum allows tightly matched filters to be used to separate informative bandwidth or to reduce disturbing effects of ambient light. LEDs usually operate at comparatively low working temperatures, simplifying heat management and dissipation. This allows using plastic lenses, filters, and diffusers. Waterproof units can also easily be designed, allowing use in harsh or wet environments (food, beverage, oil industries).

Chapter-10

Plasma Display

A typical modern plasma screen television



Ionized gases such as the ones shown here are confined to millions of tiny individual cells across the face of a plasma display, to collectively form a visual image.

A **plasma display** panel (PDP) is a type of flat panel display common to large TV displays (80 cm/30 in or larger). They are called "plasma" displays because the technology utilizes small cells containing electrically charged ionized gases, or what are in essence chambers more commonly known as fluorescent lamps.

General characteristics

A 103" plasma display panel by Panasonic

Plasma displays are bright (1,000 lux or higher for the module), have a wide color gamut, and can be produced in fairly large sizes—up to 150 inches (3.8 m) diagonally. They have a very low-luminance "dark-room" black level compared to the lighter grey of the unilluminated parts of an LCD screen (i.e. the blacks are blacker on plasmas and greyer on LCDs). LED-backlit LCD televisions have been developed to reduce this distinction. The display panel itself is about 6 cm (2.5 inches) thick, generally allowing the device's total thickness (including electronics) to be less than 10 cm (4 inches). Plasma displays use as much power per square meter as a CRT or an AMLCD television. Power consumption varies greatly with picture content, with bright scenes drawing significantly more power than darker ones - this is also true of CRTs. Typical power consumption is 400 watts for a 50-inch (127 cm) screen. 200 to 310 watts for a 50-inch (127 cm) display when set to cinema mode. Most screens are set to 'shop' mode by default, which draws at least twice the power (around 500-700 watts) of a 'home' setting of less extreme brightness. Panasonic has greatly reduced power consumption ("1/3 of 2007 models") Panasonic claims that PDPs will consume only half the power of their previous series of plasma sets to achieve the same overall brightness for a given display size. The lifetime of the latest generation of plasma displays is estimated at 100,000 hours of actual display time, or 27 years at 10 hours per day. This is the estimated time over which maximum picture brightness degrades to half the original value.

Plasma display screens are made from glass, which reflects more light than the material used to make an LCD screen. This causes glare from reflected objects in the viewing area. Companies such as Panasonic coat their newer plasma screens with an anti-glare filter material. Currently, plasma panels cannot be economically manufactured in screen sizes smaller than 32 inches. Although a few companies have been able to make plasma EDTVs this small, even fewer have made 32in plasma HDTVs. With the trend toward larger and larger displays, the 32in screen size is rapidly disappearing. Though considered bulky and thick compared to their LCD counterparts, some sets such as Panasonic's Z1 and Samsung's B860 series are as slim as one inch thick making them comparable to LCDs in this respect.

Competing display technologies include CRT, OLED, LCD, DLP, SED, LED, FED, and QLED.

Plasma display advantages and disadvantages

Advantages

- Picture quality
  - Produces deep blacks allowing for superior contrast ratio
  - Much wider viewing angles than those of LCD; images do not suffer from degradation at high angles unlike LCDs
  - No visible motion blur, thanks in large part to very high refresh rates and a faster response time, contributing to superior performance when displaying content with significant amounts of rapid motion

- Physical
  - Slim profile
  - Can be wall mounted
  - Less bulky than rear-projection televisions

Disadvantages

- Picture quality
  - Earlier generation displays were more susceptible to screen burn-in and image retention, although most recent models have a pixel orbiter that moves the entire picture faster than is noticeable to the human eye, which reduces the effect of burn-in but does not prevent it. However, turning off individual pixels does counteract screen burn-in on modern plasma displays.
  - Earlier generation displays (2006 and prior) had phosphors that lost luminosity over time, resulting in gradual decline of absolute image brightness (newer models are less susceptible to this, having lifespans exceeding 100,000 hours, far longer than older CRT technology)
  - Earlier generation (circa 2001 and earlier) models were susceptible to "large area flicker"

    - Heavier screen-door effect when compared to LCD or OLED based TVs

- Physical
    - Generally do not come in smaller sizes than 37 inches
    - Heavier than LCD due to the requirement of a glass screen to hold the gases

- Other
    - Use more electricity, on average, than an LCD TV
    - Do not work as well at high altitudes due to pressure differential between the gases inside the screen and the air pressure at altitude. It may cause a buzzing noise. Manufacturers rate their screens to indicate the altitude parameters.
    - For those who wish to listen to AM radio, or are Amateur Radio operators (Hams) or Shortwave Listeners (SWL), the Radio Frequency Interference (RFI) from these devices can be irritating or disabling.
    - Due to the strong infrared emissions inherent with the technology, standard IR repeater systems can not be used in the viewing room. A more expensive "plasma compatible" sensor must be used.

Native plasma television resolutions

Fixed-pixel displays such as plasma TVs scale the video image of each incoming signal to the native resolution of the display panel. The most common native resolutions for plasma display panels are 853×480 (EDTV), 1,366×768 or 1,920×1,080 (HDTV). As a result picture quality varies depending on the performance of the video scaling processor and the upscaling and downscaling algorithms used by each display manufacturer.

Enhanced-definition plasma television

Early plasma televisions were enhanced-definition (ED) with a native resolution of 840×480 (discontinued) or 853×480, and down-scaled their incoming high definition signals to match their native display resolution.

ED Resolutions

- 840×480
- 853×480

High-definition plasma television

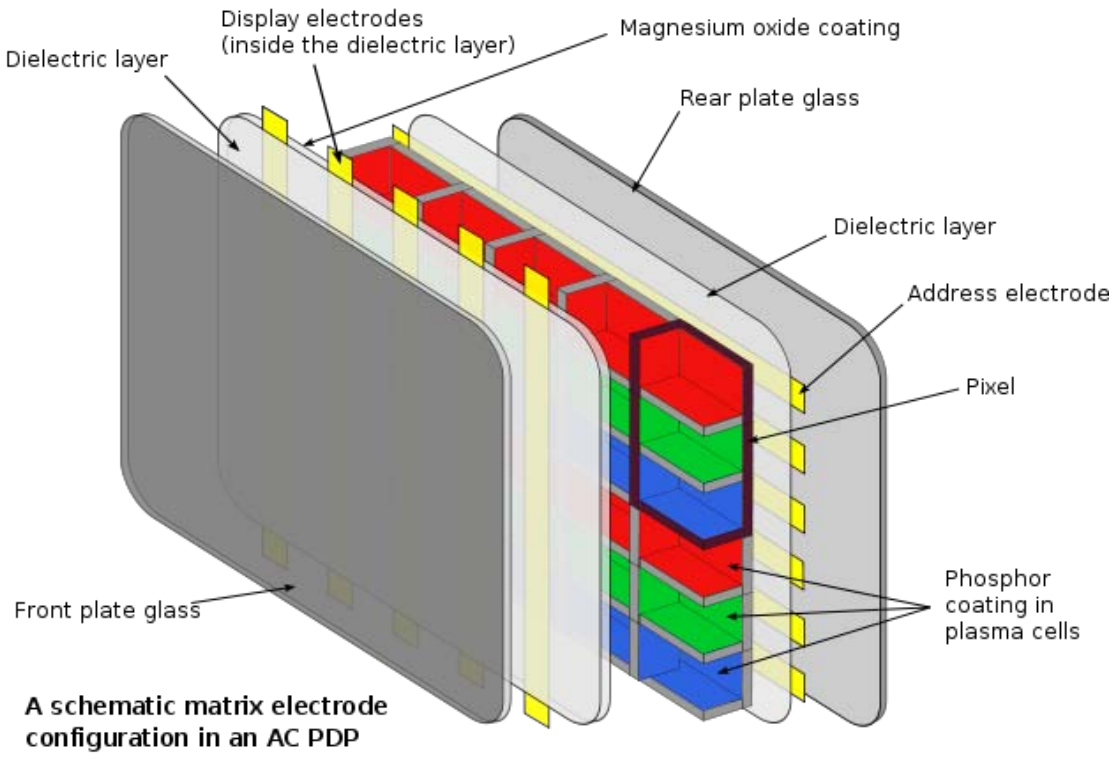
Early high-definition (HD) plasma displays had a resolution of 1024x1024 and were alternate lighting of surfaces (ALiS) panels made by Fujitsu/Hitachi. These were interlaced displays, with non-square pixels.

Modern HDTV plasma televisions usually have a resolution of 1,024×768 found on many 42 inch plasma screens, 1,280×768, 1,366×768 found on 50 in, 60 in, and 65 in plasma screens, or 1,920×1,080 found in plasma screen sizes from 42 inch to 103 inch. These displays are usually progressive displays, with square pixels, and will up-scale their incoming standard-definition signals to match their native display resolution.

HD Resolutions

- 1024×1024
- 1024×768
- 1280×768
- 1366×768
- 1280×1080
- 1920×1080

How plasma displays work



Composition of plasma display panel

A panel typically has millions of tiny cells in compartmentalized space between two panels of glass. These compartments, or "bulbs" or "cells", hold a mixture of noble gases and a minuscule amount of mercury. Just as in the fluorescent lamps over an office desk, when the mercury is vaporized and a voltage is applied across the cell, the gas in the cells

form a plasma. With flow of electricity (electrons), some of the electrons strike mercury particles as the electrons move through the plasma, momentarily increasing the energy level of the molecule until the excess energy is shed. Mercury sheds the energy as ultraviolet (UV) photons. The UV photons then strike phosphor that is painted on the inside of the cell. When the UV photon strikes a phosphor molecule, it momentarily raises the energy level of an outer orbit electron in the phosphor molecule, moving the electron from a stable to an unstable state; the electron then sheds the excess energy as a photon at a lower energy level than UV light; the lower energy photons are mostly in the infrared range but about 40% are in the visible light range. Thus the input energy is shed as mostly heat (infrared) but also as visible light. Depending on the phosphors used, different colors of visible light can be achieved. Each pixel in a plasma display is made up of three cells comprising the primary colors of visible light. Varying the voltage of the signals to the cells thus allows different perceived colors.

A plasma display panel is an array of hundreds of thousands of small, luminous cells positioned between two plates of glass. Each cell is essentially a tiny neon lamp filled with rarefied neon, xenon, and other inert gases; the cells are luminous when they are electrified through "electrodes".

The long electrodes are stripes of electrically conducting material that also lie between the glass plates, in front of and behind the cells. The "address electrodes" sit behind the cells, along the rear glass plate, and can be opaque. The transparent display electrodes are mounted in front of the cell, along the front glass plate. As can be seen in the illustration, the electrodes are covered by an insulating protective layer. Control circuitry charges the electrodes that cross paths at a cell, creating a voltage difference between front and back. Some of the atoms in the gas of a cell then lose electrons and become ionized, which creates an electrically conducting plasma of atoms, free electrons, and ions. The collisions of the flowing electrons in the plasma with the inert gas atoms leads to light emission; such light-emitting plasmas are known as glow discharges.

In a monochrome plasma panel, the gas is usually mostly neon, and the color is the characteristic orange of a neon-filled lamp (or sign). Once a glow discharge has been initiated in a cell, it can be maintained by applying a low-level voltage between all the horizontal and vertical electrodes–even after the ionizing voltage is removed. To erase a cell all voltage is removed from a pair of electrodes. This type of panel has inherent memory. A small amount of nitrogen is added to the neon to increase hysteresis.

In color panels, the back of each cell is coated with a phosphor. The ultraviolet photons emitted by the plasma excite these phosphors, which give off visible light with colors determined by the phosphor materials. This aspect is comparable to fluorescent lamps and to the neon signs that use colored phosphors.

Every pixel is made up of three separate subpixel cells, each with different colored phosphors. One subpixel has a red light phosphor, one subpixel has a green light phosphor and one subpixel has a blue light phosphor. These colors blend together to create the overall color of the pixel, the same as a triad of a shadow mask CRT or color

LCD. Plasma panels use pulse-width modulation (PWM) to control brightness: by varying the pulses of current flowing through the different cells thousands of times per second, the control system can increase or decrease the intensity of each subpixel color to create billions of different combinations of red, green and blue. In this way, the control system can produce most of the visible colors. Plasma displays use the same phosphors as CRTs, which accounts for the extremely accurate color reproduction when viewing television or computer video images (which use an RGB color system designed for CRT display technology).

Plasma displays should not be confused with liquid crystal displays (LCDs), another lightweight flat-screen display using very different technology. LCDs may use one or two large fluorescent lamps as a backlight source, but the different colors are controlled by LCD units, which in effect behave as gates that allow or block the passage of light from the backlight to red, green, or blue paint on the front of the LCD panel.

Contrast ratio

Contrast ratio is the difference between the brightest and darkest parts of an image, measured in discrete steps, at any given moment. Generally, the higher the contrast ratio, the more realistic the image is (though the "realism" of an image depends on many factors including color accuracy, luminance linearity, and spatial linearity.) Contrast ratios for plasma displays are often advertised as high as 5,000,000:1. On the surface, this is a significant advantage of plasma over most other current display technologies, a notable exception being organic light-emitting diode. Although there are no industry-wide guidelines for reporting contrast ratio, most manufacturers follow either the ANSI standard or perform a full-on-full-off test. The ANSI standard uses a checkered test pattern whereby the darkest blacks and the lightest whites are simultaneously measured, yielding the most accurate "real-world" ratings. In contrast, a full-on-full-off test measures the ratio using a pure black screen and a pure white screen, which gives higher values but does not represent a typical viewing scenario. Some displays, using many different technologies, have some "leakage" of light, through either optical or electronic means, from lit pixels to adjacent pixels so that dark pixels that are near bright ones appear less dark than they do during a full-off display. Manufacturers can further artificially improve the reported contrast ratio by increasing the contrast and brightness settings to achieve the highest test values. However, a contrast ratio generated by this method is misleading, as content would be essentially unwatchable at such settings.

Plasma is often cited as having better (i.e. darker) black levels (and higher contrast ratios), although both plasma and LCD each have their own technological challenges.

Each cell on a plasma display has to be precharged before it is due to be illuminated (otherwise the cell would not respond quickly enough) and this precharging means the cells cannot achieve a true black, whereas an LED backlit LCD panel can actually turn off parts of the screen. Some manufacturers have worked hard to reduce the precharge and the associated background glow, to the point where black levels on modern plasmas are starting to rival CRT. With LCD technology, black pixels are generated by a light

polarization method; many panels are unable to completely block the underlying backlight. However, more recent LCD panels (particularly those using white LED illumination) can compensate by automatically reducing the backlighting on darker scenes, though this method — analogous to the strategy of noise reduction on analog audio tape — obviously cannot be used in high-contrast scenes, leaving some light showing from black parts of an image with bright parts, such as (at the extreme) a solid black screen with one fine intense bright line. This is called a "halo" effect which has been almost completely minimized on newer LED backlit LCD's with local dimming. Edgelit models cannot compete with this as the light is reflected via a light funnell to distribute the light behind the panel.

Screen burn-in

An example of a plasma display that has suffered severe burn-in from stationary text

Image burn-in occurs on CRTs and plasma panels when the same picture is displayed for long periods of time. This causes the phosphors to overheat, losing some of their luminosity and producing a "shadow" image that is visible with the power off. Burn-in cannot be repaired (except on monochrome CRTs), and is especially a problem on plasma panels because they run hotter than CRTs. Early plasma televisions were plagued by burn-in, making it impossible to use video games or anything else that displayed static images.

Plasma displays also exhibit another image retention issue which is sometimes confused with screen burn-in damage. In this mode, when a group of pixels are run at high brightness (when displaying white, for example) for an extended period of time, a charge build-up in the pixel structure occurs and a ghost image can be seen. However, unlike burn-in, this charge build-up is transient and self corrects after the image condition that caused the effect has been removed and a long enough period of time has passed (with the display either off or on).

Plasma manufacturers have tried various ways of reducing burn-in such as using gray pillarboxes, pixel orbiters and image washing routines, but none to date have eliminated the problem and all plasma manufacturers continue to exclude burn-in from their warranties.

Environmental impact

Nitrogen trifluoride, cited as a very potent greenhouse gas, is used during production of plasma screens, which are therefore alleged to contribute to climate change. Plasma screens have also been lagging behind CRT and LCD screens in terms of energy consumption. To reduce the energy consumption, new technologies are also being found. Although it can be expected that plasma screens will continue to become more energy efficient in the future, a growing problem is that people tend to keep their old TVs running and an increasing trend to escalating screen sizes.

History

Plasma displays were first used in PLATO computer terminals. This PLATO V model illustrates the display's monochromatic orange glow as seen in 1988.

In 1936 Kálmán Tihanyi described the principle of "plasma television" and conceived the first flat-panel television system.

The monochrome plasma video display was co-invented in 1964 at the University of Illinois at Urbana-Champaign by Donald Bitzer, H. Gene Slottow, and graduate student Robert Willson for the PLATO Computer System. The original neon orange monochrome Digivue display panels built by glass producer Owens-Illinois were very popular in the early 1970s because they were rugged and needed neither memory nor

circuitry to refresh the images. A long period of sales decline occurred in the late 1970s because semiconductor memory made CRT displays cheaper than the US$2500 512 x 512 PLATO plasma displays. Nonetheless, the plasma displays' relatively large screen size and 1 inch thickness made them suitable for high-profile placement in lobbies and stock exchanges.

Electrical engineering student Larry F. Weber became interested in plasma displays while studying at the University of Illinois at Urbana-Champaign in the 1960s, and pursued postgraduate work in the field under Bitzer and Slottow. His research eventually earned him 15 patents relating to plasma displays. One of his early contributions was development of the power-saving "energy recovery sustain circuit", now included in every color plasma display.

Burroughs Corporation, a maker of adding machines and computers, developed the Panaplex display in the early 1970s. The Panaplex display, generically referred to as a gas-discharge or gas-plasma display, uses the same technology as later plasma video displays, but began life as seven-segment display for use in adding machines. They became popular for their bright orange luminous look and found nearly ubiquitous use in cash registers, calculators, pinball machines, aircraft avionics such as radios, navigational instruments, and stormscopes; test equipment such as frequency counters and multimeters; and generally anything that previously used nixie tube or numitron displays with a high digit-count throughout the late 1970s and into the 1990s. These displays remained popular until LEDs gained popularity because of their low-current draw and module-flexibility, but are still found in some applications where their high-brightness is desired, such as pinball machines and avionics. Pinball displays started with six- and seven-digit seven-segment displays and later evolved into 16-segment alphanumeric displays, and later into 128x32 dot-matrix displays in 1990, which are still used today.

1983

In 1983, IBM introduced a 19-inch (48 cm) orange-on-black monochrome display (model 3290 'information panel') which was able to show up to four simultaneous IBM 3270 terminal sessions. Due to heavy competition from monochrome LCD's, in 1987 IBM planned to shut down its factory in upstate New York, the largest plasma plant in the world, in favor of manufacturing mainframe computers. Consequently, Larry Weber co-founded a startup company Plasmaco with Stephen Globus, as well as James Kehoe, who was the IBM plant manager, and bought the plant from IBM. Weber stayed in Urbana as CTO until 1990, then moved to upstate New York to work at Plasmaco.

1992

In 1992, Fujitsu introduced the world's first 21-inch (53 cm) full-color display. It was a hybrid, the plasma display created at the University of Illinois at Urbana-Champaign and NHK Science & Technology Research Laboratories.

1994

In 1994, Weber demonstrated color plasma technology at an industry convention in San Jose. Panasonic Corporation began a joint development project with Plasmaco, which led in 1996 to the purchase of Plasmaaco, its color AC technology, and its American factory.

1997

In 1997, Fujitsu introduced the first 42-inch (107 cm) plasma display; it had 852x480 resolution and was progressively scanned. Also in 1997, Philips introduced a 42-inch (107 cm) display, with 852x480 resolution. It was the only plasma to be displayed to the retail public in 4 Sears locations in the US. The price was US$14,999 and included in-home installation. Later in 1997, Pioneer started selling their first plasma television to the public, and others followed.

2006 - Present

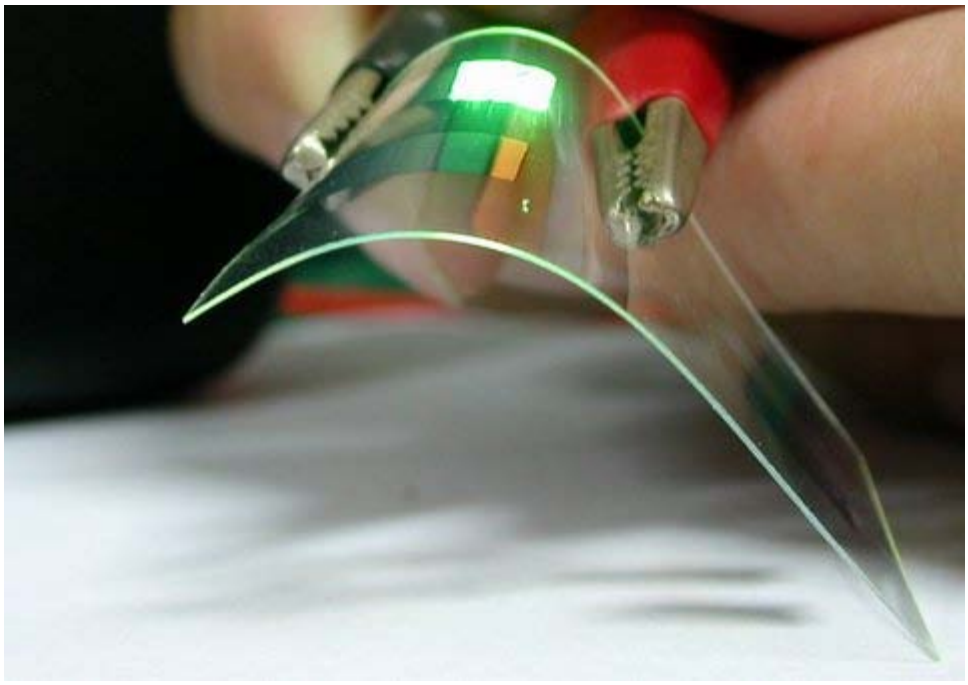
In late 2006, analysts noted that LCDs overtook plasmas, particularly in the 40-inch (1.0 m) and above segment where plasma had previously gained market share. Another industry trend is the consolidation of manufacturers of plasma displays, with around fifty brands available but only five manufacturers. In the first quarter of 2008 a comparison of worldwide TV sales breaks down to 22.1 million for direct-view CRT, 21.1 million for LCD, 2.8 million for Plasma, and 0.1 million for rear-projection.

Until the early 2000s, plasma displays were the most popular choice for HDTV flat panel display as they had many benefits over LCDs. Beyond plasma's deeper blacks, increased contrast, faster response time, greater color spectrum, and wider viewing angle; they were also much bigger than LCDs, and it was believed that LCD technology was suited only to smaller sized televisions. However, improvements in VLSI fabrication technology have since narrowed the technological gap. The increased size, lower weight, falling prices, and often lower electrical power consumption of LCDs now make them competitive with plasma television sets.

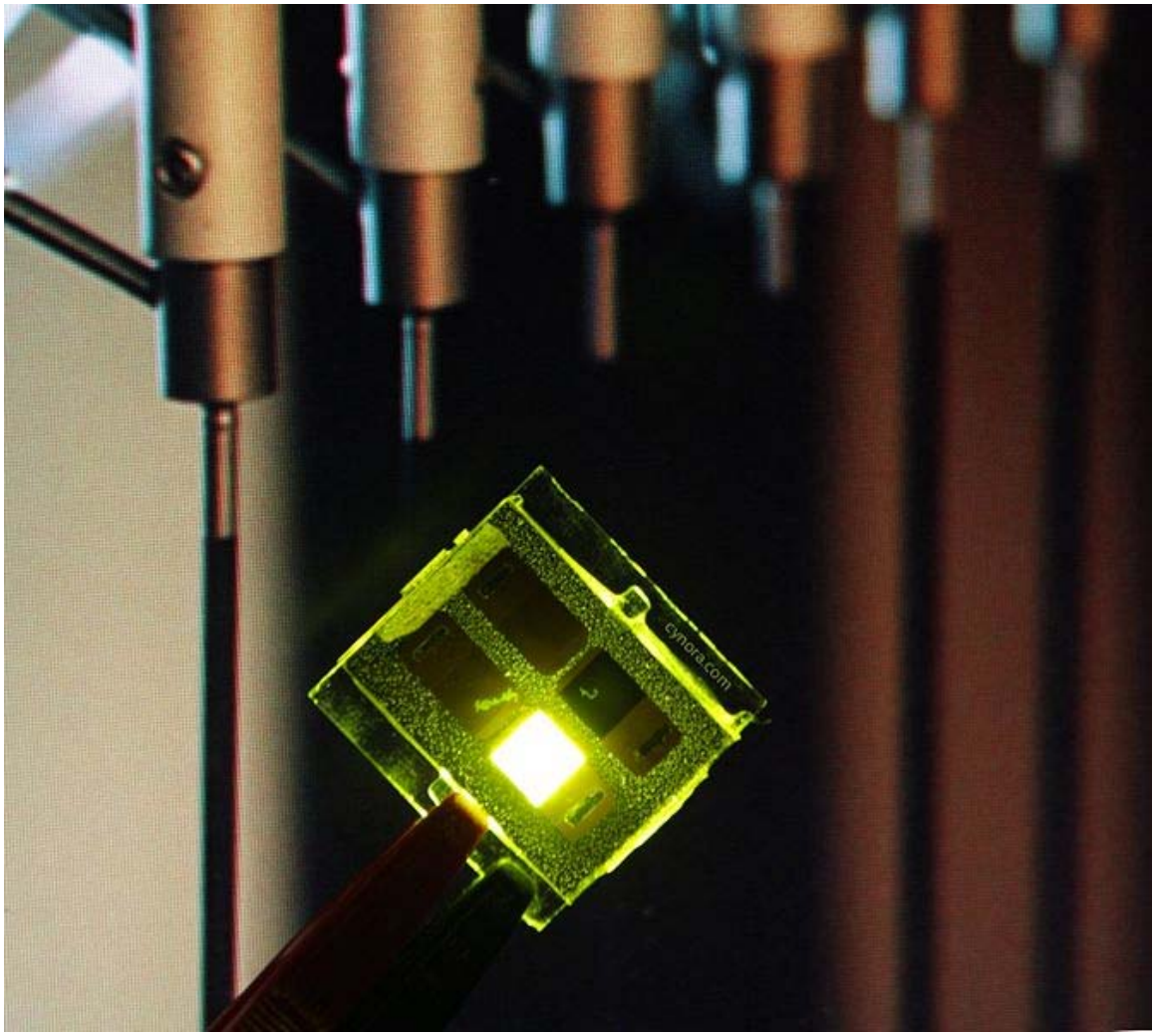
Screen sizes have increased since the introduction of plasma displays. The largest plasma video display in the world at the 2008 Consumer Electronics Show in Las Vegas, Nevada, was a 150-inch (381 cm) unit manufactured by Matsushita Electric Industrial (Panasonic) standing 6 ft (180 cm) tall by 11 ft (330 cm) wide. At the 2010 Consumer Electronics Show in Las Vegas, Panasonic introduced their 152" 2160p 3D plasma. In 2010 Panasonic shipped 19.1 million plasma TV panels.

Chapter-11

Organic Light-emitting Diode



Demonstration of a flexible OLED device



A green emitting OLED device

An **organic light emitting diode** (**OLED**) is a light-emitting diode (LED) in which the emissive electroluminescent layer is a film of organic compounds which emit light in response to an electric current. This layer of organic semiconductor material is situated between two electrodes. Generally, at least one of these electrodes is transparent.

OLEDs are used in television screens, computer monitors, small, portable system screens such as mobile phones and PDAs, watches, advertising, information, and indication. OLEDs are also used in light sources for space illumination and in large-area light-emitting elements. Due to their early stage of development, they typically emit less light per unit area than inorganic solid-state based LED point-light sources.

An OLED display functions without a backlight. Thus, it can display deep black levels and can be thinner and lighter than liquid crystal displays. In low ambient light conditions such as dark rooms, an OLED screen can achieve a higher contrast ratio than an LCD

using either cold cathode fluorescent lamps or the more recently developed LED backlight.

There are two main families of OLEDs: those based upon small molecules and those employing polymers. Adding mobile ions to an OLED creates a Light-emitting Electrochemical Cell or LEC, which has a slightly different mode of operation.

OLED displays can use either passive-matrix (PMOLED) or active-matrix addressing schemes. Active-matrix OLEDs (AMOLED) require a thin-film transistor backplane to switch each individual pixel on or off, and can make higher resolution and larger size displays possible.

History

The first observations of electroluminescence in organic materials were in the early 1950s by A. Bernanose and co-workers at the Nancy-Université, France. They applied high-voltage alternating current (AC) fields in air to materials such as acridine orange, either deposited on or dissolved in cellulose or cellophane thin films. The proposed mechanism was either direct excitation of the dye molecules or excitation of electrons.

In 1960, Martin Pope and co-workers at New York University developed ohmic dark-injecting electrode contacts to organic crystals. They further described the necessary energetic requirements (work functions) for hole and electron injecting electrode contacts. These contacts are the basis of charge injection in all modern OLED devices. Pope's group also first observed direct current (DC) electroluminescence under vacuum on a pure single crystal of anthracene and on anthracene crystals doped with tetracene in 1963 using a small area silver electrode at 400V. The proposed mechanism was field-accelerated electron excitation of molecular fluorescence.

Pope's group reported in 1965 that in the absence of an external electric field, the electroluminescence in anthracene crystals is caused by the recombination of a thermalized electron and hole, and that the conducting level of anthracene is higher in energy than the exciton energy level. Also in 1965, W. Helfrich and W. G. Schneider of the National Research Council in Canada produced double injection recombination electroluminescence for the first time in an anthracene single crystal using hole and electron injecting electrodes, the forerunner of modern double injection devices. In the same year, Dow Chemical researchers patented a method of preparing electroluminescent cells using high voltage (500–1500 V) AC-driven (100–3000 Hz) electrically-insulated one millimetre thin layers of a melted phosphor consisting of ground anthracene powder, tetracene, and graphite powder. Their proposed mechanism involved electronic excitation at the contacts between the graphite particles and the anthracene molecules.

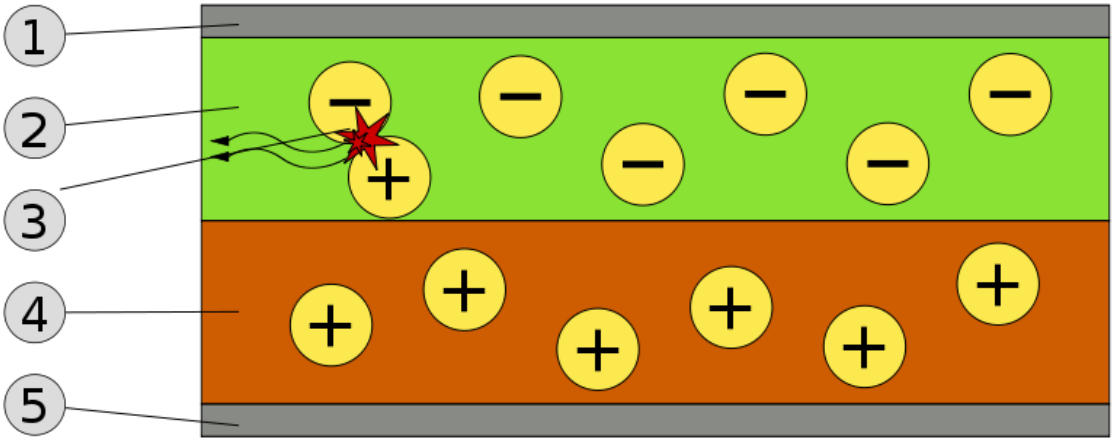
Device performance was limited by the poor electrical conductivity of contemporary organic materials. This was overcome by the discovery and development of highly conductive polymers.

Electroluminescence from polymer films was first observed by Roger Partridge at the National Physical Laboratory in the United Kingdom. The device consisted of a film of poly(n-vinylcarbazole) up to 2.2 micrometres thick located between two charge injecting electrodes. The results of the project were patented in 1975 and published in 1983.

The first diode device was reported at Eastman Kodak by Ching W. Tang and Steven Van Slyke in 1987. This device used a novel two-layer structure with separate hole transporting and electron transporting layers such that recombination and light emission occurred in the middle of the organic layer. This resulted in a reduction in operating voltage and improvements in efficiency and led to the current era of OLED research and device production.

Research into polymer electroluminescence culminated in 1990 with J. H. Burroughes *et al.* at the Cavendish Laboratory in Cambridge reporting a high efficiency green light-emitting polymer based device using 100 nm thick films of poly(p-phenylene vinylene).

Working principle



Schematic of a bilayer OLED: 1. Cathode (−), 2. Emissive Layer, 3. Emission of radiation, 4. Conductive Layer, 5. Anode (+)

A typical OLED is composed of a layer of organic materials situated between two electrodes, the anode and cathode, all deposited on a substrate. The organic molecules are electrically conductive as a result of delocalization of pi electrons caused by conjugation over all or part of the molecule. These materials have conductivity levels ranging from insulators to conductors, and therefore are considered organic semiconductors. The highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO) of organic semiconductors are analogous to the valence and conduction bands of inorganic semiconductors.

Originally, the most basic polymer OLEDs consisted of a single organic layer. One example was the first light-emitting device synthesised by J. H. Burroughes *et al.*, which involved a single layer of poly(p-phenylene vinylene). However multilayer OLEDs can

be fabricated with two or more layers in order to improve device efficiency. As well as conductive properties, different materials may be chosen to aid charge injection at electrodes by providing a more gradual electronic profile, or block a charge from reaching the opposite electrode and being wasted. Many modern OLEDs incorporate a simple bilayer structure, consisting of a conductive layer and an emissive layer.

During operation, a voltage is applied across the OLED such that the anode is positive with respect to the cathode. A current of electrons flows through the device from cathode to anode, as electrons are injected into the LUMO of the organic layer at the cathode and withdrawn from the HOMO at the anode. This latter process may also be described as the injection of electron holes into the HOMO. Electrostatic forces bring the electrons and the holes towards each other and they recombine forming an exciton, a bound state of the electron and hole. This happens closer to the emissive layer, because in organic semiconductors holes are generally more mobile than electrons. The decay of this excited state results in a relaxation of the energy levels of the electron, accompanied by emission of radiation whose frequency is in the visible region. The frequency of this radiation depends on the band gap of the material, in this case the difference in energy between the HOMO and LUMO.

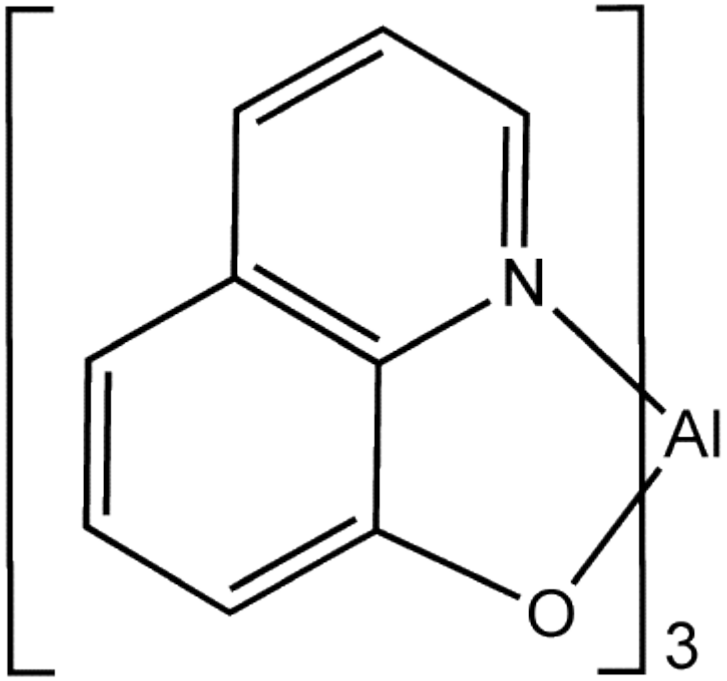
As electrons and holes are fermions with half integer spin, an exciton may either be in a singlet state or a triplet state depending on how the spins of the electron and hole have been combined. Statistically three triplet excitons will be formed for each singlet exciton. Decay from triplet states (phosphorescence) is spin forbidden, increasing the timescale of the transition and limiting the internal efficiency of fluorescent devices. Phosphorescent organic light-emitting diodes make use of spin–orbit interactions to facilitate intersystem crossing between singlet and triplet states, thus obtaining emission from both singlet and triplet states and improving the internal efficiency.

Indium tin oxide (ITO) is commonly used as the anode material. It is transparent to visible light and has a high work function which promotes injection of holes into the HOMO level of the organic layer. A typical conductive layer may consist of PEDOT:PSS as the HOMO level of this material generally lies between the workfunction of ITO and the HOMO of other commonly used polymers, reducing the energy barriers for hole injection. Metals such as barium and calcium are often used for the cathode as they have low work functions which promote injection of electrons into the LUMO of the organic layer. Such metals are reactive, so require a capping layer of aluminium to avoid degradation.

Single carrier devices are typically used to study the kinetics and charge transport mechanisms of an organic material and can be useful when trying to study energy transfer processes. As current through the device is composed of only one type of charge carrier, either electrons or holes, recombination does not occur and no light is emitted. For example, electron only devices can be obtained by replacing ITO with a lower work function metal which increases the energy barrier of hole injection. Similarly, hole only devices can be made by using a cathode comprised solely of aluminium, resulting in an energy barrier too large for efficient electron injection.

Material technologies

Small molecules



Alq_3, commonly used in small molecule OLEDs.

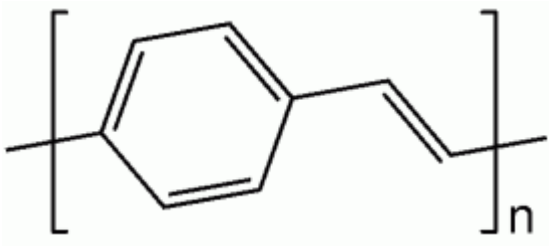
Efficient OLEDs using small molecules were first developed by Dr. Ching W. Tang *et al.* at Eastman Kodak. The term OLED traditionally refers specifically to this type of device, though the term SM-OLED is also in use.

Molecules commonly used in OLEDs include organometallic chelates (for example Alq_3, used in the organic light-emitting device reported by Tang *et al.*), fluorescent and phosphorescent dyes and conjugated dendrimers. A number of materials are used for their charge transport properties, for example triphenylamine and derivatives are commonly used as materials for hole transport layers. Fluorescent dyes can be chosen to obtain light emission at different wavelengths, and compounds such as perylene, rubrene and quinacridone derivatives are often used. Alq_3 has been used as a green emitter, electron transport material and as a host for yellow and red emitting dyes.

The production of small molecule devices and displays usually involves thermal evaporation in a vacuum. This makes the production process more expensive and of limited use for large-area devices than other processing techniques. However, contrary to polymer-based devices, the vacuum deposition process enables the formation of well controlled, homogeneous films, and the construction of very complex multi-layer structures. This high flexibility in layer design, enabling distinct charge transport and charge blocking layers to be formed, is the main reason for the high efficiencies of the small molecule OLEDs.

Coherent emission from a laser dye-doped tandem SM-OLED device, excited in the pulsed regime, has been demonstrated. The emission is nearly diffraction limited with a spectral width similar to that of broadband dye lasers.

Polymer light-emitting diodes



poly(*p*-phenylene vinylene), used in the first PLED.

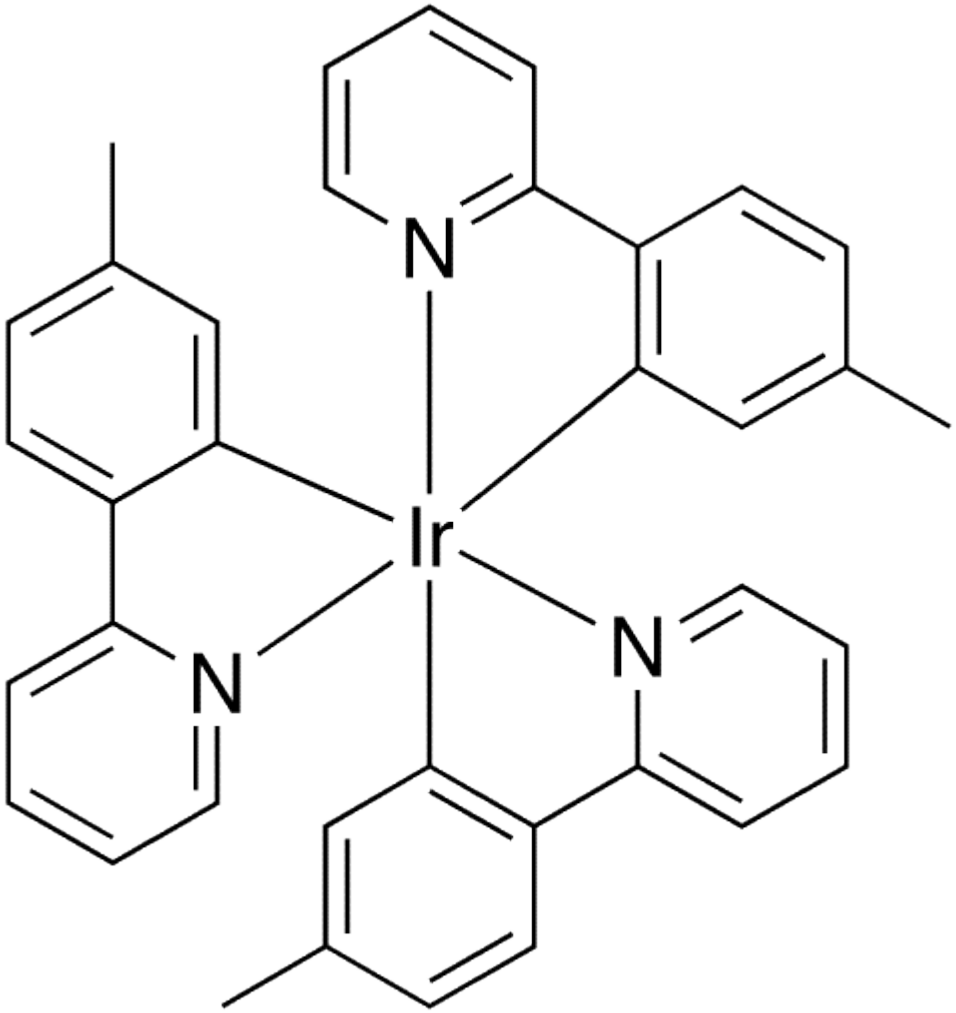
Polymer light-emitting diodes (PLED), also light-emitting polymers (LEP), involve an electroluminescent conductive polymer that emits light when connected to an external voltage. They are used as a thin film for full-spectrum colour displays. Polymer OLEDs are quite efficient and require a relatively small amount of power for the amount of light produced.

Vacuum deposition is not a suitable method for forming thin films of polymers. However, polymers can be processed in solution, and spin coating is a common method of depositing thin polymer films. This method is more suited to forming large-area films than thermal evaporation. No vacuum is required, and the emissive materials can also be applied on the substrate by a technique derived from commercial inkjet printing. However, as the application of subsequent layers tends to dissolve those already present, formation of multilayer structures is difficult with these methods. The metal cathode may still need to be deposited by thermal evaporation in vacuum.

Typical polymers used in PLED displays include derivatives of poly(*p*-phenylene vinylene) and polyfluorene. Substitution of side chains onto the polymer backbone may determine the colour of emitted light or the stability and solubility of the polymer for performance and ease of processing.

While unsubstituted poly(p-phenylene vinylene) (PPV) is typically insoluble, a number of PPVs and related poly(naphthalene vinylene)s (PNVs) that are soluble in organic solvents or water have been prepared via ring opening metathesis polymerization.

Phosphorescent materials



$Ir(mppy)_3$, a phosphorescent dopant which emits green light.

Phosphorescent organic light emitting diodes use the principle of electrophosphorescence to convert electrical energy in an OLED into light in a highly efficient manner, with the internal quantum efficiencies of such devices approaching 100%.

Typically, a polymer such as poly(n-vinylcarbazole) is used as a host material to which an organometallic complex is added as a dopant. Iridium complexes such as $Ir(mppy)_3$ are currently the focus of research, although complexes based on other heavy metals such as platinum have also been used.

The heavy metal atom at the centre of these complexes exhibits strong spin-orbit coupling, facilitating intersystem crossing between singlet and triplet states. By using these phosphorescent materials, both singlet and triplet excitons will be able to decay radiatively, hence improving the internal quantum efficiency of the device compared to a standard PLED where only the singlet states will contribute to emission of light.

Applications of OLEDs in solid state lighting require the achievement of high brightness with good CIE coordinates (for white emission). The use of macromolecular species like polyhedral oligomeric silsesquioxanes (POSS) in conjunction with the use of phosphorescent species such as Ir for printed OLEDs have exhibited brightnesses as high as 10,000 cd/m^2.

Device Architectures

Structure

- **Bottom or top emission:** Bottom emission devices use a transparent or semi-transparent bottom electrode to get the light through a transparent substrate. Top emission devices use a transparent or semi-transparent top electrode emitting light directly. Top-emitting OLEDs are better suited for active-matrix applications as they can be more easily integrated with a non-transparent transistor backplane.

- **Transparent OLEDs** use transparent or semi-transparent contacts on both sides of the device to create displays that can be made to be both top and bottom emitting (transparent). TOLEDs can greatly improve contrast, making it much easier to view displays in bright sunlight. This technology can be used in Head-up displays, smart windows or augmented reality applications. Novaled's OLED panel presented in Finetech Japan 2010, boasts a transparency of 60–70%.

- **Stacked OLEDs** use a pixel architecture that stacks the red, green, and blue subpixels on top of one another instead of next to one another, leading to substantial increase in gamut and color depth, and greatly reducing pixel gap. Currently, other display technologies have the RGB (and RGBW) pixels mapped next to each other decreasing potential resolution.

- **Inverted OLED:** In contrast to a conventional OLED, in which the anode is placed on the substrate, an Inverted OLED uses a bottom cathode that can be connected to the drain end of an n-channel TFT especially for the low cost amorphous silicon TFT backplane useful in the manufacturing of AMOLED displays.

Patterning technologies

Patternable organic light-emitting devices use a light or heat activated electroactive layer. A latent material (PEDOT-TMA) is included in this layer that, upon activation, becomes highly efficient as a hole injection layer. Using this process, light-emitting devices with arbitrary patterns can be prepared.

Colour patterning can be accomplished by means of laser, such as radiation-induced sublimation transfer (RIST).

Organic vapour jet printing (OVJP) uses an inert carrier gas, such as argon or nitrogen, to transport evaporated organic molecules (as in Organic Vapor Phase Deposition). The gas is expelled through a micron sized nozzle or nozzle array close to the substrate as it is being translated. This allows printing arbitrary multilayer patterns without the use of solvents.

Conventional OLED displays are formed by vapor thermal evaporation (VTE) and are patterned by shadow-mask. A mechanical mask has openings allowing the vapor to pass only on the desired location.

Backplane technologies

For a high resolution display like a TV, a TFT backplane is necessary to drive the pixels correctly. Currently, Low Temperature Polycrystalline silicon LTPS-TFT is used for commercial AMOLED displays. LTPS-TFT has variation of the performance in a display, so various compensation circuits have been reported. Due to the size limitation of the excimer laser used for LTPS, the AMOLED size was limited. To cope with the hurdle related to the panel size, amorphous-silicon/microcrystalline-silicon backplanes have been reported with large display prototype demonstrations.

Advantages

Demonstration of a 4.1" prototype flexible display from Sony

The different manufacturing process of OLEDs lends itself to several advantages over flat-panel displays made with LCD technology.

- **Lower cost in the future:** OLEDs can be printed onto any suitable substrate by an inkjet printer or even by screen printing, theoretically making them cheaper to produce than LCD or plasma displays. However, fabrication of the OLED substrate is more costly than that of a TFT LCD, until mass production methods lower cost through scalability. Roll-roll vapour-deposition methods for organic devices do allow mass production of thousands of devices per minute for minimal cost, although this technique also induces problems in that multi-layer devices can be challenging to make.

- **Light weight & flexible plastic substrates:** OLED displays can be fabricated on flexible plastic substrates leading to the possibility of flexible organic light-emitting diodes being fabricated or other new applications such as roll-up displays embedded in fabrics or clothing. As the substrate used can be flexible such as PET., the displays may be produced inexpensively.

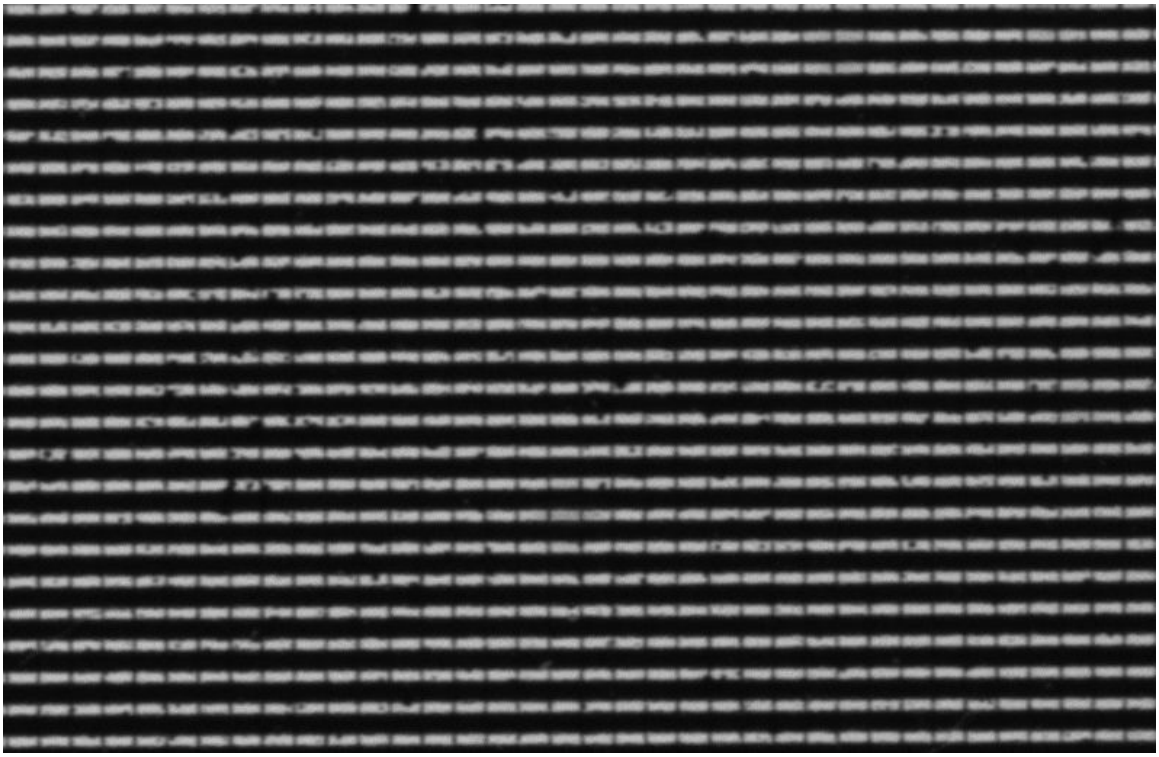
- **Wider viewing angles & improved brightness:** OLEDs can enable a greater artificial contrast ratio (both dynamic range and static, measured in purely dark conditions) and viewing angle compared to LCDs because OLED pixels directly emit light. OLED pixel colours appear correct and unshifted, even as the viewing angle approaches 90° from normal.

- **Better power efficiency:** LCDs filter the light emitted from a backlight, allowing a small fraction of light through so they cannot show true black, while an inactive OLED element does not produce light or consume power.

- **Response time:** OLEDs can also have a faster response time than standard LCD screens. Whereas LCD displays are capable of between 2 and 8 ms response time offering a frame rate of +/-200 Hz, an OLED can theoretically have less than 0.01 ms response time enabling 100,000 Hz refresh rates.

Disadvantages

LEP display showing partial failure



An old OLED display showing wear

- **Current costs:** OLED manufacture currently requires process steps that make it extremely expensive. Specifically, it requires the use of Low-Temperature Polysilicon backplanes; LTPS backplanes in turn require laser annealing from an amorphous silicon start, so this part of the manufacturing process for AMOLEDs starts with the process costs of standard LCD, and then adds an expensive, time-consuming process that cannot currently be used on large-area glass substrates.

- **Lifespan:** The biggest technical problem for OLEDs was the limited lifetime of the organic materials. In particular, blue OLEDs historically have had a lifetime of around 14,000 hours to half original brightness (five years at 8 hours a day) when used for flat-panel displays. This is lower than the typical lifetime of LCD, LED or PDP technology—each currently rated for about 25,000 – 40,000 hours to half brightness, depending on manufacturer and model. However, some manufacturers' displays aim to increase the lifespan of OLED displays, pushing their expected life past that of LCD displays by improving light outcoupling, thus achieving the same brightness at a lower drive current. In 2007, experimental OLEDs were created which can sustain 400 cd/m^2 of luminance for over 198,000 hours for green OLEDs and 62,000 hours for blue OLEDs.

- **Color balance issues:** Additionally, as the OLED material used to produce blue light degrades significantly more rapidly than the materials that produce other colors, blue light output will decrease relative to the other colors of light. This differential color output change will change the color balance of the display and is

much more noticeable than a decrease in overall luminance. This can be partially avoided by adjusting colour balance but this may require advanced control circuits and interaction with the user, which is unacceptable for some users. In order to delay the problem, manufacturers bias the colour balance towards blue so that the display initially has an artificially blue tint, leading to complaints of artificial-looking, over-saturated colors. More commonly, though, manufacturers optimize the size of the R, G and B subpixels to reduce the current density through the subpixel in order to equalize lifetime at full luminance. For example, a blue subpixel may be 100% larger than the green subpixel. The red subpixel may be 10% smaller than the green.

- **Efficiency of blue OLEDs:** Improvements to the efficiency and lifetime of blue OLEDs is vital to the success of OLEDs as replacements for LCD technology. Considerable research has been invested in developing blue OLEDs with high external quantum efficiency as well as a deeper blue color. External quantum efficiency values of 20% and 19% have been reported for red (625 nm) and green (530 nm) diodes, respectively. However, blue diodes (430 nm) have only been able to achieve maximum external quantum efficiencies in the range between 4% to 6%.

- **Water damage:** Water can damage the organic materials of the displays. Therefore, improved sealing processes are important for practical manufacturing. Water damage may especially limit the longevity of more flexible displays.

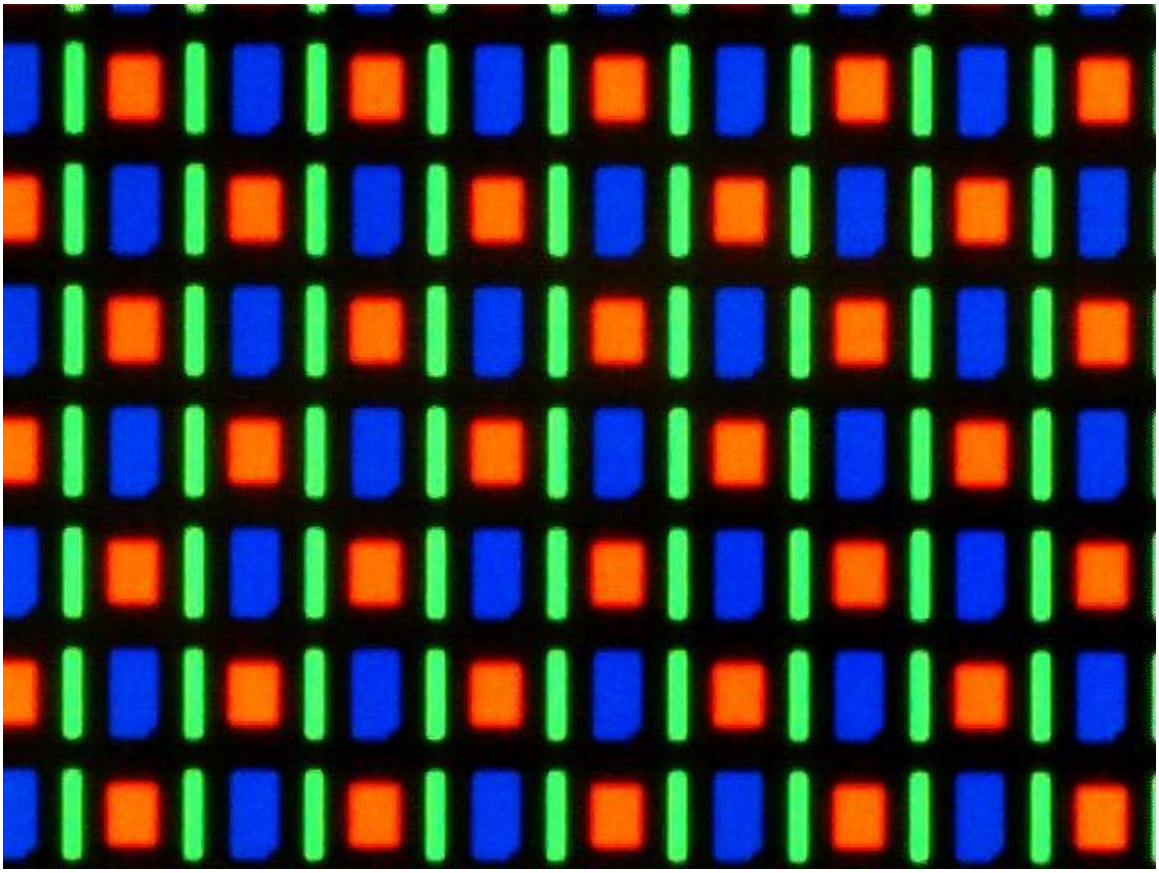
- **Outdoor performance:** As an emissive display technology, OLEDs rely completely upon converting electricity to light, unlike most LCDs which are to some extent reflective; e-ink leads the way in efficiency with ~ 33% ambient light reflectivity, enabling the display to be used without any internal light source. The metallic cathode in an OLED acts as a mirror, with reflectance approaching 80%, leading to poor readability in bright ambient light such as outdoors. However, with the proper application of a circular polarizer and anti-reflective coatings, the diffuse reflectance can be reduced to less than 0.1%. With 10,000 fc incident illumination (typical test condition for simulating outdoor illumination), that yields an approximate photopic contrast of 5:1.

- **Power consumption:** While an OLED will consume around 40% of the power of an LCD displaying an image which is primarily black, for the majority of images it will consume 60–80% of the power of an LCD – however it can use over three times as much power to display an image with a white background such as a document or website. This can lead to reduced real-world battery life in mobile devices.

- **Screen burn-in:** Unlike displays with a common light source, the brightness of each OLED pixel fades depending on the content displayed. The varied lifespan of the organic dyes can cause a discrepancy between red, green, and blue intensity. This leads to image persistence, also known as burn-in.

- **UV sensitivity:** OLED displays can be damaged by prolonged exposure to UV light. The most pronounced example of this can be seen with a near UV laser (such as a Bluray pointer) and can damage the display almost instantly with more than 20 mW leading to dim or dead spots where the beam is focused. This is usually avoided by installing a UV blocking filter over the panel and this can easily be seen as a clear plastic layer on the glass. Removal of this filter can lead to severe damage and an unusable display after only a few months of room light exposure.

Manufacturers and commercial uses



Magnified image of the AMOLED screen on the Google Nexus One smartphone using the RGBG system of the PenTile Matrix Family.

A 3.8 cm (1.5 in) OLED display from a Creative ZEN V media player

OLED technology is used in commercial applications such as displays for mobile phones and portable digital media players, car radios and digital cameras among others. Such portable applications favor the high light output of OLEDs for readability in sunlight and their low power drain. Portable displays are also used intermittently, so the lower lifespan of organic displays is less of an issue. Prototypes have been made of flexible and rollable displays which use OLEDs' unique characteristics. Applications in flexible signs and lighting are also being developed. Philips Lighting have made OLED lighting samples under the brand name 'Lumiblade' available online.

OLEDs have been used in most Motorola and Samsung colour cell phones, as well as some HTC, LG and Sony Ericsson models. Nokia has also recently introduced some OLED products including the N85 and the N86 8MP, both of which feature an AMOLED display. OLED technology can also be found in digital media players such as the Creative ZEN V, the iriver clix, the Zune HD and the Sony Walkman X Series.

The Google and HTC Nexus One smartphone includes an AMOLED screen, as does HTC's own Desire and Legend phones. However due to supply shortages of the Samsung-produced displays, certain HTC models will use Sony's SLCD displays in the future, while the Google and Samsung Nexus S smartphone will use "Super Clear LCD" instead in some countries.

Other manufacturers of OLED panels include Anwell Technologies Limited, Chi Mei Corporation, LG, and others.

DuPont stated in a press release in May 2010 that they can produce a 50-inch OLED TV in two minutes with a new printing technology. If this can be scaled up in terms of manufacturing, then the total cost of OLED TVs would be greatly reduced. Dupont also states that OLED TVs made with this less expensive technology can last up to 15 years if left on for a normal eight hour day.

The use of OLEDs may be subject to patents held by Eastman Kodak, DuPont, General Electric, Royal Philips Electronics, numerous universities and others. There are by now literally thousands of patents associated with OLEDs, both from larger corporations and smaller technology companies .

Samsung applications

By 2004 Samsung, South Korea's largest conglomerate, was the world's largest OLED manufacturer, producing 40% of the OLED displays made in the world, and as of 2010 has a 98% share of the global AMOLED market. The company is leading the world OLED industry, generating $100.2 million out of the total $475 million revenues in the global OLED market in 2006. As of 2006, it held more than 600 American patents and more than 2800 international patents, making it the largest owner of AMOLED technology patents.

Samsung SDI announced in 2005 the world's largest OLED TV at the time, at 21 inches (53 cm). This OLED featured the highest resolution at the time, of 6.22 million pixels. In addition, the company adopted active matrix based technology for its low power consumption and high-resolution qualities. This was exceeded in January 2008, when Samsung showcased the world's largest and thinnest OLED TV at the time, at 31 inches and 4.3 mm.

In May 2008, Samsung unveiled an ultra-thin 12.1 inch laptop OLED display concept, with a 1,280×768 resolution with infinite contrast ratio. According to Woo Jong Lee, Vice President of the Mobile Display Marketing Team at Samsung SDI, the company expected OLED displays to be used in notebook PCs as soon as 2010.

In October 2008, Samsung showcased the world's thinnest OLED display, also the first to be 'flappable' and bendable. It measures just 0.05 mm (thinner than paper), yet a Samsung staff member said that it is "technically possible to make the panel thinner". To achieve this thickness, Samsung etched an OLED panel that uses a normal glass substrate. The drive circuit was formed by low-temperature polysilicon TFTs. Also, low-molecular organic EL materials were employed. The pixel count of the display is 480 × 272. The contrast ratio is 100,000:1, and the luminance is 200 cd/m². The colour reproduction range is 100% of the NTSC standard.

In the same month, Samsung unveiled what was then the world's largest OLED Television at 40-inch with a Full HD resolution of 1920×1080 pixel. In the FPD International, Samsung stated that its 40-inch OLED Panel is the largest size currently possible. The panel has a contrast ratio of 1,000,000:1, a colour gamut of 107% NTSC, and a luminance of 200 cd/m² (peak luminance of 600 cd/m²).

At the Consumer Electronics Show (CES) in January 2010, Samsung demonstrated a laptop computer with a large, transparent OLED display featuring up to 40% transparency and an animated OLED display in a photo ID card.

Samsung's latest AMOLED smartphones use their Super AMOLED trademark, with the Samsung Wave S8500 and Samsung i9000 Galaxy S being launched in June 2010. In January 2011 Samsung announced their Super AMOLED Plus displays - which offer several advances over the older Super AMOLED displays - real stripe matrix (50% more sub pixels), thinner form factor, brighter image and a 18% reduction in energy consumption.

Sony applications

Sony XEL-1, the world's first OLED TV. (front)

Sony XEL-1 (side)

The Sony CLIÉ PEG-VZ90 was released in 2004, being the first PDA to feature an OLED screen. Other Sony products to feature OLED screens include the MZ-RH1 portable minidisc recorder, released in 2006 and the Walkman X Series.

At the Las Vegas CES 2007, Sony showcased 11-inch (28 cm, resolution 960×540) and 27-inch (68.5 cm, full HD resolution at 1920×1080) OLED TV models. Both claimed 1,000,000:1 contrast ratios and total thicknesses (including bezels) of 5 mm. In April 2007, Sony announced it would manufacture 1000 11-inch OLED TVs per month for market testing purposes. On October 1, 2007, Sony announced that the 11-inch model, now called the XEL-1, would be released commercially; the XEL-1 was first released in Japan in December 2007.

In May 2007, Sony publicly unveiled a video of a 2.5-inch flexible OLED screen which is only 0.3 millimeters thick. At the Display 2008 exhibition, Sony demonstrated a 0.2 mm thick 3.5 inch display with a resolution of 320×200 pixels and a 0.3 mm thick 11 inch display with 960×540 pixels resolution, one-tenth the thickness of the XEL-1.

In July 2008, a Japanese government body said it would fund a joint project of leading firms, which is to develop a key technology to produce large, energy-saving organic displays. The project involves one laboratory and 10 companies including Sony Corp.

NEDO said the project was aimed at developing a core technology to mass-produce 40 inch or larger OLED displays in the late 2010s.

In October 2008, Sony published results of research it carried out with the Max Planck Institute over the possibility of mass-market bending displays, which could replace rigid LCDs and plasma screens. Eventually, bendable, transparent OLED screens could be stacked to produce 3D images with much greater contrast ratios and viewing angles than existing products.

Sony exhibited a 24.5" prototype OLED 3D television during the Consumer Electronics Show in January 2010.

In January 2011, Sony announced the Next Generation Portable handheld game console (the successor to the PSP) will feature a 5-inch OLED screen.

On February 17, 2011, Sony announced its 25" OLED Professional Reference Monitor aimed at the Cinema and high end Drama Post Production market.

LG applications

As of 2010, LG produces one model of OLED television, the 15 inch 15EL9500 and has announced a 31" OLED 3D television for March 2011.