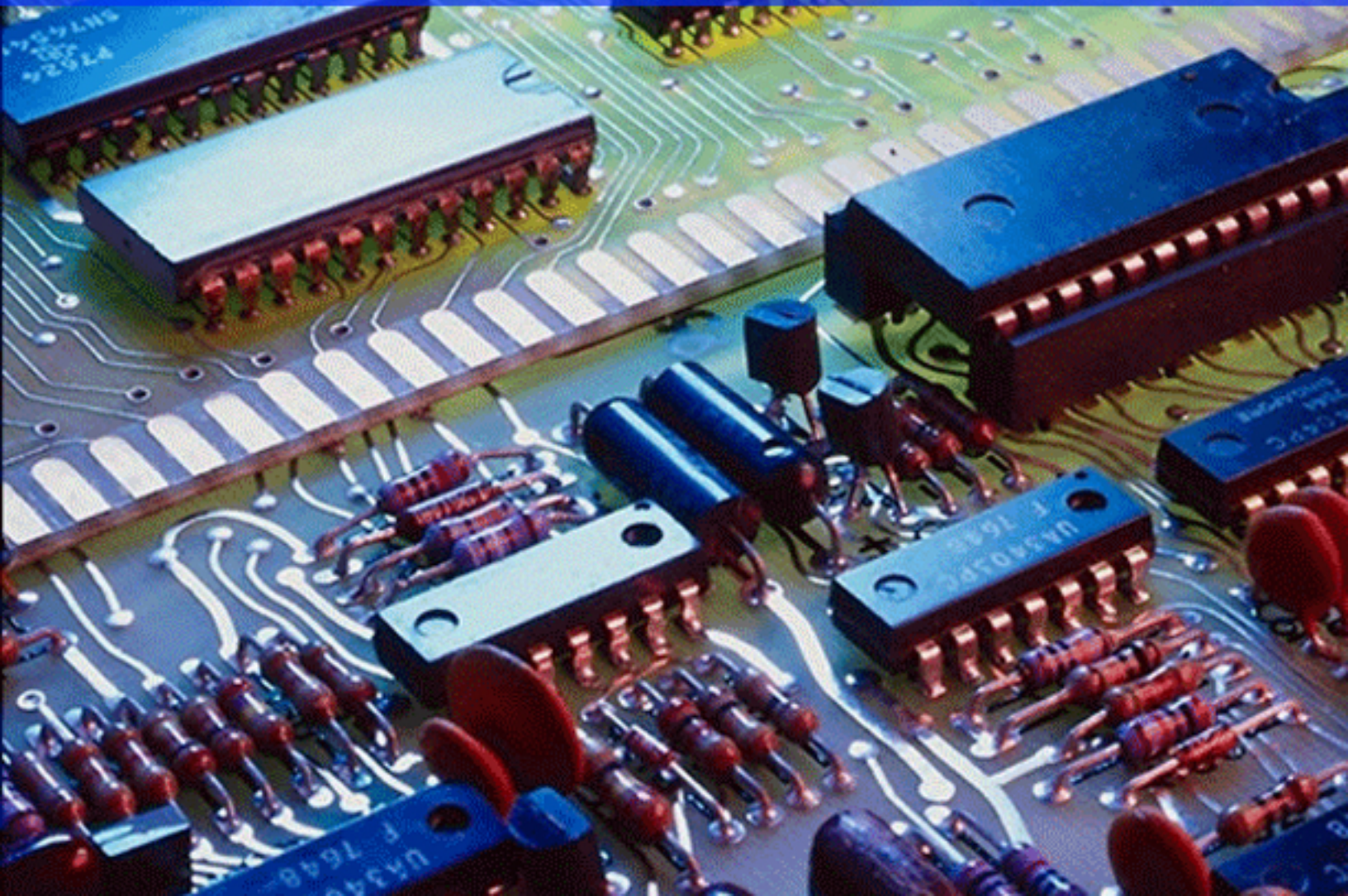
Microelectronics Engineering

Cletus Schilling



First Edition, 2012

ISBN 978-81-323-3492-7

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

Table of Contents

Introduction

Microelectronics is a subfield of electronics. Microelectronics, as the name suggests, is related to the study and manufacture, or microfabrication, of electronic components which are very small (usually micrometre-scale or smaller, but not always). These devices are made from semiconductors. Many components of normal electronic design are available in microelectronic equivalent: transistors, capacitors, inductors, resistors, diodes and of course insulators and conductors can all be found in microelectronic devices. Unique wiring techniques such as wire bonding are also often used in microelectronics because of the unusually small size of the components, leads and pads. This technique requires specialized equipment.

Digital integrated circuits (ICs) consist mostly of transistors. Analog circuits commonly contain resistors and capacitors as well. Inductors are used in some high frequency analog circuits, but tend to occupy large chip area if used at low frequencies; gyrators can replace them in many applications.

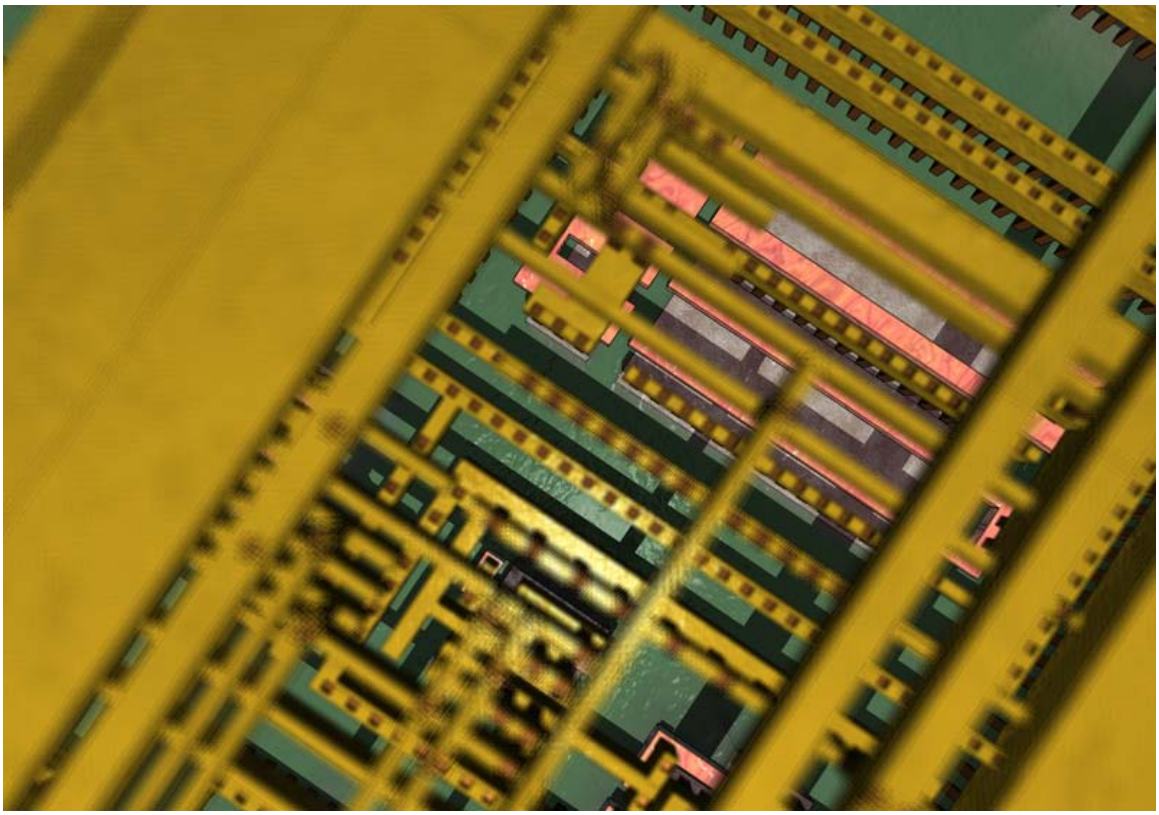
As techniques improve, the scale of microelectronic components continues to decrease. At smaller scales, the relative impact of intrinsic circuit properties such as interconnections may become more significant. These are called **parasitic effects**, and the goal of the microelectronics design engineer is to find ways to compensate for or to minimize these effects, while always delivering smaller, faster, and cheaper devices.

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.

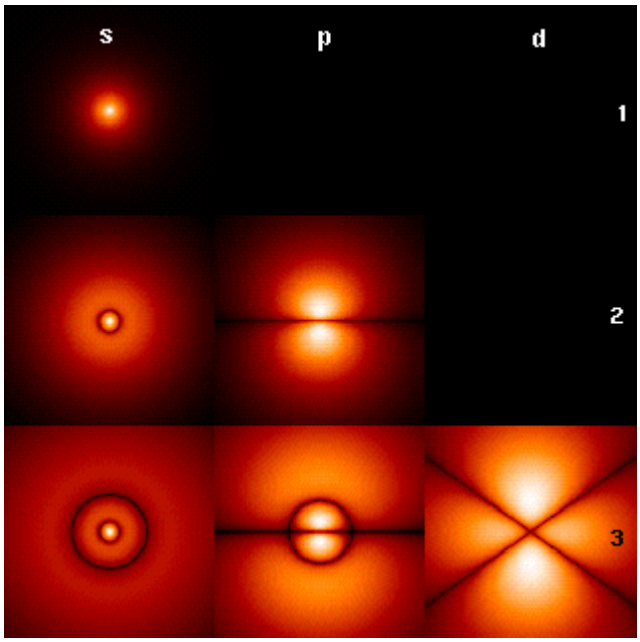
Silicon is used to create most semiconductors commercially. Dozens of other materials are used, including germanium, gallium arsenide, and silicon carbide. A pure semiconductor is often called an “intrinsic” semiconductor. The electronic properties and the conductivity of a semiconductor can be changed 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

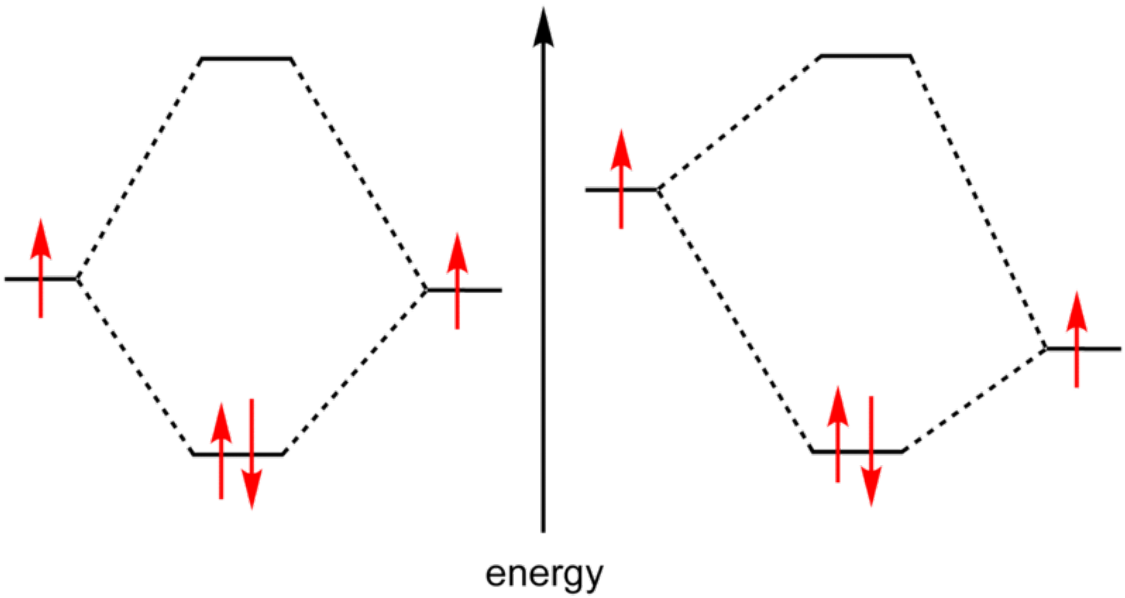
There are three popular ways to classify the electronic structure of a crystal.

- Band structure

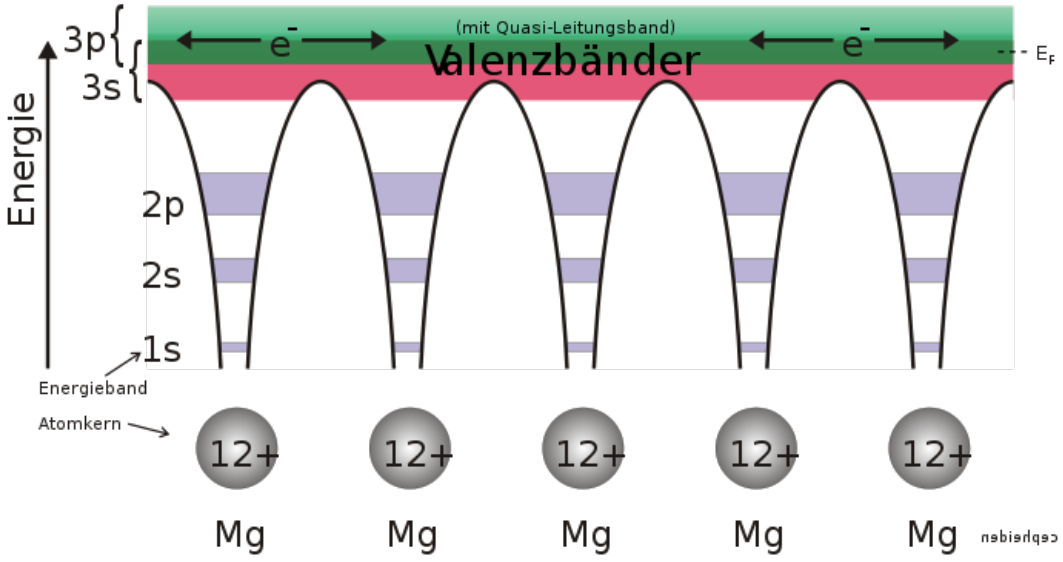
atoms – crystal – vacuum



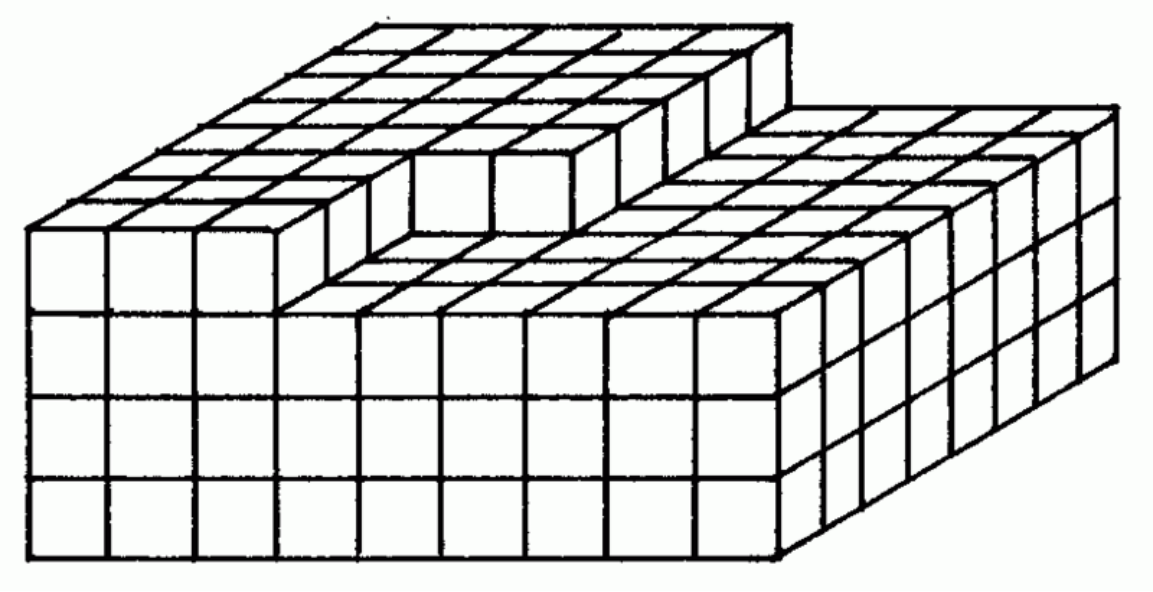
In a single H-atom an electron resides in well known orbitals. Note that the orbitals are called s,p,d in order of increasing circular current.



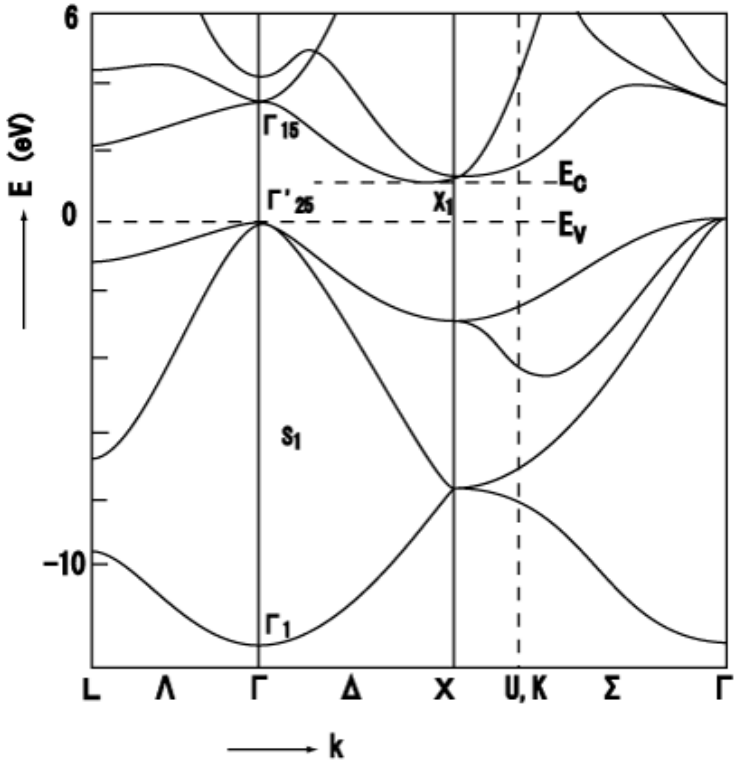
Putting two atoms together leads to delocalized orbitals across two atoms, yielding a covalent bond. Due to the Pauli exclusion principle, every state can contain only one electron.



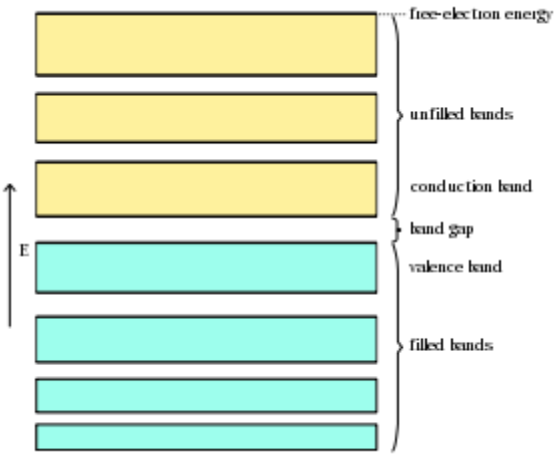
This can be continued with more atoms. Note: This picture shows a metal, not an actual semiconductor.



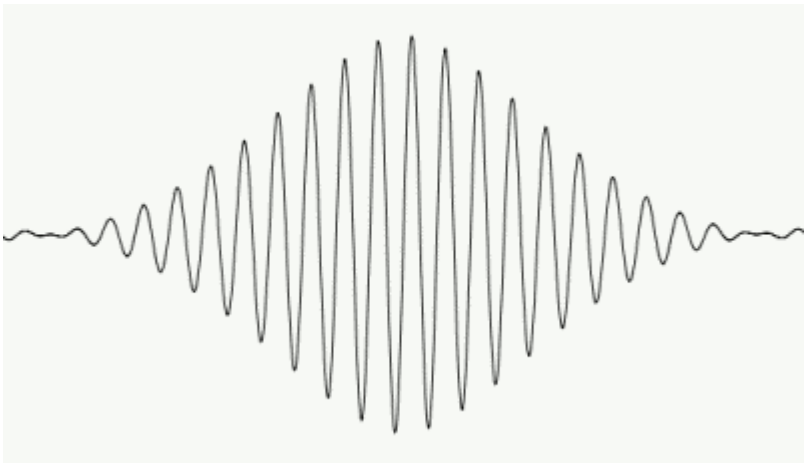
Continuing to add creates a crystal, which may then be cut into a tape and fused together at the ends to allow circular currents.



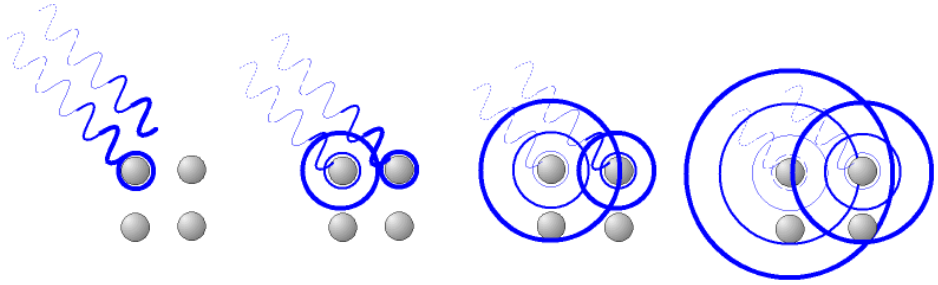
For this regular solid the band structure can be calculated or measured.



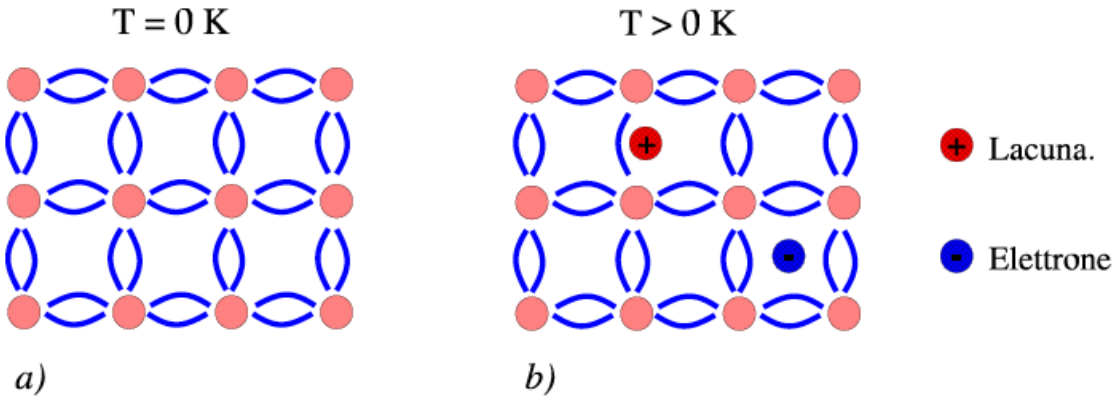
Integrating over the k axis gives the bands of a semiconductor showing a full valence band and an empty conduction band. Generally stopping at the vacuum level is undesirable, because some people want to calculate: photoemission, inverse photoemission



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.

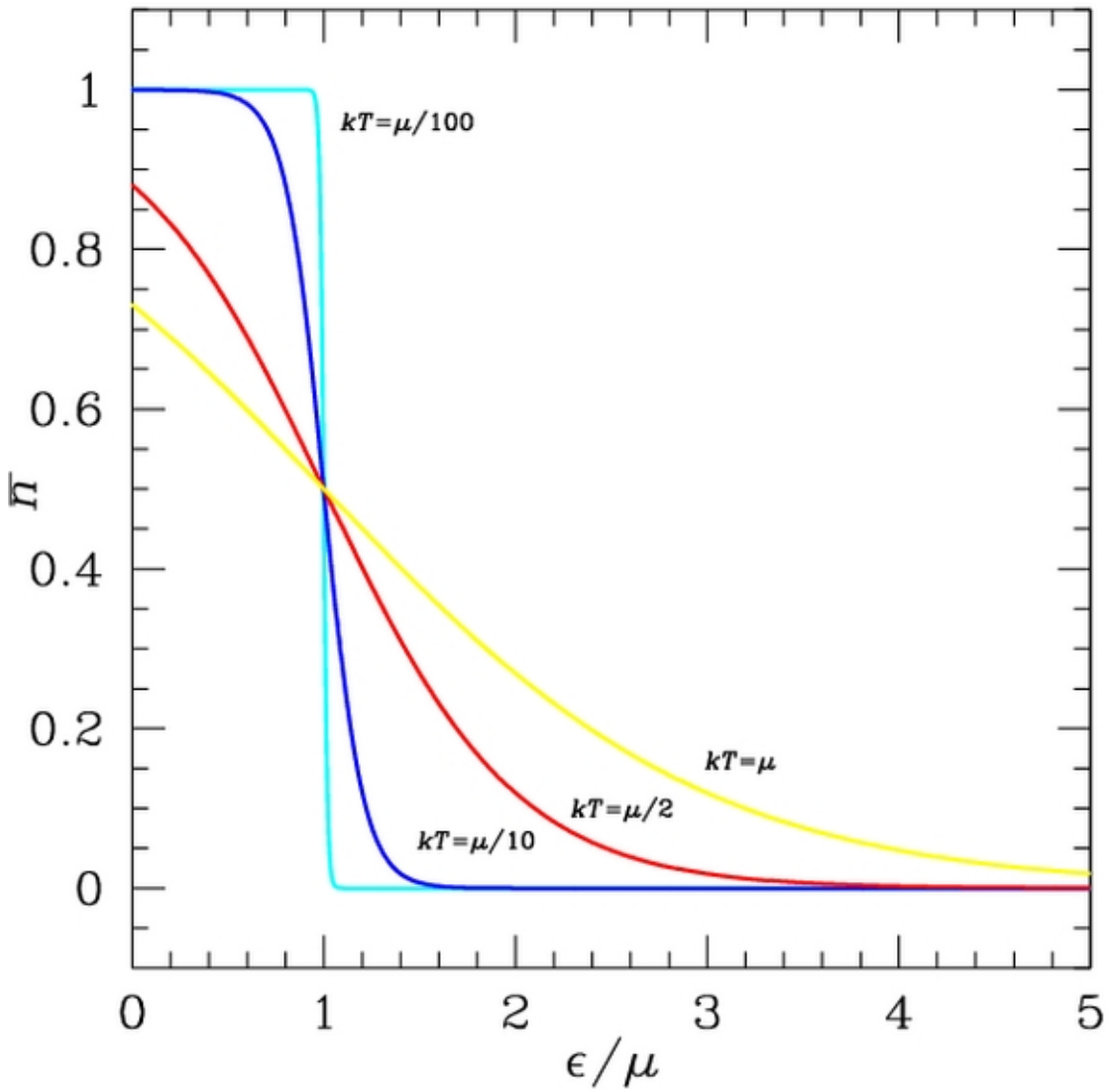
With covalent bonds, an electron moves by hopping to a neighboring bond. The Pauli exclusion principle requires the electron to be lifted into the higher anti-bonding state of that bond. For delocalized states, for example in one dimension – that is in a nanowire, for every energy there is a state with 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,

$$n = p = n_i.$$

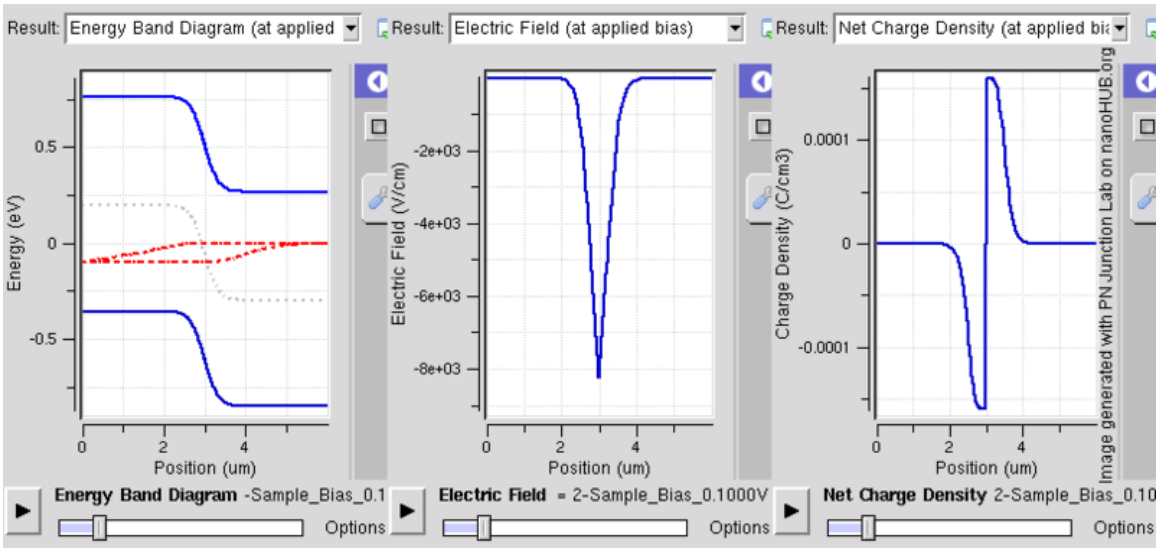
If we have a non-intrinsic semiconductor in thermal equilibrium the relation becomes:

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

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/cm³. Doping concentration for silicon semiconductors may range anywhere from 10^{13} cm^{-3} to 10^{18} cm^{-3}. Doping concentration above about 10^{18} cm^{-3} is considered degenerate at room temperature. Degenerately doped silicon contains a proportion of impurity to silicon on the order of parts per thousand. This proportion may be reduced to parts per billion in very lightly doped silicon. Typical concentration values fall somewhere in this range and are tailored to produce the desired properties in the device that the semiconductor is intended for.

Effect on band structure



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

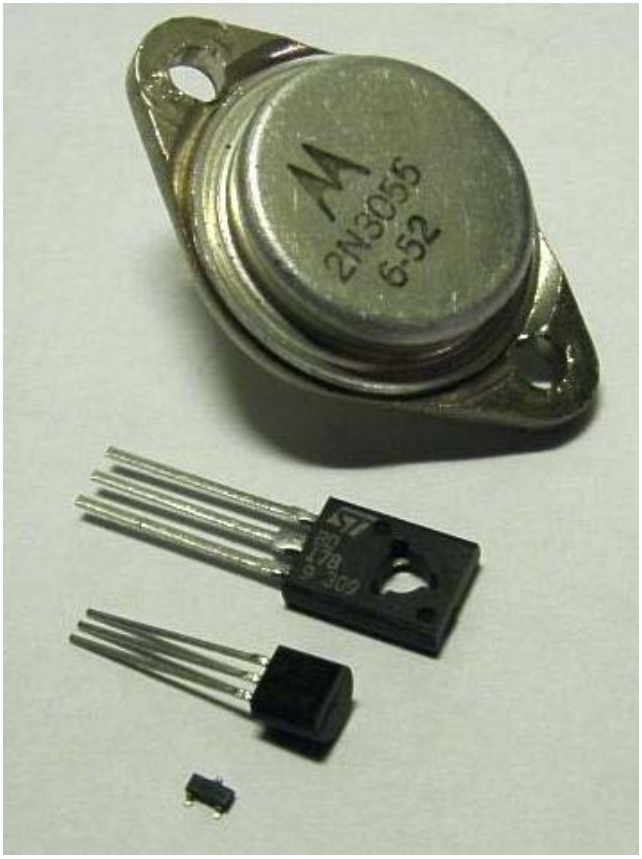
Semiconductors with predictable, reliable electronic properties are necessary for mass production. The level of chemical purity needed is extremely high because the presence of impurities even in very small proportions can have large effects on the properties of the material. A high degree of crystalline 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.

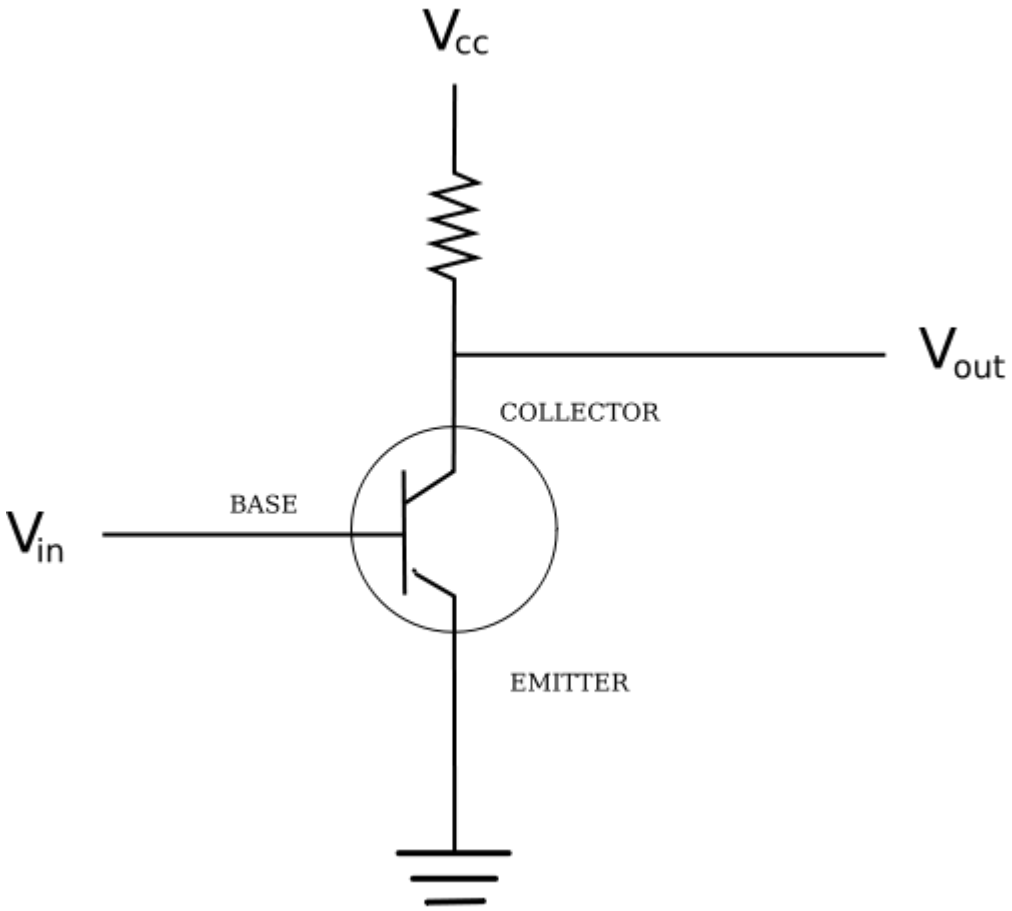
Although several companies each produce over a billion individually packaged (known as *discrete*) transistors every year, the vast majority of transistors now produced are in integrated circuits (often shortened to *IC*, *microchips* or simply *chips*), along with diodes, resistors, capacitors and other 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.

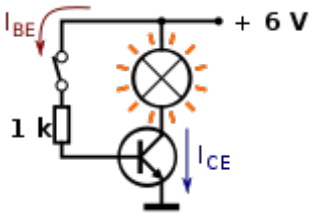
The essential usefulness of a transistor comes from its ability to use a small signal applied between one pair of its terminals to control a much larger signal at another pair of terminals. This property is called gain. A transistor can control its output in proportion to the input signal; that is, it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

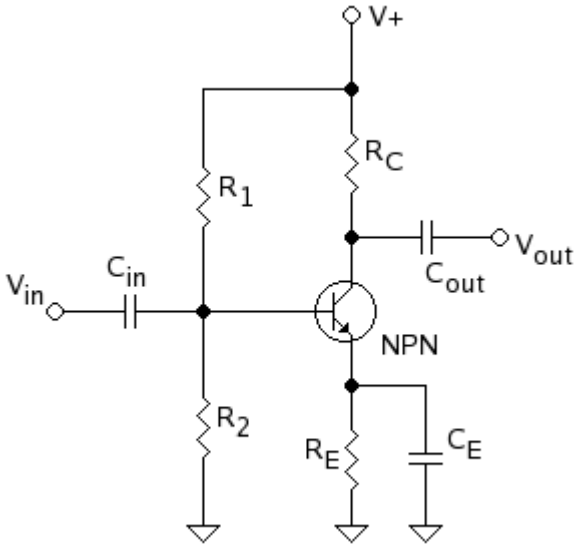
Transistors are commonly used as electronic switches, both for high-power applications such as switched-mode power supplies and for low-power applications such as logic gates.

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

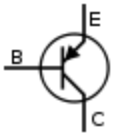
The key advantages that have allowed transistors to replace their vacuum tube predecessors in most applications are

- 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

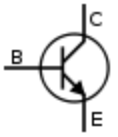
- Silicon transistors do not operate at voltages higher than about 1,000 volts (SiC devices can be operated as high as 3,000 volts). In contrast, vacuum tubes have been developed that can be operated at tens of thousands of volts.
- High-power, high-frequency operation, such as that used in over-the-air 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

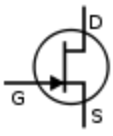


PNP

P-channel



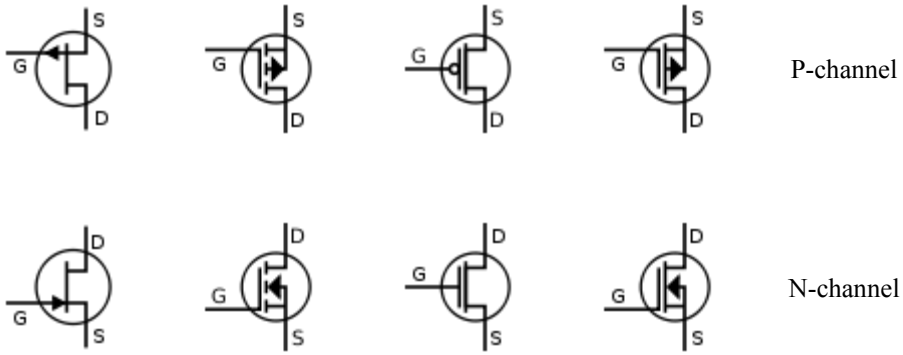
NPN



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:

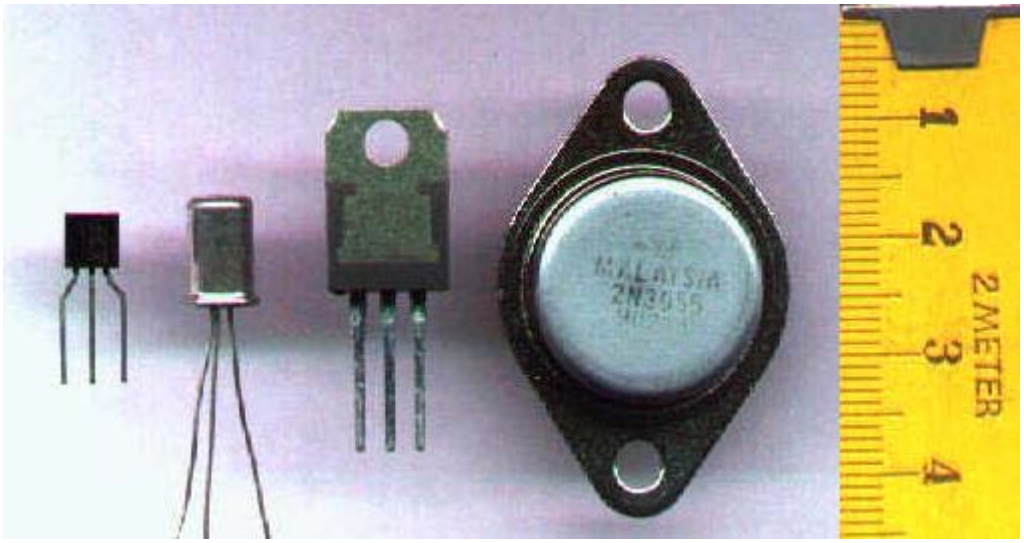
- Its maximum temperature is limited;
- it has relatively high leakage current;
- it cannot withstand high voltages;
- it is less suitable for fabricating integrated circuits.

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.

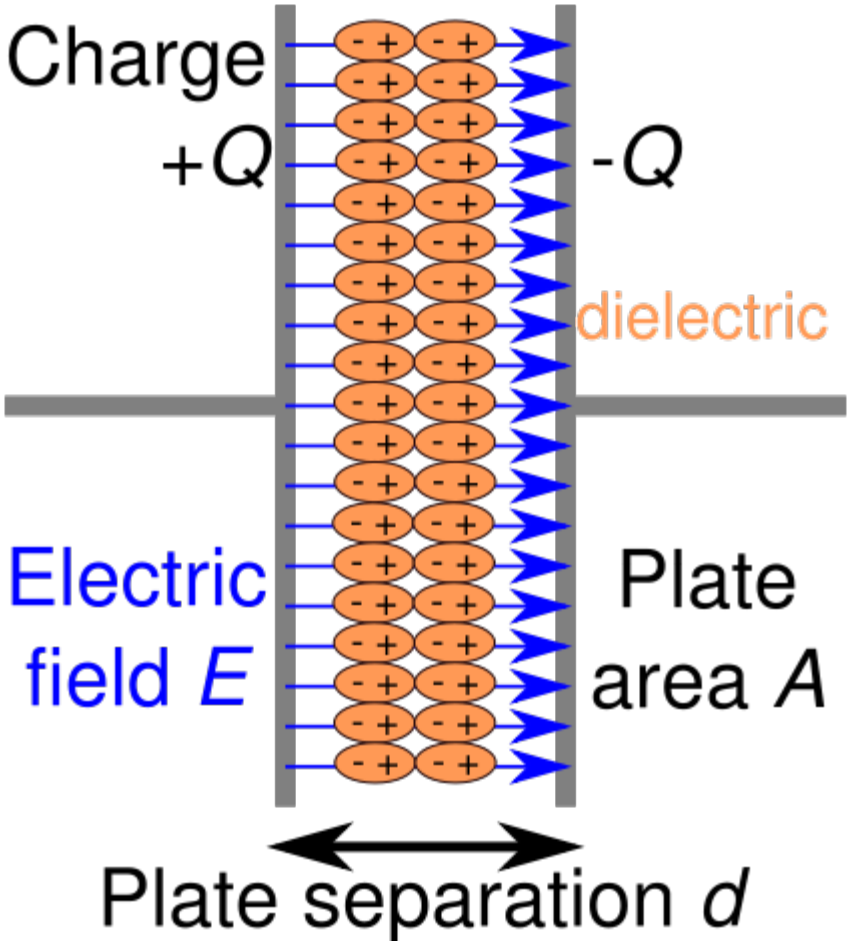
In October 1745, Ewald Georg von Kleist of Pomerania in Germany found that charge could be stored by connecting a high voltage electrostatic generator by a wire to a volume of water in a hand-held glass jar. Von Kleist's hand and the water acted as conductors and the jar as a dielectric (although details 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.

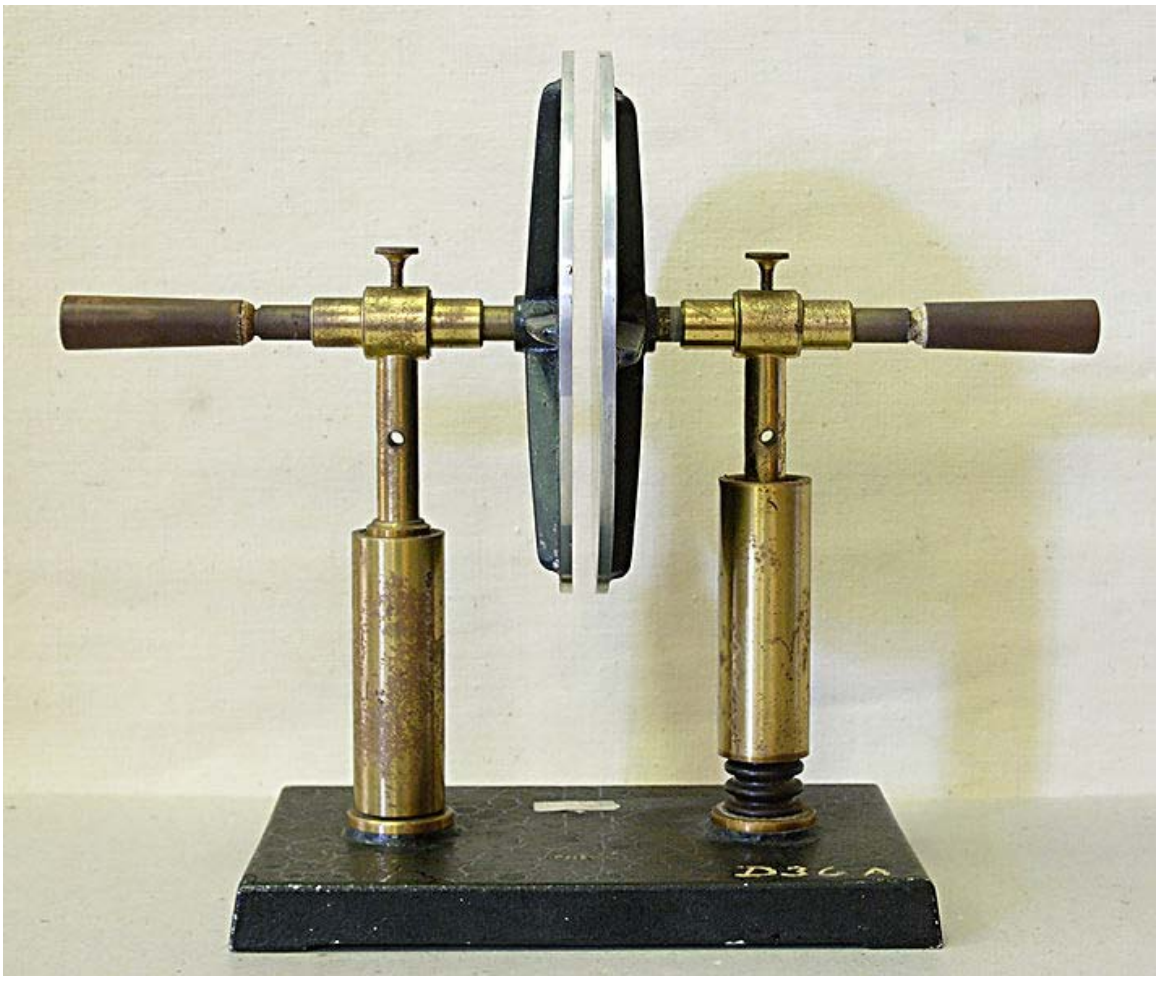
Leyden jars or more powerful devices employing flat glass plates alternating with foil conductors were used exclusively up until about 1900, when the invention of wireless (radio) created a demand for standard capacitors, and the steady move to higher frequencies required capacitors with lower inductance. 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,

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

.

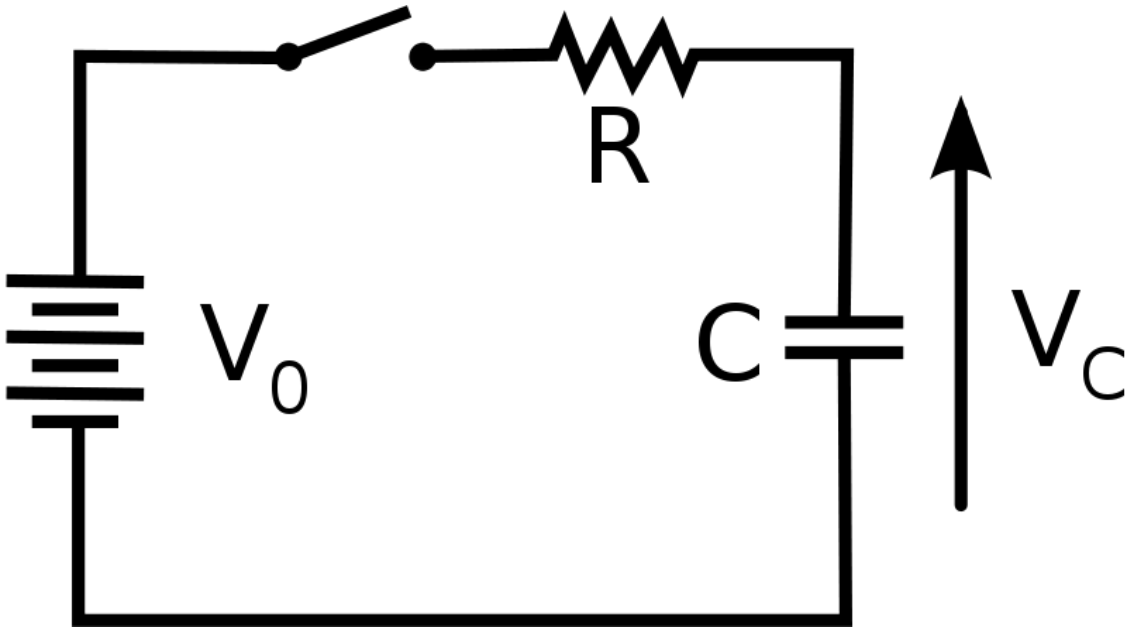
Taking the derivative of this, and multiplying by C, yields the derivative form,

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

.

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

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

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

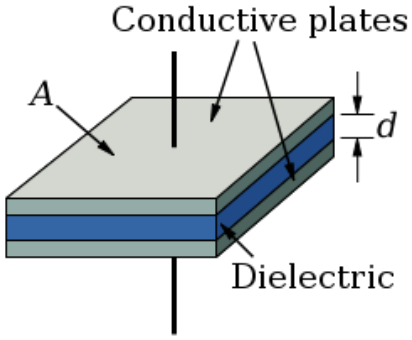
$$X = -\frac{1}{\omega C} = -\frac{1}{2\pi f C}$$
$$Z = \frac{1}{j\omega C} = -\frac{j}{\omega C} = -\frac{j}{2\pi f C}$$

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.

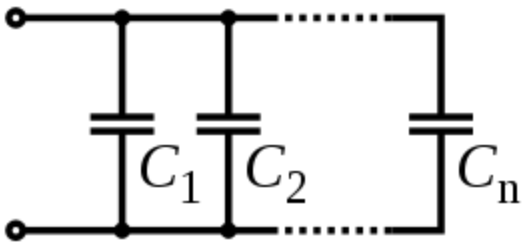
The simplest capacitor consists of two parallel conductive plates separated by a dielectric with permittivity ε (such as air). The model may also be used to make qualitative predictions for other device geometries. The plates are considered to extend uniformly over an area A and a charge density $\pm\rho = \pm Q/A$ 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.

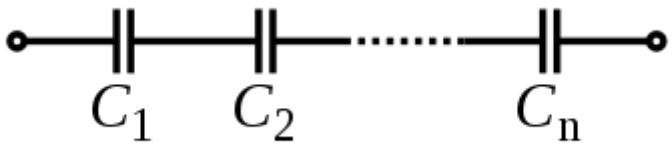
Networks

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,

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

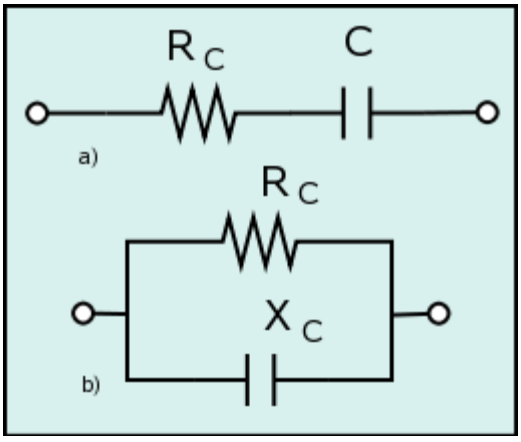
The maximum energy that can be stored safely in a capacitor is limited by the breakdown voltage. Due to the scaling of capacitance and breakdown voltage with dielectric

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.

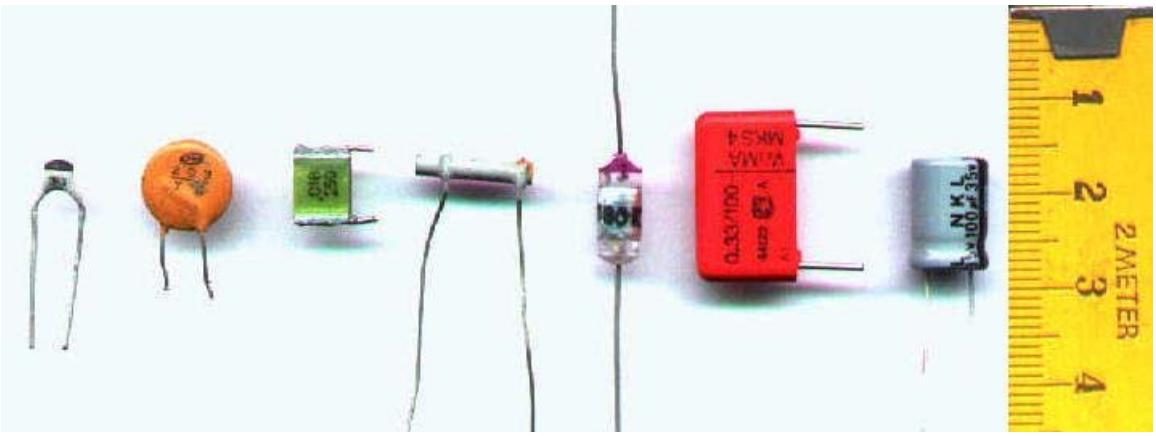
Capacitor types

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.

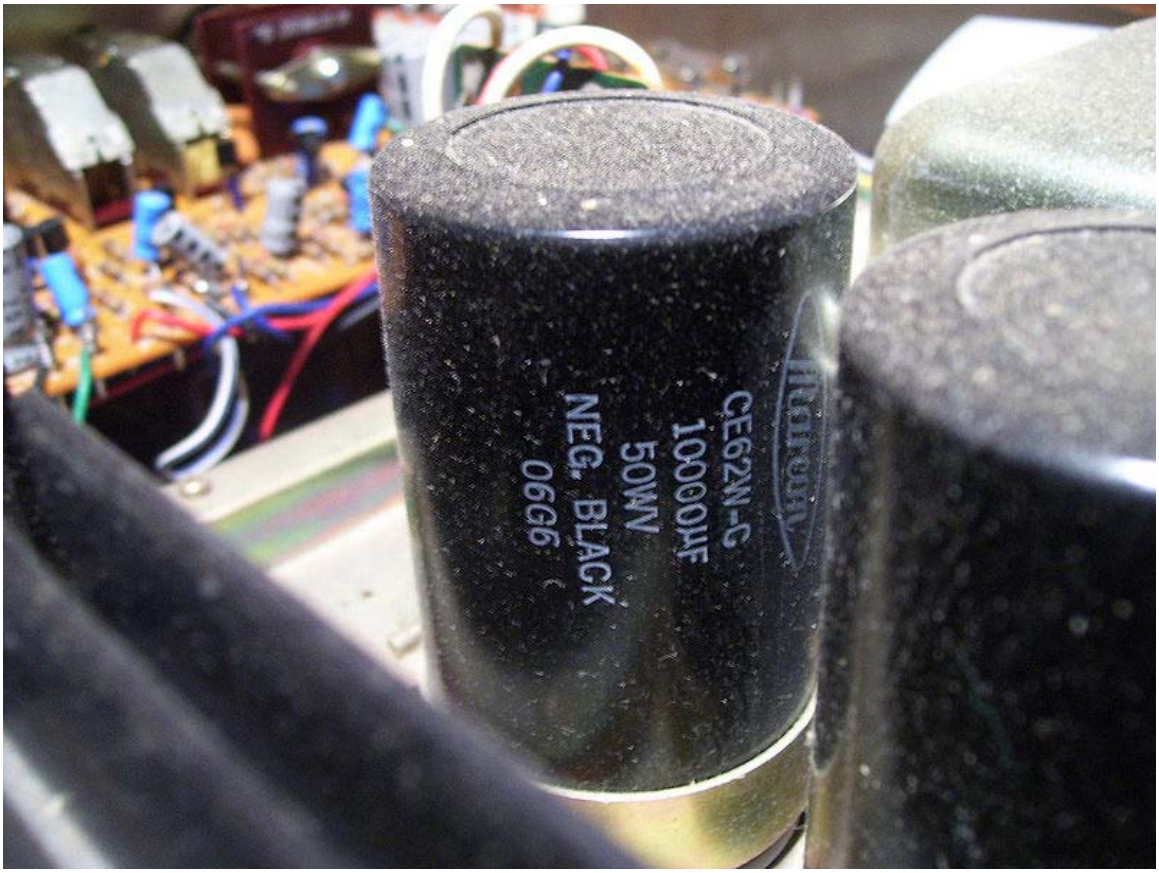
In car audio systems, large capacitors store energy for the amplifier to use on demand. Also for a flash tube a capacitor is used to hold the high voltage.

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.

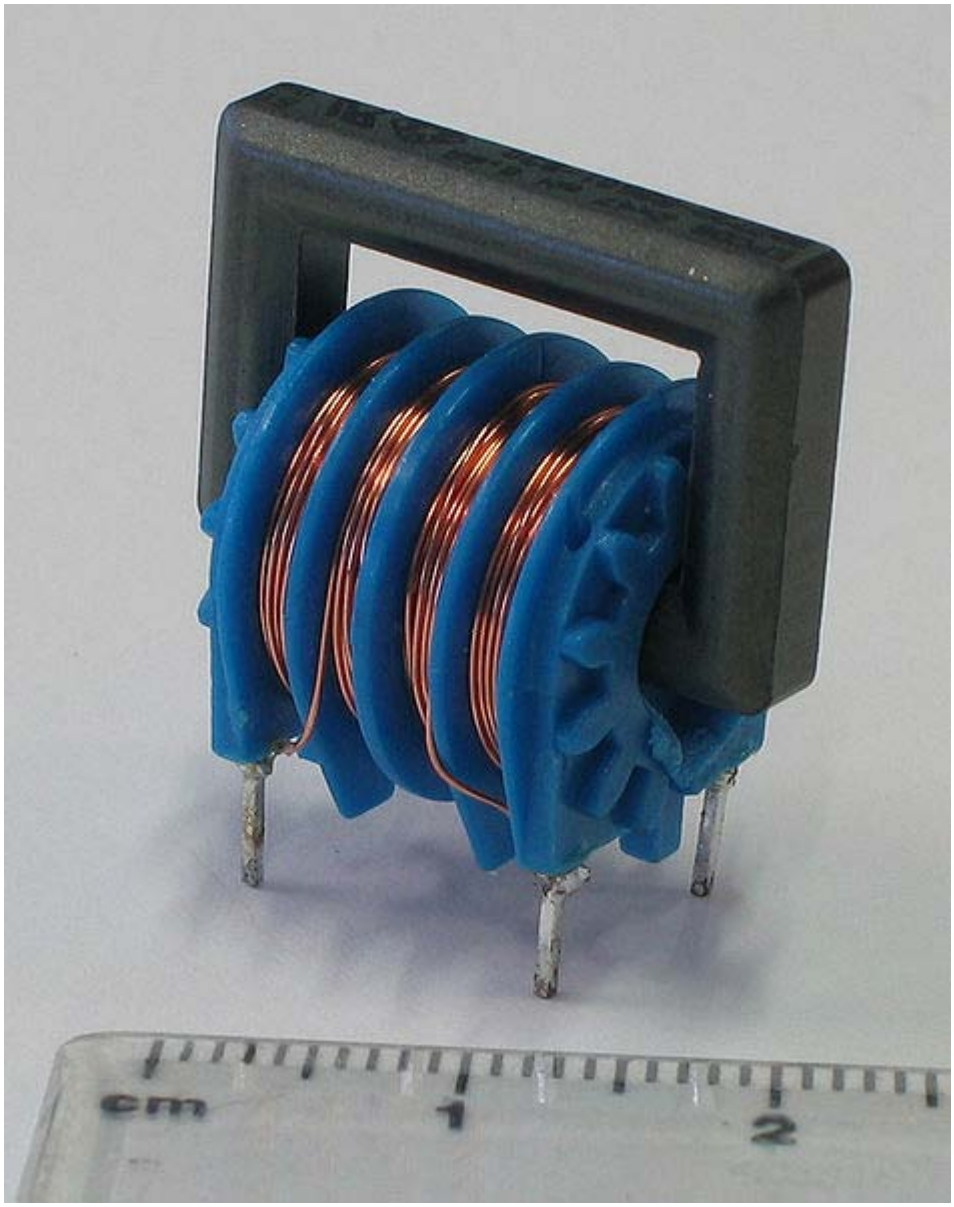
Overview

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.

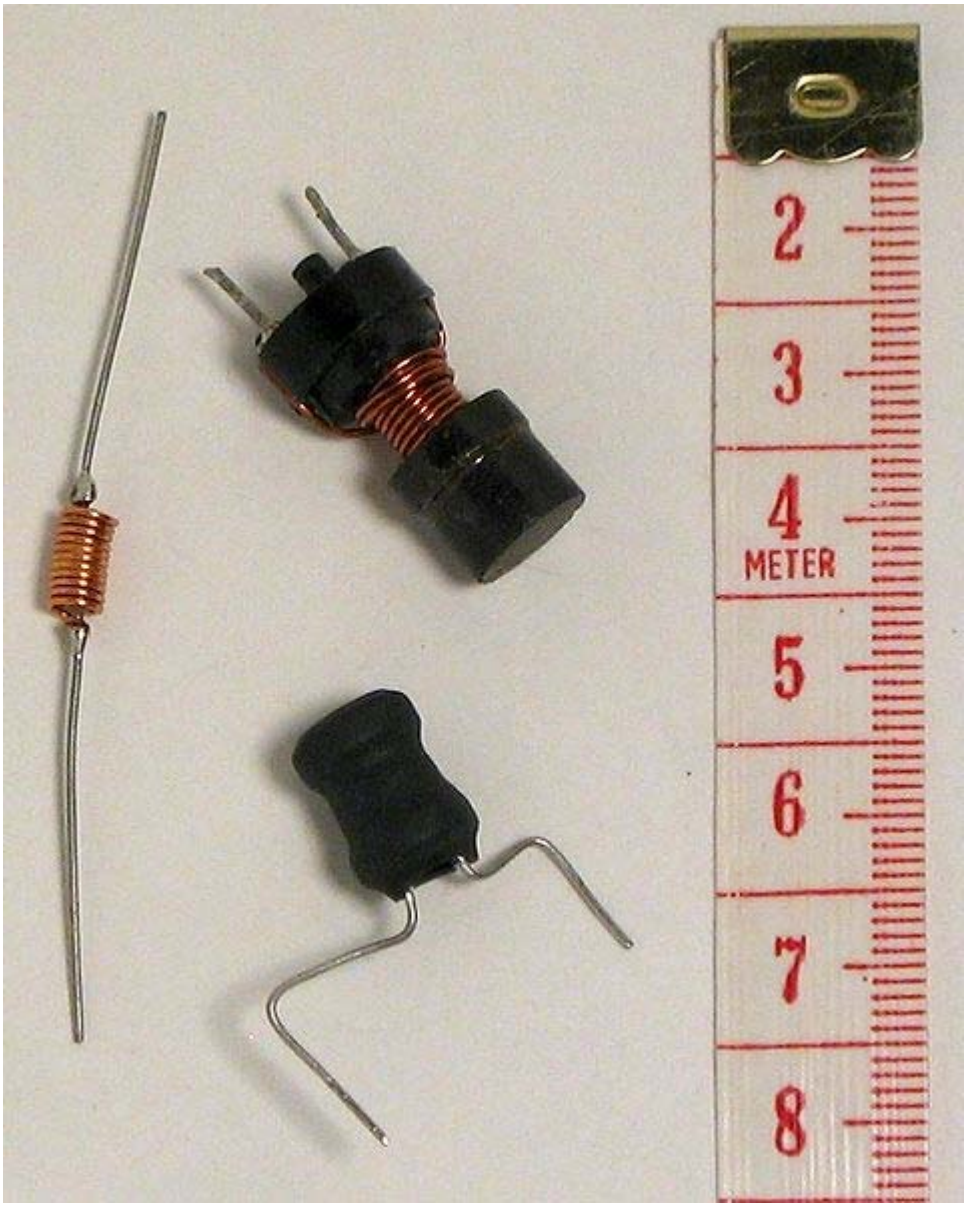
Two (or more) inductors that have coupled magnetic flux form a transformer, which is a fundamental component of every electric utility power grid. The efficiency of a transformer may decrease as the frequency increases due to eddy currents in the core material and skin effect on the windings. Size of 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*)

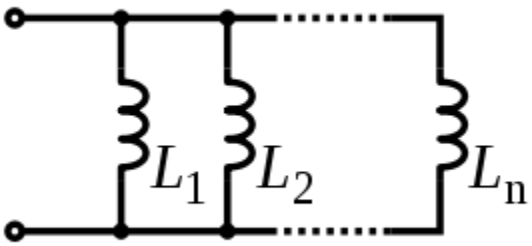
- or by adding a current source in parallel with the inductor, having the value:

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

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

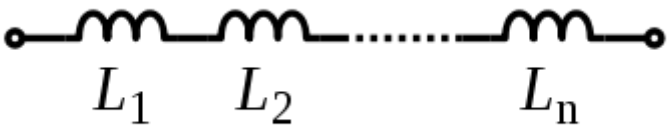
Inductor networks

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)	

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

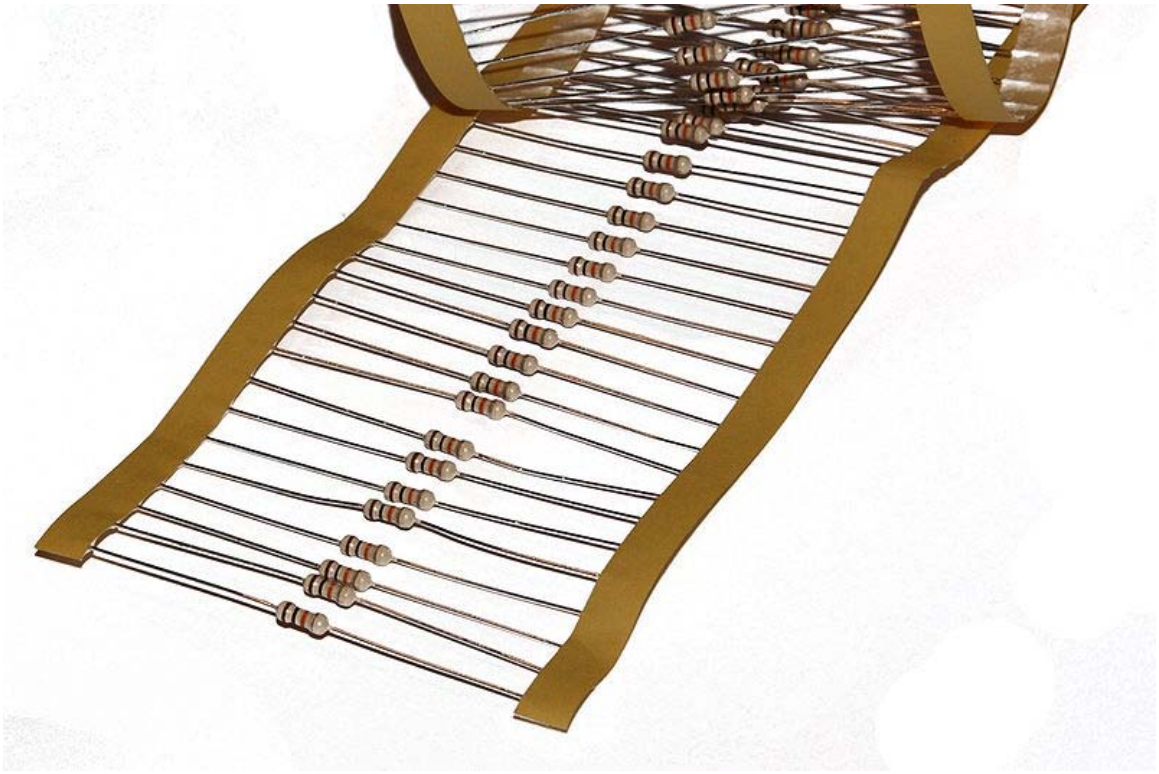
		(in) N = number of turns d = depth of coil (outer radius minus inner radius) (in) L = inductance (H) μ_0 = permeability of free space = $4\pi \times 10^{-7}$ H/m μ_r = relative permeability of core material r = radius of coil winding (m) N = number of turns D = overall diameter of toroid (m)
Toroidal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

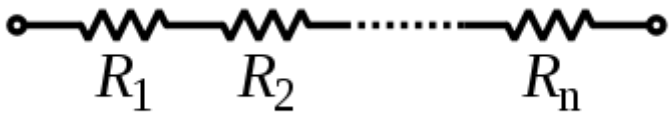
$$I = \frac{V}{R}$$

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

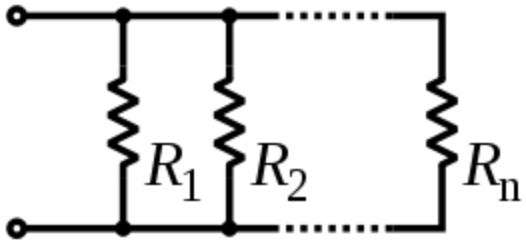
In a series configuration, the current through all of the resistors is the same, but the voltage across each resistor will be in proportion to its resistance. The potential difference (voltage) seen across the network is the sum of those voltages, thus the total resistance can be found as the sum of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



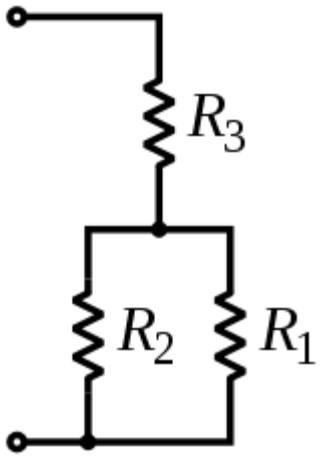
$$\frac{1}{R_{\text{eq}}} = \frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_n}$$

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:

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

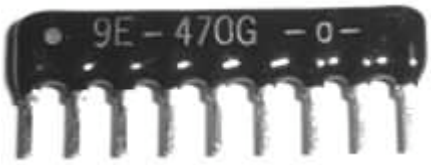
Practical resistors are rated according to their maximum power dissipation. The vast majority of resistors used in electronic circuits absorb much less than a watt of electrical power and require no attention to their power rating. Such resistors in their discrete form, including most of the packages 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.

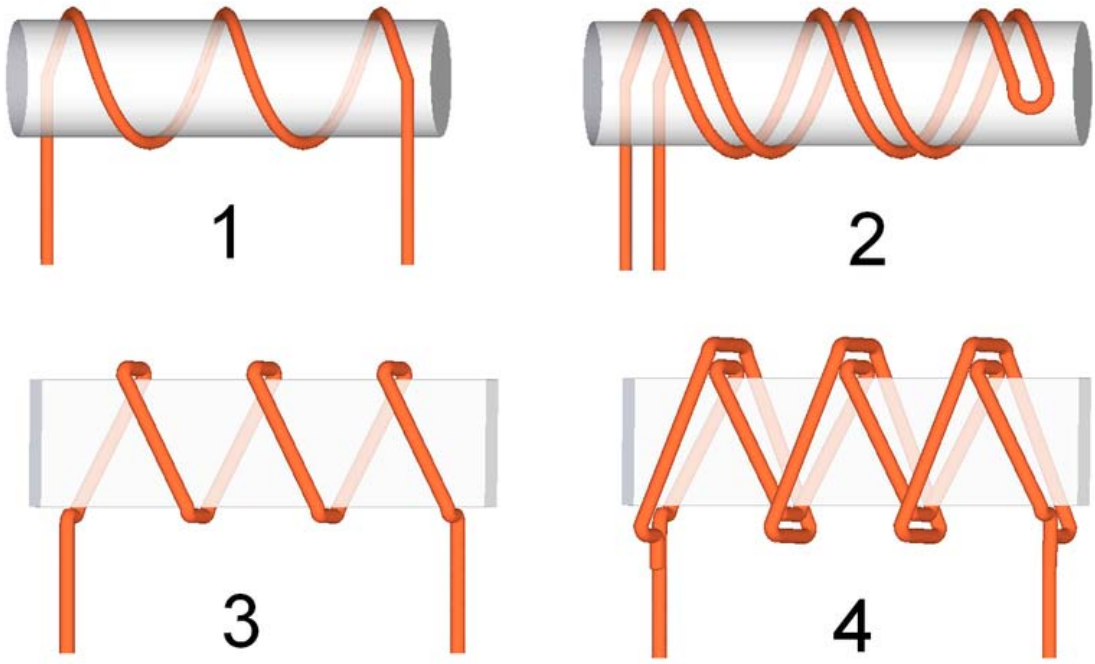
Metal film resistors are usually coated with nickel chromium (NiCr), but might be coated with any of the cermet materials listed above for thin film resistors. Unlike thin film resistors, the material may be applied using different techniques than sputtering (though that is one such technique). Also,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

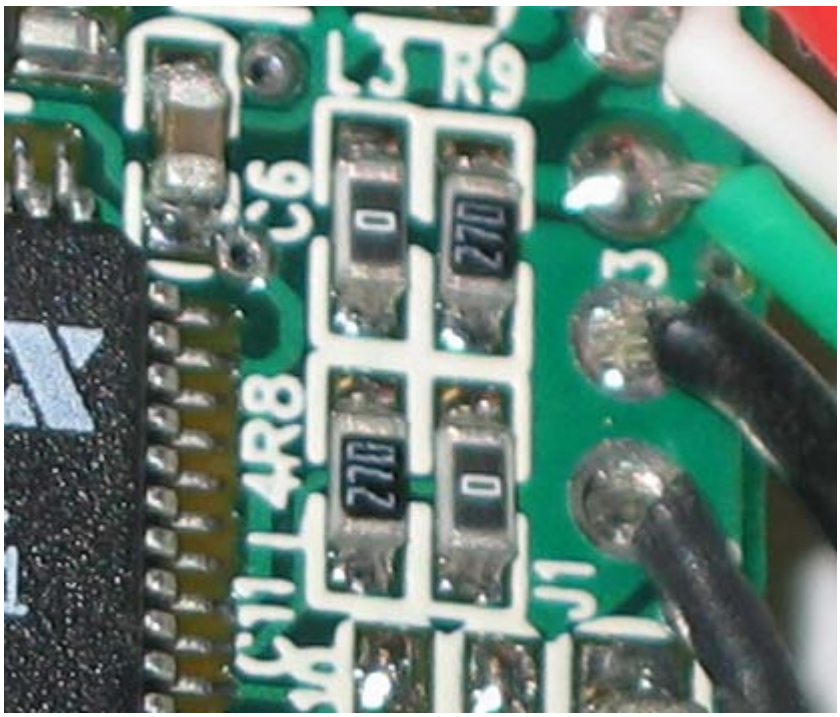
tolerance, with 12, 24, 96, and 192 different values within each decade. The actual values used are in the IEC 60063 lists of preferred numbers.

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

Diode

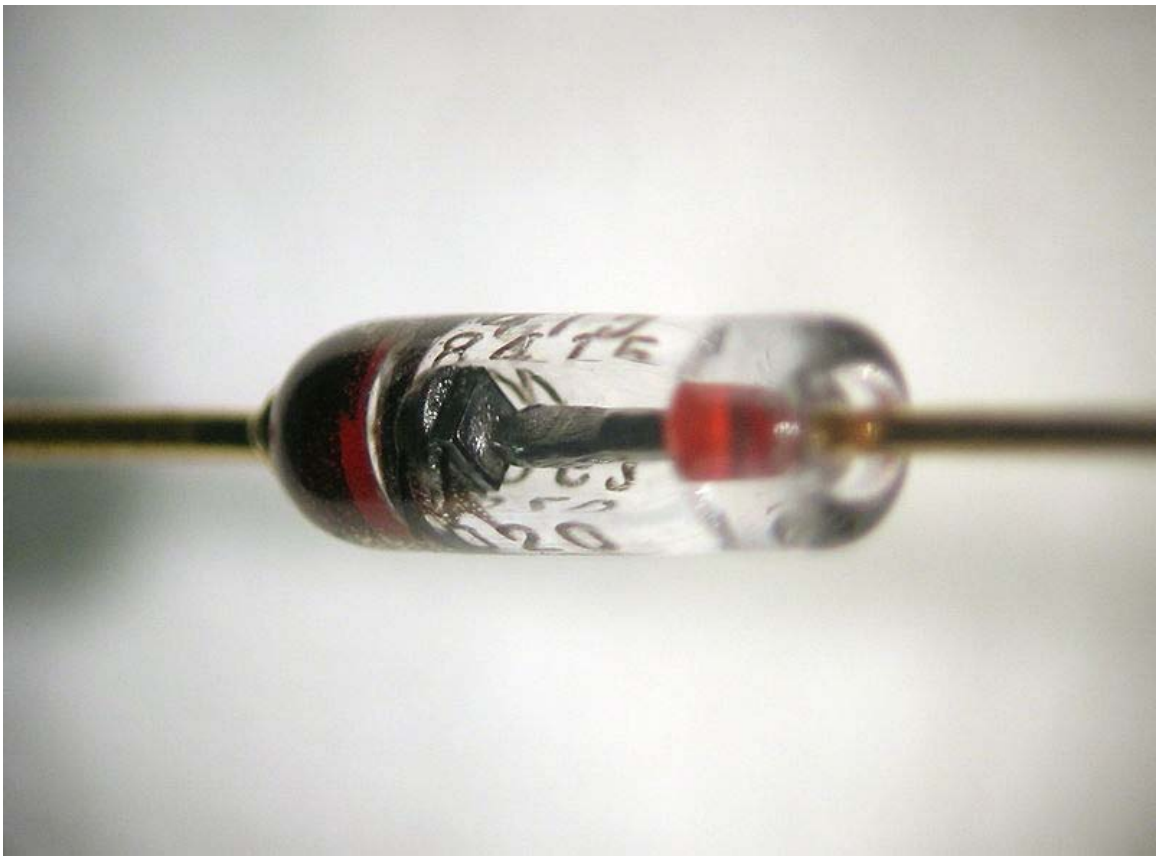


Figure 1: Closeup of a diode, showing the square shaped semiconductor crystal *(black object on left)*.

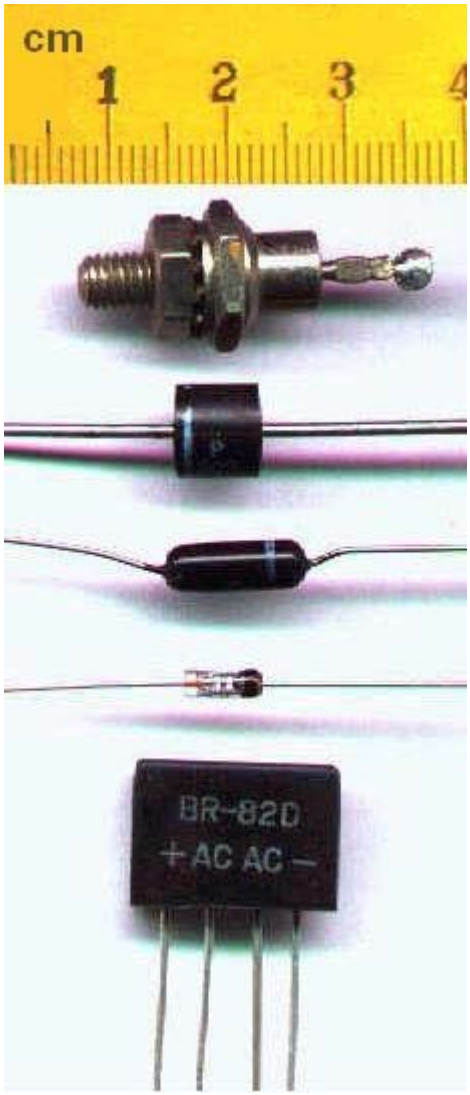


Figure 2: Various semiconductor diodes. Bottom: A bridge rectifier. In most diodes, a white or black painted band identifies the cathode terminal, that is, the terminal which conventional current flows out of when the diode is conducting.

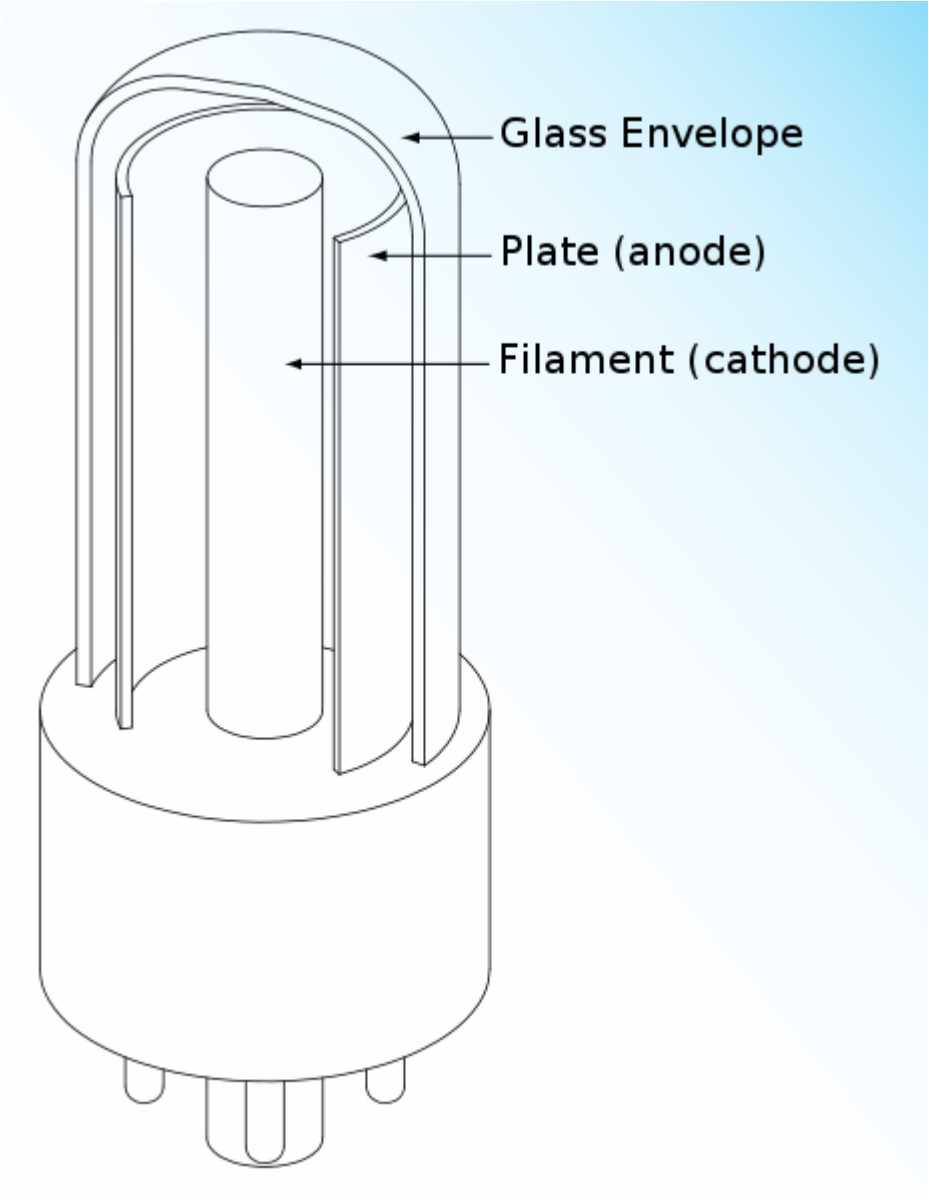


Figure 3: Structure of a vacuum tube diode. The filament may be bare, or more commonly (as shown here), embedded within and insulated from an enclosing cathode

In electronics, a **diode** is a two-terminal electronic component that conducts electric current in only one direction. The term usually refers to a **semiconductor diode**, the most common type today. This is a crystalline piece of semiconductor material connected to two electrical terminals. A **vacuum tube diode** (now little used except in some high-power technologies) is a vacuum tube with two electrodes: a plate and a cathode.

The most common function of a diode is to allow an electric current to pass in one direction (called the diode's *forward* direction) while blocking current in the opposite direction (the *reverse* direction). Thus, the diode can be thought of as an electronic version of a check valve. This unidirectional behavior is called rectification, and is used

to convert alternating current to direct current, and to extract modulation from radio signals in radio receivers.

However, diodes can have more complicated behavior than this simple on-off action. This is due to their complex non-linear electrical characteristics, which can be tailored by varying the construction of their P-N junction. These are exploited in special purpose diodes that perform many different functions. For example, specialized diodes are used to regulate voltage (Zener diodes), to electronically tune radio and TV receivers (varactor diodes), to generate radio frequency oscillations (tunnel diodes), and to produce light (light emitting diodes). Tunnel diodes exhibit negative resistance, which makes them useful in some types of circuits.

Diodes were the first semiconductor electronic devices. The discovery of crystals' rectifying abilities was made by German physicist Ferdinand Braun in 1874. The first semiconductor diodes, called cat's whisker diodes, developed around 1906, were made of mineral crystals such as galena. Today most diodes are made of silicon, but other semiconductors such as germanium are sometimes used.

History

Although the crystal semiconductor diode was popular before the thermionic diode, thermionic and solid state diodes were developed in parallel.

In 1873 Frederick Guthrie discovered the basic principle of operation of thermionic diodes. Guthrie discovered that a positively charged electroscope could be discharged by bringing a grounded piece of white-hot metal close to it (but not actually touching it). The same did not apply to a negatively charged electroscope, indicating that the current flow was only possible in one direction.

Thomas Edison independently rediscovered the principle on February 13, 1880. At the time, Edison was investigating why the filaments of his carbon-filament light bulbs nearly always burned out at the positive-connected end. He had a special bulb made with a metal plate sealed into the glass envelope. Using this device, he confirmed that an invisible current flowed from the glowing filament through the vacuum to the metal plate, but only when the plate was connected to the positive supply.

Edison devised a circuit where his modified light bulb effectively replaced the resistor in a DC voltmeter. Edison was awarded a patent for this invention in 1884. There was no apparent practical use for such a device at the time. So, the patent application was most likely simply a precaution in case someone else did find a use for the so-called Edison effect.

About 20 years later, John Ambrose Fleming (scientific adviser to the Marconi Company and former Edison employee) realized that the Edison effect could be used as a precision radio detector. Fleming patented the first true thermionic diode in Britain on November 16, 1904 (followed by U.S. Patent 803,684 in November 1905).

In 1874 German scientist Karl Ferdinand Braun discovered the "unilateral conduction" of crystals. Braun patented the crystal rectifier in 1899. Copper oxide and selenium rectifiers were developed for power applications in the 1930s.

Indian scientist Jagadish Chandra Bose was the first to use a crystal for detecting radio waves in 1894. The crystal detector was developed into a practical device for wireless radio reception by Greenleaf Whittier Pickard, who invented a silicon crystal detector in 1903 and received a patent for it on November 20, 1906. Other experimenters tried a variety of other substances, of which the most widely used was the mineral galena (lead sulfide). Other substances offered slightly better performance, but galena was most widely used because it had the advantage of being cheap and easy to obtain. The crystal detector in these early radio sets consisted of an adjustable wire point-contact (the so-called "cat's whisker") which could be manually moved over the face of the crystal in order to obtain optimum signal. This troublesome device was quickly superseded by thermionic diodes, but the crystal detector later returned to dominant use with the advent of inexpensive fixed-germanium diodes in the 1950s.

At the time of their invention, such devices were known as rectifiers. In 1919, William Henry Eccles coined the term ***diode*** from the Greek roots *dia*, meaning “through”, and *ode* (from ὁδος), meaning “path”.

Thermionic and gaseous state diodes

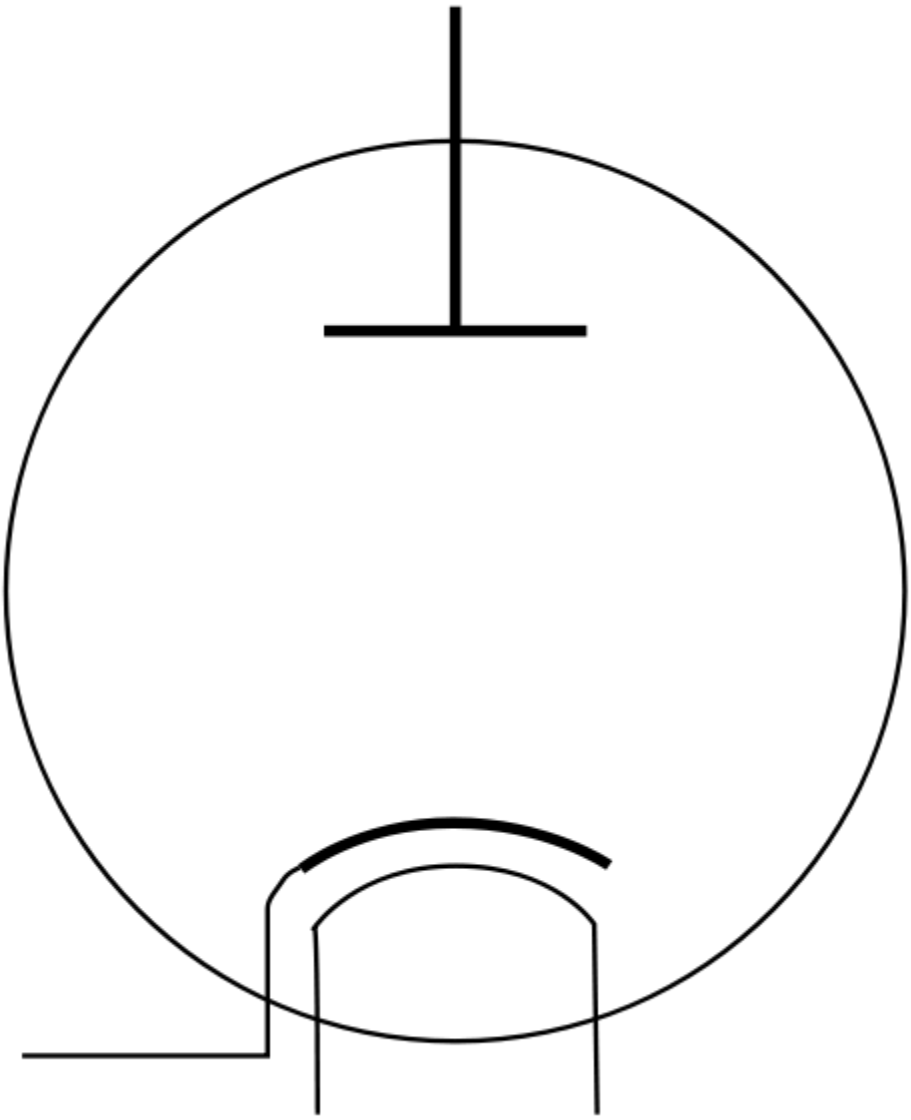


Figure 4: The symbol for an indirect heated vacuum tube diode. From top to bottom, the components are the anode, the cathode, and the heater filament.

Thermionic diodes are thermionic-valve devices (also known as vacuum tubes, tubes, or valves), which are arrangements of electrodes surrounded by a vacuum within a glass envelope. Early examples were fairly similar in appearance to incandescent light bulbs.

In thermionic valve diodes, a current through the heater filament indirectly heats the cathode, another internal electrode treated with a mixture of barium and strontium oxides, which are oxides of alkaline earth metals; these substances are chosen because they have a small work function. (Some valves use direct heating, in which a tungsten filament acts

as both heater and cathode.) The heat causes thermionic emission of electrons into the vacuum. In forward operation, a surrounding metal electrode called the anode is positively charged so that it electrostatically attracts the emitted electrons. However, electrons are not easily released from the unheated anode surface when the voltage polarity is reversed. Hence, any reverse flow is negligible.

For much of the 20th century, thermionic valve diodes were used in analog signal applications, and as rectifiers in many power supplies. Today, valve diodes are only used in niche applications such as rectifiers in electric guitar and high-end audio amplifiers as well as specialized high-voltage equipment.

Semiconductor diodes

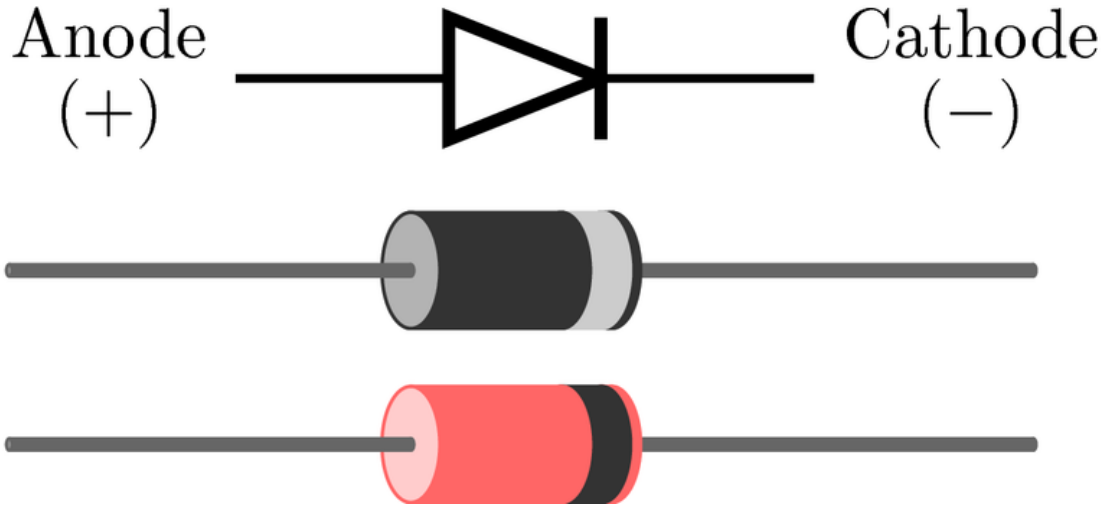


Figure 7: Typical diode packages in same alignment as diode symbol. Thin bar depicts the cathode.

A modern semiconductor diode is made of a crystal of semiconductor like silicon that has impurities added to it to create a region on one side that contains negative charge carriers (electrons), called n-type semiconductor, and a region on the other side that contains positive charge carriers (holes), called p-type semiconductor. The diode's terminals are attached to each of these regions. The boundary within the crystal between these two regions, called a PN junction, is where the action of the diode takes place. The crystal conducts a current of electrons in a direction from the N-type side (called the cathode) to the P-type side (called the anode), but not in the opposite direction; that is, a conventional current flows from anode to cathode (opposite to the electron flow, since electrons have negative charge).

Another type of semiconductor diode, the Schottky diode, is formed from the contact between a metal and a semiconductor rather than by a p-n junction.

Current–voltage characteristic

A semiconductor diode's behavior in a circuit is given by its current–voltage characteristic, or I–V graph. The shape of the curve is determined by the transport of charge carriers through the so-called *depletion layer* or *depletion region* that exists at the p-n junction between differing semiconductors. When a p-n junction is first created, conduction band (mobile) electrons from the N-doped region diffuse into the P-doped region where there is a large population of holes (vacant places for electrons) with which the electrons "recombine". When a mobile electron recombines with a hole, both hole and electron vanish, leaving behind an immobile positively charged donor (dopant) on the N-side and negatively charged acceptor (dopant) on the P-side. The region around the p-n junction becomes depleted of charge carriers and thus behaves as an insulator.

However, the width of the depletion region (called the depletion width) cannot grow without limit. For each electron-hole pair that recombines, a positively charged dopant ion is left behind in the N-doped region, and a negatively charged dopant ion is left behind in the P-doped region. As recombination proceeds more ions are created, an increasing electric field develops through the depletion zone which acts to slow and then finally stop recombination. At this point, there is a "built-in" potential across the depletion zone.

If an external voltage is placed across the diode with the same polarity as the built-in potential, the depletion zone continues to act as an insulator, preventing any significant electric current flow (unless electron/hole pairs are actively being created in the junction by, for instance, light). This is the *reverse bias* phenomenon. However, if the polarity of the external voltage opposes the built-in potential, recombination can once again proceed, resulting in substantial electric current through the p-n junction (i.e. substantial numbers of electrons and holes recombine at the junction). For silicon diodes, the built-in potential is approximately 0.7 V (0.3 V for Germanium and 0.2 V for Schottky). Thus, if an external current is passed through the diode, about 0.7 V will be developed across the diode such that the P-doped region is positive with respect to the N-doped region and the diode is said to be "turned on" as it has a *forward bias*.

A diode's '***I–V characteristic'*** can be approximated by four regions of operation.

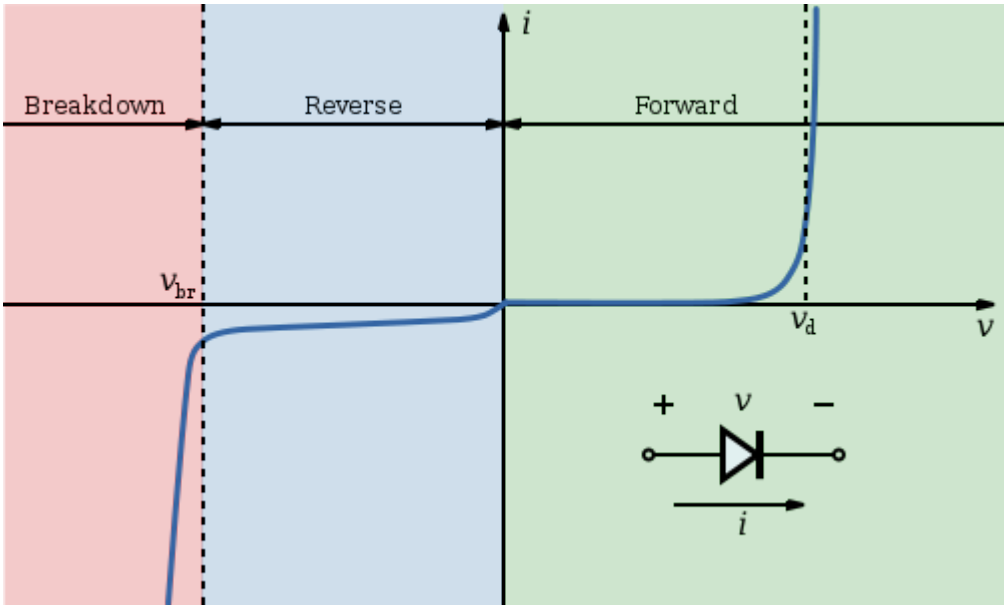


Figure 5: I–V characteristics of a P-N junction diode (not to scale).

At very large reverse bias , beyond the peak inverse voltage or PIV, a process called reverse breakdown occurs which causes a large increase in current (i.e. a large number of electrons and holes are created at, and move away from the pn junction) that usually damages the device permanently. The avalanche diode is deliberately designed for use in the avalanche region. In the zener diode, the concept of PIV is not applicable. A zener diode contains a heavily doped p-n junction allowing electrons to tunnel from the valence band of the p-type material to the conduction band of the n-type material, such that the reverse voltage is "clamped" to a known value (called the *zener voltage*), and avalanche does not occur. Both devices, however, do have a limit to the maximum current and power in the clamped reverse voltage region. Also, following the end of forward conduction in any diode, there is reverse current for a short time. The device does not attain its full blocking capability until the reverse current ceases.

The second region, at reverse biases more positive than the PIV, has only a very small reverse saturation current. In the reverse bias region for a normal P-N rectifier diode, the current through the device is very low (in the μA range). However, this is temperature dependent, and at sufficiently high temperatures, a substantial amount of reverse current can be observed (mA or more).

The third region is forward but small bias, where only a small forward current is conducted.

As the potential difference is increased above an arbitrarily defined "cut-in voltage" or "on-voltage" or "diode forward voltage drop (V_d)", the diode current becomes appreciable (the level of current considered "appreciable" and the value of cut-in voltage depends on the application), and the diode presents a very low resistance. The current–voltage curve is exponential. In a normal silicon diode at rated currents, the arbitrary

"cut-in" voltage is defined as 0.6 to 0.7 volts. The value is different for other diode types — Schottky diodes can be rated as low as 0.2 V, Germanium diodes 0.25 to 0.3 V, and red or blue light-emitting diodes (LEDs) can have values of 1.4 V and 4.0 V respectively.

At higher currents the forward voltage drop of the diode increases. A drop of 1 V to 1.5 V is typical at full rated current for power diodes.

Shockley diode equation

The *Shockley ideal diode equation* or the *diode law* (named after transistor co-inventor William Bradford Shockley, not to be confused with tetrode inventor Walter H. Schottky) gives the I–V characteristic of an ideal diode in either forward or reverse bias (or no bias). The equation is:

$$I = I_{\mathrm{S}} \left(e^{V_{\mathrm{D}}/(nV_{\mathrm{T}})} - 1 \right),$$

where

I is the diode current,
I_{S} is the reverse bias saturation current (or scale current),
V_{D} is the voltage across the diode,
V_{T} is the thermal voltage, and
n is the *ideality factor*, also known as the *quality factor* or sometimes *emission coefficient*. The ideality factor n varies from 1 to 2 depending on the fabrication process and semiconductor material and in many cases is assumed to be approximately equal to 1 (thus the notation n is omitted).

The thermal voltage V_{T} is approximately 25.85 mV at 300 K, a temperature close to "room temperature" commonly used in device simulation software. At any temperature it is a known constant defined by:

$$V_{\mathrm{T}} = \frac{kT}{q},$$

where k is the Boltzmann constant, T is the absolute temperature of the p-n junction, and q is the magnitude of charge on an electron (the elementary charge).

The *Shockley ideal diode equation* or the *diode law* is derived with the assumption that the only processes giving rise to the current in the diode are drift (due to electrical field), diffusion, and thermal recombination-generation. It also assumes that the recombination-generation (R-G) current in the depletion region is insignificant. This means that the Shockley equation doesn't account for the processes involved in reverse breakdown and photon-assisted R-G. Additionally, it doesn't describe the "leveling off" of the I–V curve at high forward bias due to internal resistance.

Under *reverse bias* voltages the exponential in the diode equation is negligible, and the current is a constant (negative) reverse current value of $-I_S$. The reverse *breakdown region* is not modeled by the Shockley diode equation.

For even rather small *forward bias* voltages the exponential is very large because the thermal voltage is very small, so the subtracted '1' in the diode equation is negligible and the forward diode current is often approximated as

$$I = I_S e^{V_D/(nV_T)}$$

The use of the diode equation in circuit problems is illustrated in diode modeling.

Reverse-recovery effect

Following the end of forward conduction in a PN type diode, a reverse current flows for a short time. The device does not attain its full blocking capability until the reverse current ceases.

The effect can be significant when switching large currents very quickly (di/dt on the order of 100 A/µs or more). A certain amount of "reverse recovery time" t_r (on the order of tens of nanoseconds) may be required to remove the "reverse recovery charge" Q_r (on the order of tens of nanocoulombs) from the diode. During this recovery time, the diode can actually conduct in the reverse direction. In certain real-world cases it can be important to consider the losses incurred by this non-ideal diode effect. However, when the slew rate of the current is not so severe (di/dt on the order of 10 A/µs or less), the effect can be safely ignored. For most applications, the effect is also negligible for Schottky diodes.

The reverse current ceases abruptly when the stored charge is depleted, which is exploited in step recovery diodes for generation of extremely short pulses.

Types of semiconductor diode

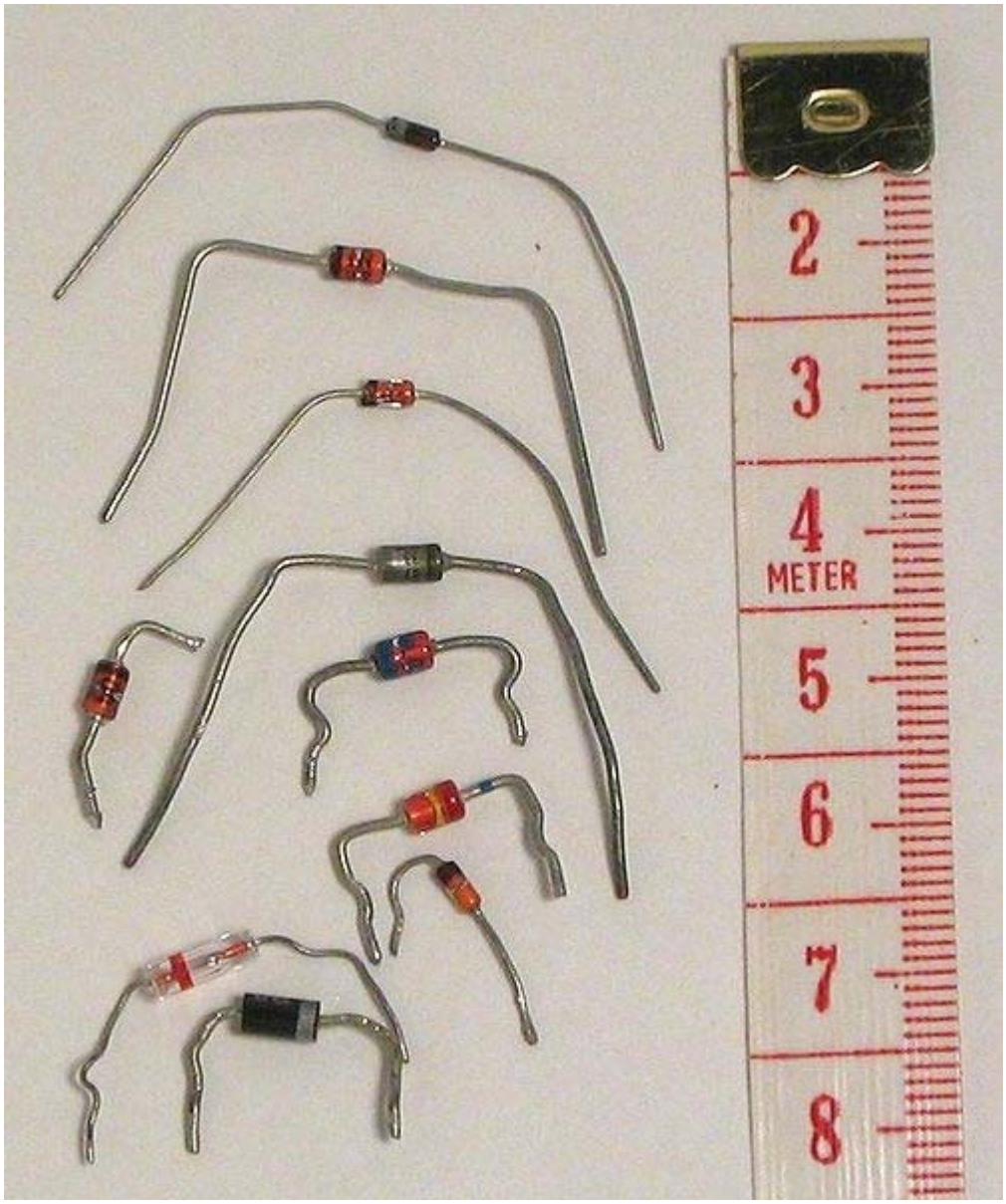


Figure 8: Several types of diodes. The scale is centimeters.

There are several types of junction diodes, which either emphasize a different physical aspect of a diode often by geometric scaling, doping level, choosing the right electrodes, are just an application of a diode in a special circuit, or are really different devices like the Gunn and laser diode and the MOSFET:

Normal (p-n) diodes, which operate as described above, are usually made of doped silicon or, more rarely, germanium. Before the development of modern silicon power rectifier diodes, cuprous oxide and later selenium was used; its low efficiency gave it a much higher forward voltage drop (typically 1.4 to 1.7 V per “cell”, with multiple cells

stacked to increase the peak inverse voltage rating in high voltage rectifiers), and required a large heat sink (often an extension of the diode's metal substrate), much larger than a silicon diode of the same current ratings would require. The vast majority of all diodes are the p-n diodes found in CMOS integrated circuits, which include two diodes per pin and many other internal diodes.

Avalanche diodes

Diodes that conduct in the reverse direction when the reverse bias voltage exceeds the breakdown voltage. These are electrically very similar to Zener diodes, and are often mistakenly called Zener diodes, but break down by a different mechanism, the *avalanche effect*. This occurs when the reverse electric field across the p-n junction causes a wave of ionization, reminiscent of an avalanche, leading to a large current. Avalanche diodes are designed to break down at a well-defined reverse voltage without being destroyed. The difference between the avalanche diode (which has a reverse breakdown above about 6.2 V) and the Zener is that the channel length of the former exceeds the "mean free path" of the electrons, so there are collisions between them on the way out. The only practical difference is that the two types have temperature coefficients of opposite polarities.

Cat's whisker or crystal diodes

These are a type of point-contact diode. The cat's whisker diode consists of a thin or sharpened metal wire pressed against a semiconducting crystal, typically galena or a piece of coal. The wire forms the anode and the crystal forms the cathode. Cat's whisker diodes were also called crystal diodes and found application in crystal radio receivers. Cat's whisker diodes are generally obsolete, but may be available from a few manufacturers.

Constant current diodes

These are actually a JFET with the gate shorted to the source, and function like a two-terminal current-limiter analog to the Zener diode, which is limiting voltage. They allow a current through them to rise to a certain value, and then level off at a specific value. Also called *CLDs*, *constant-current diodes*, *diode-connected transistors*, or *current-regulating diodes*.

Esaki or tunnel diodes

These have a region of operation showing negative resistance caused by quantum tunneling, allowing amplification of signals and very simple bistable circuits. Due to the high carrier concentration, tunnel diodes are very fast, may be used at low (mK) temperatures, high magnetic fields, and in high radiation environments. Because of these properties, they are often used in spacecraft.

Gunn diodes

These are similar to tunnel diodes in that they are made of materials such as GaAs or InP that exhibit a region of negative differential resistance. With appropriate biasing, dipole domains form and travel across the diode, allowing high frequency microwave oscillators to be built.

Light-emitting diodes (LEDs)

In a diode formed from a direct band-gap semiconductor, such as gallium arsenide, carriers that cross the junction emit photons when they recombine with the majority carrier on the other side. Depending on the material, wavelengths (or colors) from the infrared to the near ultraviolet may be produced. The forward potential of these diodes depends on the wavelength of the emitted photons: 1.2 V corresponds to red, 2.4 V to violet. The first LEDs were red and yellow, and higher-frequency diodes have been developed over time. All LEDs produce incoherent, narrow-spectrum light; “white” LEDs are actually combinations of three LEDs of a different color, or a blue LED with a yellow scintillator coating. LEDs can also be used as low-efficiency photodiodes in signal applications. An LED may be paired with a photodiode or phototransistor in the same package, to form an opto-isolator.

Laser diodes

When an LED-like structure is contained in a resonant cavity formed by polishing the parallel end faces, a laser can be formed. Laser diodes are commonly used in optical storage devices and for high speed optical communication.

Thermal diodes

This term is used both for conventional PN diodes used to monitor temperature due to their varying forward voltage with temperature, and for Peltier heat pumps for thermoelectric heating and cooling.. Peltier heat pumps may be made from semiconductor, though they do not have any rectifying junctions, they use the differing behaviour of charge carriers in N and P type semiconductor to move heat.

Photodiodes

All semiconductors are subject to optical charge carrier generation. This is typically an undesired effect, so most semiconductors are packaged in light blocking material. Photodiodes are intended to sense light(photodetector), so they are packaged in materials that allow light to pass, and are usually PIN (the kind of diode most sensitive to light). A photodiode can be used in solar cells, in photometry, or in optical communications. Multiple photodiodes may be packaged in a single device, either as a linear array or as a two-dimensional array. These arrays should not be confused with charge-coupled devices.

Point-contact diodes

These work the same as the junction semiconductor diodes described above, but their construction is simpler. A block of n-type semiconductor is built, and a conducting sharp-point contact made with some group-3 metal is placed in contact with the semiconductor. Some metal migrates into the semiconductor to make a small region of p-type semiconductor near the contact. The long-popular 1N34 germanium version is still used in radio receivers as a detector and occasionally in specialized analog electronics.

PIN diodes

A PIN diode has a central un-doped, or *intrinsic*, layer, forming a p-type/intrinsic/n-type structure. They are used as radio frequency switches and attenuators. They are also used as large volume ionizing radiation detectors and as photodetectors. PIN diodes are also used in power electronics, as their central layer can withstand high voltages. Furthermore, the PIN structure can be found in many power semiconductor devices, such as IGBTs, power MOSFETs, and thyristors.

Schottky diodes

Schottky diodes are constructed from a metal to semiconductor contact. They have a lower forward voltage drop than p-n junction diodes. Their forward voltage drop at forward currents of about 1 mA is in the range 0.15 V to 0.45 V, which makes them useful in voltage clamping applications and prevention of transistor saturation. They can also be used as low loss rectifiers although their reverse leakage current is generally higher than that of other diodes. Schottky diodes are majority carrier devices and so do not suffer from minority carrier storage problems that slow down many other diodes — so they have a faster "reverse recovery" than p-n junction diodes. They also tend to have much lower junction capacitance than p-n diodes which provides for high switching speeds and their use in high-speed circuitry and RF devices such as switched-mode power supply, mixers and detectors.

Super barrier diodes

Super barrier diodes are rectifier diodes that incorporate the low forward voltage drop of the Schottky diode with the surge-handling capability and low reverse leakage current of a normal p-n junction diode.

Gold-doped diodes

As a dopant, gold (or platinum) acts as recombination centers, which help a fast recombination of minority carriers. This allows the diode to operate at signal frequencies, at the expense of a higher forward voltage drop. Gold doped diodes are faster than other p-n diodes (but not as fast as Schottky diodes). They also have less reverse-current leakage than Schottky diodes (but not as good as other p-n diodes). A typical example is the 1N914.

Snap-off or Step recovery diodes

The term *step recovery* relates to the form of the reverse recovery characteristic of these devices. After a forward current has been passing in an SRD and the current is interrupted or reversed, the reverse conduction will cease very abruptly (as in a step waveform). SRDs can therefore provide very fast voltage transitions by the very sudden disappearance of the charge carriers.

Transient voltage suppression diode (TVS)

These are avalanche diodes designed specifically to protect other semiconductor devices from high-voltage transients. Their p-n junctions have a much larger cross-sectional area than those of a normal diode, allowing them to conduct large currents to ground without sustaining damage.

Varicap or varactor diodes

These are used as voltage-controlled capacitors. These are important in PLL (phase-locked loop) and FLL (frequency-locked loop) circuits, allowing tuning circuits, such as those in television receivers, to lock quickly, replacing older designs that took a long time to warm up and lock. A PLL is faster than an FLL, but prone to integer harmonic locking (if one attempts to lock to a broadband signal). They also enabled tunable oscillators in early discrete tuning of radios, where a cheap and stable, but fixed-frequency, crystal oscillator provided the reference frequency for a voltage-controlled oscillator.

Zener diodes

Diodes that can be made to conduct backwards. This effect, called Zener breakdown, occurs at a precisely defined voltage, allowing the diode to be used as a precision voltage reference. In practical voltage reference circuits Zener and switching diodes are connected in series and opposite directions to balance the temperature coefficient to near zero. Some devices labeled as high-voltage Zener diodes are actually avalanche diodes. Two (equivalent) Zeners in series and in reverse order, in the same package, constitute a transient absorber (or Transorb, a registered trademark). The Zener diode is named for Dr. Clarence Melvin Zener of Carnegie Mellon University, inventor of the device.

Other uses for semiconductor diodes include sensing temperature, and computing analog logarithms.

Diode | Zener diode | Schottky diode | Tunnel diode

Light-emitting diode | Photodiode | Varicap | Silicon controlled rectifier

Figure 6: Some diode symbols.

Numbering and coding schemes

There are a number of common, standard and manufacturer-driven numbering and coding schemes for diodes; the two most common being the EIA/JEDEC standard and the European Pro Electron standard:

EIA/JEDEC

A standardized 1N-series numbering system was introduced in the US by EIA/JEDEC (Joint Electron Device Engineering Council) about 1960. Among the most popular in this series were: 1N34A/1N270 (Germanium signal), 1N914/1N4148 (Silicon signal), 1N4001-1N4007 (Silicon 1A power rectifier) and 1N54xx (Silicon 3A power rectifier)

Pro Electron

The European Pro Electron coding system for active components was introduced in 1966 and comprises two letters followed by the part code. The first letter represents the semiconductor material used for the component (A = Germanium and B = Silicon) and the second letter represents the general function of the part (for diodes: A = low-power/signal, B = Variable capacitance, X = Multiplier, Y = Rectifier and Z = Voltage reference), for example:

- AA-series germanium low-power/signal diodes (e.g.: AA119)
- BA-series silicon low-power/signal diodes (e.g.: BAT18 Silicon RF Switching Diode)
- BY-series silicon rectifier diodes (e.g.: BY127 1250V, 1A rectifier diode)
- BZ-series silicon zener diodes (e.g.: BZY88C4V7 4.7V zener diode)

Other common numbering / coding systems (generally manufacturer-driven) include:

- GD-series germanium diodes (ed: GD9) — this is a very old coding system
- OA-series germanium diodes (e.g.: OA47) — a coding sequence developed by Mullard, a UK company

As well as these common codes, many manufacturers or organisations have their own systems too — for example:

- HP diode 1901-0044 = JEDEC 1N4148
- UK military diode CV448 = Mullard type OA81 = GEC type GEX23

Related devices

- Rectifier
- Transistor
- Thyristor or silicon controlled rectifier (SCR)
- TRIAC
- Diac
- Varistor

In optics, an equivalent device for the diode but with laser light would be the Optical isolator, also known as an Optical Diode, that allows light to only pass in one direction. It uses a Faraday rotator as the main component.

Applications

Radio demodulation

The first use for the diode was the demodulation of amplitude modulated (AM) radio broadcasts. The history of this discovery is treated in depth in the radio article. In summary, an AM signal consists of alternating positive and negative peaks of voltage, whose amplitude or "envelope" is proportional to the original audio signal. The diode (originally a crystal diode) rectifies the AM radio frequency signal, leaving an audio signal which is the original audio signal, minus atmospheric noise. The audio is extracted using a simple filter and fed into an audio amplifier or transducer, which generates sound waves.

Power conversion

Rectifiers are constructed from diodes, where they are used to convert alternating current (AC) electricity into direct current (DC). Automotive alternators are a common example, where the diode, which rectifies the AC into DC, provides better performance than the commutator of earlier dynamo. Similarly, diodes are also used in **Cockcroft–Walton voltage multipliers** to convert AC into higher DC voltages.

Over-voltage protection

Diodes are frequently used to conduct damaging high voltages away from sensitive electronic devices. They are usually reverse-biased (non-conducting) under normal circumstances. When the voltage rises above the normal range, the diodes become forward-biased (conducting). For example, diodes are used in (stepper motor and H-

bridge) motor controller and relay circuits to de-energize coils rapidly without the damaging voltage spikes that would otherwise occur. (Any diode used in such an application is called a flyback diode). Many integrated circuits also incorporate diodes on the connection pins to prevent external voltages from damaging their sensitive transistors. Specialized diodes are used to protect from over-voltages at higher power.

Logic gates

Diodes can be combined with other components to construct AND and OR logic gates. This is referred to as diode logic.

Ionizing radiation detectors

In addition to light, mentioned above, semiconductor diodes are sensitive to more energetic radiation. In electronics, cosmic rays and other sources of ionizing radiation cause noise pulses and single and multiple bit errors. This effect is sometimes exploited by particle detectors to detect radiation. A single particle of radiation, with thousands or millions of electron volts of energy, generates many charge carrier pairs, as its energy is deposited in the semiconductor material. If the depletion layer is large enough to catch the whole shower or to stop a heavy particle, a fairly accurate measurement of the particle's energy can be made, simply by measuring the charge conducted and without the complexity of a magnetic spectrometer or etc. These semiconductor radiation detectors need efficient and uniform charge collection and low leakage current. They are often cooled by liquid nitrogen. For longer range (about a centimetre) particles they need a very large depletion depth and large area. For short range particles, they need any contact or un-depleted semiconductor on at least one surface to be very thin. The back-bias voltages are near breakdown (around a thousand volts per centimetre). Germanium and silicon are common materials. Some of these detectors sense position as well as energy. They have a finite life, especially when detecting heavy particles, because of radiation damage. Silicon and germanium are quite different in their ability to convert gamma rays to electron showers.

Semiconductor detectors for high energy particles are used in large numbers. Because of energy loss fluctuations, accurate measurement of the energy deposited is of less use.

Temperature measurements

A diode can be used as a temperature measuring device, since the forward voltage drop across the diode depends on temperature, as in a Silicon bandgap temperature sensor. From the Shockley ideal diode equation given above, it appears the voltage has a positive temperature coefficient (at a constant current) but depends on doping concentration and operating temperature (Sze 2007). The temperature coefficient can be negative as in typical thermistors or positive for temperature sense diodes down to about 20 kelvins. Typically, silicon diodes have approximately −2 mV/°C temperature coefficient at room temperature.

Current steering

Diodes will prevent currents in unintended directions. To supply power to an electrical circuit during a power failure, the circuit can draw current from a battery. An Uninterruptible power supply may use diodes in this way to ensure that current is only drawn from the battery when necessary. Similarly, small boats typically have two circuits each with their own battery/batteries: one used for engine starting; one used for domestics. Normally both are charged from a single alternator, and a heavy duty split charge diode is used to prevent the higher charge battery (typically the engine battery) from discharging through the lower charged battery when the alternator is not running.

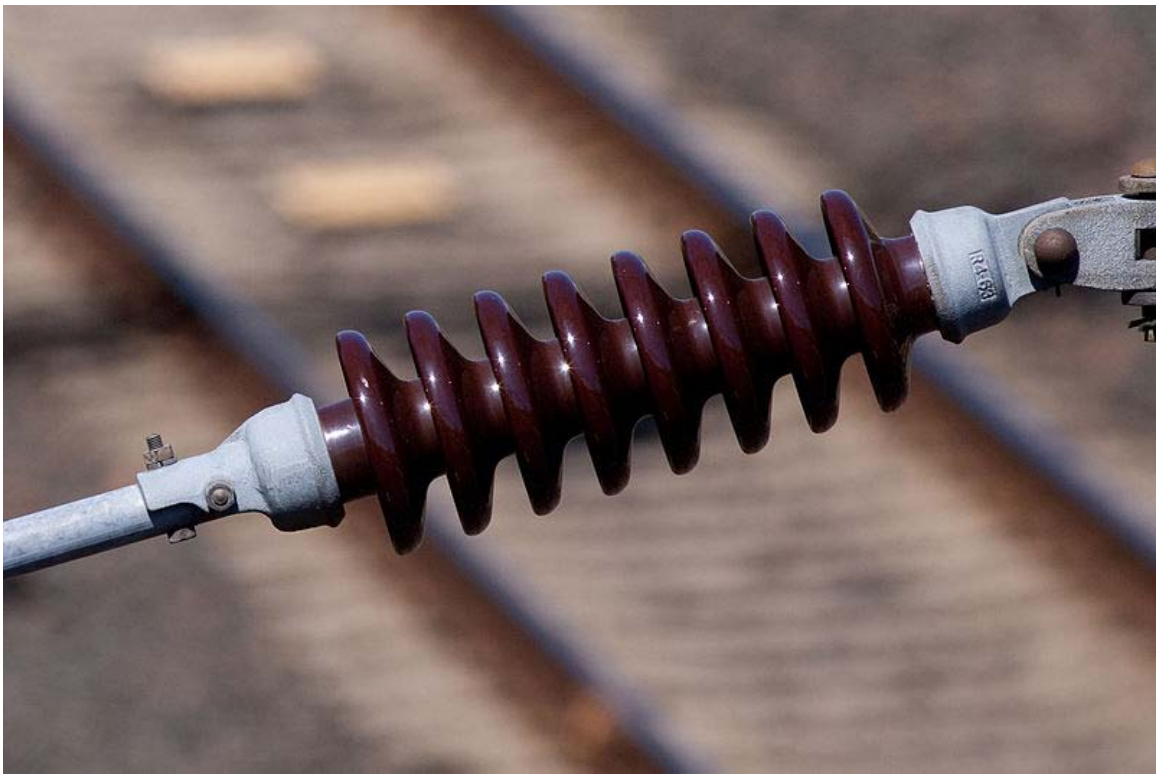
Diodes are also used in electronic musical keyboards. To reduce the amount of wiring needed in electronic musical keyboards, these instruments often use keyboard matrix circuits. The keyboard controller scans the rows and columns to determine which note the player has pressed. The problem with matrix circuits is that when several notes are pressed at once, the current can flow backwards through the circuit and trigger "phantom keys" that cause “ghost” notes to play. To avoid triggering unwanted notes, most keyboard matrix circuits have diodes soldered with the switch under each key of the musical keyboard. The same principle is also used for the switch matrix in solid state pinball machines.

Abbreviations

Diodes are usually referred to as *D* for diode on PCBs. Sometimes the abbreviation *CR* for **crystal rectifier** is used.

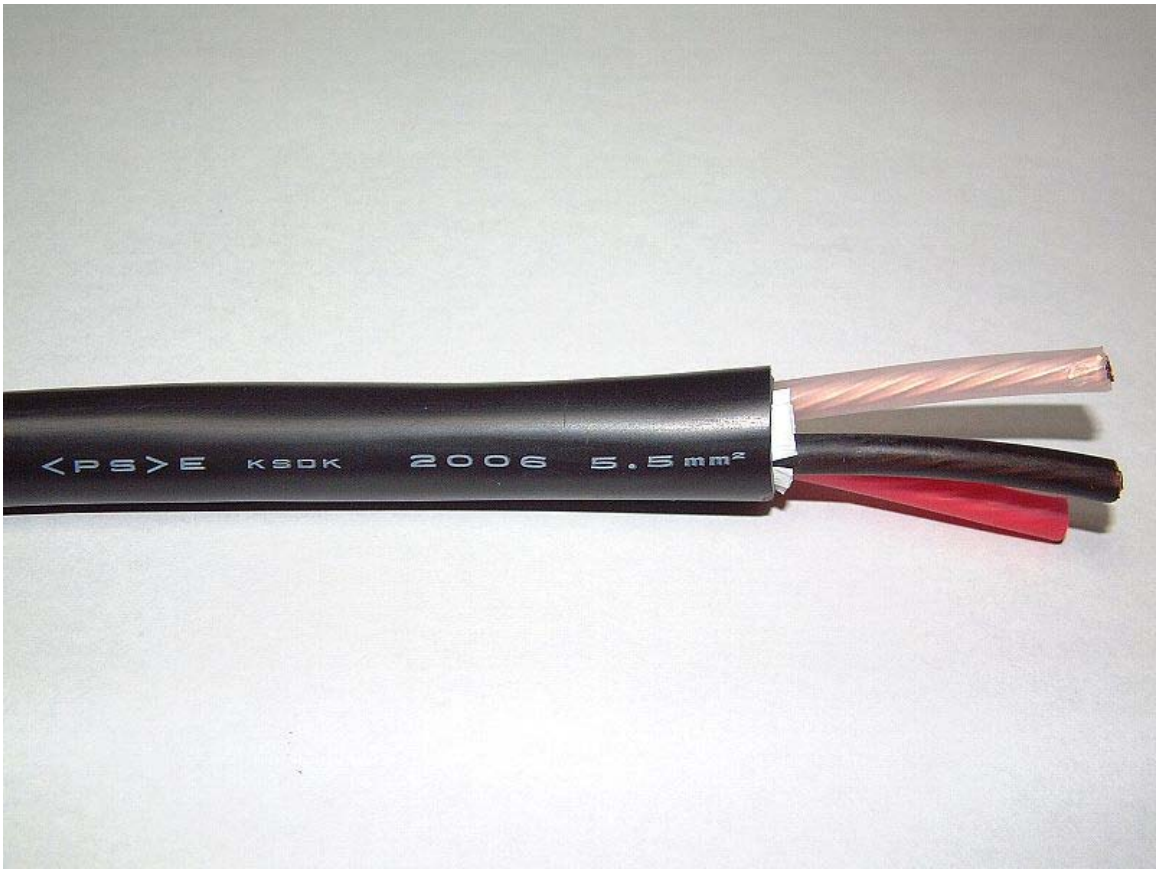
Chapter-8

Insulator

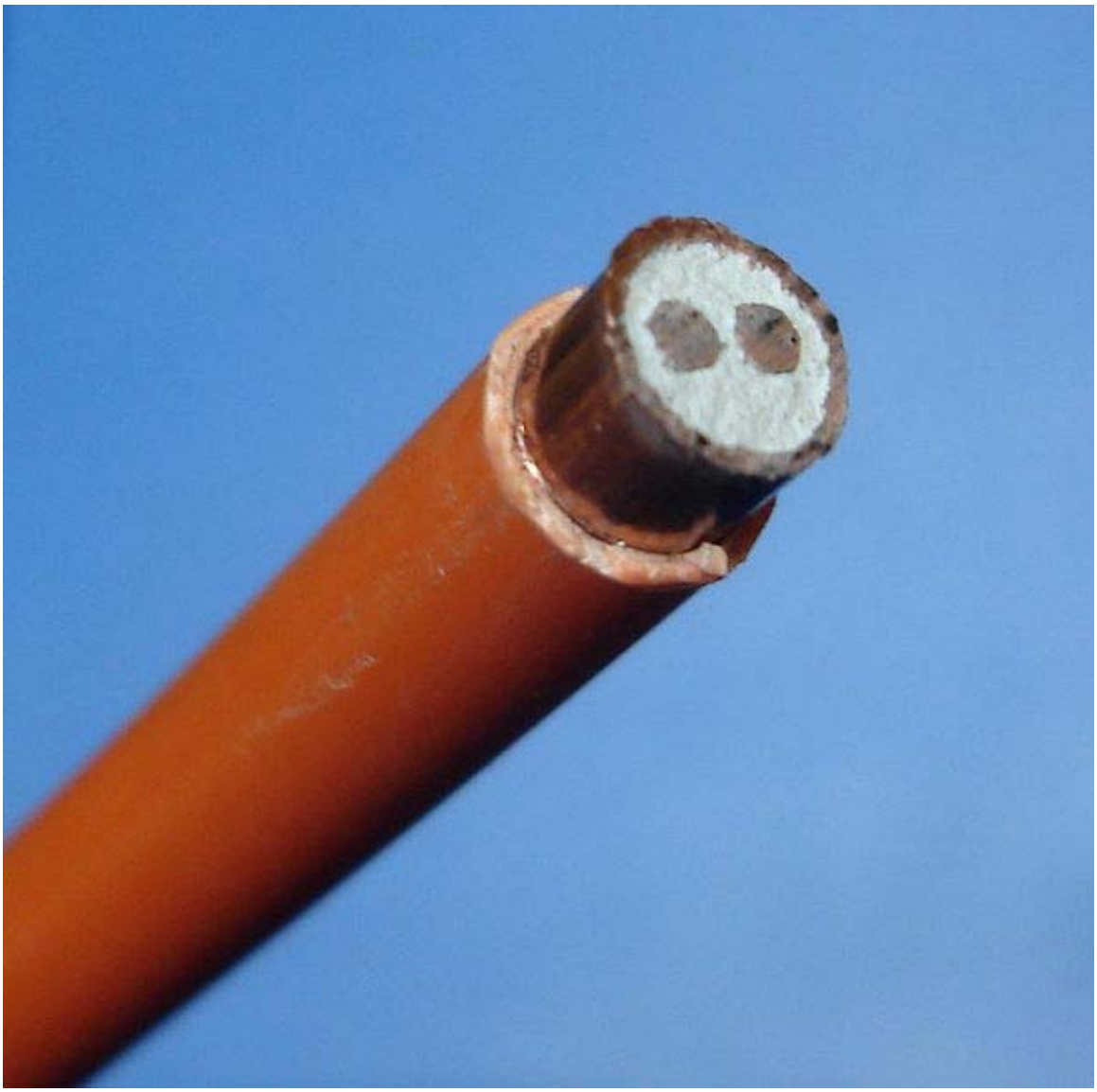


Ceramic insulator at railways

Conducting copper wire insulated by an outer layer of polyethylene



3-core copper wire power cable, each core with individual colour coded insulating sheaths all contained within an outer protective sheath



PVC-sheathed Mineral insulated copper cable with 2 conducting cores

Coaxial Cable with dielectric insulator supporting a central core

An **insulator**, also called a *dielectric*, is a material that resists the flow of electric charge. In insulating materials valence electrons are tightly bonded to their atoms. These materials are used in electrical equipment as *insulators* or *insulation*. Their function is to support or separate electrical conductors without allowing current through themselves. The term also refers to insulating supports that attach electric power transmission wires to utility poles or pylons.

Some materials such as glass, paper or Teflon are very good electrical insulators. Even though they may have lower bulk resistivity, a much larger class of materials are still "good enough" to insulate electrical wiring and cables. Examples include rubber-like polymers and most plastics. Such materials can serve as practical and safe insulators for low to moderate voltages (hundreds, or even thousands, of volts).

Physics of conduction in solids

Electrical insulation is the absence of electrical conduction. Electronic band theory (a branch of physics) says that a charge will flow if states are available into which electrons can be excited. This allows electrons to gain energy and thereby move through a conductor such as a metal. If no such states are available, the material is an insulator.

Most insulators have a large band gap. This occurs because the "valence" band containing the highest energy electrons is full, and a large energy gap separates this band from the next band above it. There is always some voltage (called the breakdown voltage) that will give the electrons enough energy to be excited into this band. Once this voltage is exceeded, the material ceases being an insulator, and charge will begin to pass through it. However, it is usually accompanied by physical or chemical changes that permanently degrade the material's insulating properties.

Materials that lack electron conduction are insulators if they lack other mobile charges as well. For example, if a liquid or gas contains ions, then the ions can be made to flow as an electric current, and the material is a conductor. Electrolytes and plasmas contain ions and will act as conductors whether or not electron flow is involved.

Breakdown

Insulators suffer from the phenomenon of electrical breakdown. When the electric field applied across an insulating substance exceeds in any location the threshold breakdown field for that substance, which is proportional to the band gap energy, the insulator suddenly turns into a resistor, sometimes with catastrophic results. During electrical breakdown, any free charge carrier being accelerated by the strong e-field will have enough velocity to knock electrons from (ionize) any atom it strikes. These freed electrons and ions are in turn accelerated and strike other atoms, creating more charge carriers, in a chain reaction. Rapidly the insulator becomes filled with mobile carriers, and its resistance drops to a low level. In air, "corona discharge" is normal current near a high-voltage conductor; an "arc" is an unusual and undesired current. Similar breakdown can occur within any insulator, even within the bulk solid of a material. Even a vacuum can suffer a sort of breakdown, but in this case the breakdown or vacuum arc involves charges ejected from the surface of metal electrodes rather than produced by the vacuum itself.

Uses

Insulators are commonly used as a flexible coating on electric wire and cable. Since air is an insulator, no other substance is needed to keep power where it should be. High-voltage power lines commonly use just air, since a solid (e.g., plastic) coating would be impractical. However, wires which touch each other will produce cross connections, short circuits, and fire hazards. In coaxial cable the center conductor must be supported exactly in the middle of the hollow shield in order to prevent EM wave reflections. And any wires which present voltages higher than 60V can cause human shock and electrocution hazards. Insulating coatings helps to prevent all of these problems.

Some wires have a mechanical covering which has no voltage rating; e.g: service-drop, welding, doorbell, thermostat.

An insulated wire or cable has a voltage rating and a maximum conductor temperature rating. It does not have an ampacity rating, since such is dependent upon the wire or cables environment where installed.

In electronic systems, printed circuit boards are made from epoxy plastic and fibreglass. The nonconductive boards support layers of copper foil conductors. In electronic devices, the tiny and delicate active components are embedded within nonconductive epoxy or phenolic plastics, or within baked glass or ceramic coatings.

In microelectronic components such as transistors and ICs, the silicon material is normally a conductor because of doping, but it can easily be selectively transformed into a good insulator by the application of heat and oxygen. Oxidized silicon is quartz, i.e. silicon dioxide.

In high voltage systems containing transformers and capacitors, liquid insulator oil is the typical method used for preventing arcs. The oil replaces the air in any spaces which must support significant voltage without electrical breakdown. Other methods of insulating high voltage systems are ceramic or glass wire holders, gas, vacuum, and simply placing the wires with a large separation, using the air as insulation.

Telegraph and power transmission insulators

Suspended glass disk insulator unit used in *cap and pin* insulator strings for high voltage transmission lines

Suspended wires for electric power transmission are bare, except where they enter buildings, and are insulated by the surrounding air. Insulators are required at the points at which they are supported by utility poles or pylons. Insulators are also required where the wire enters buildings or electrical devices, such as transformers or circuit breakers, to insulate the wire from the case. These hollow insulators with a conductor inside them are called bushings.

10 kV ceramic insulator, showing sheds

Material

Insulators used for high-voltage power transmission are made from glass, porcelain, or composite polymer materials. Porcelain insulators are made from clay, quartz or alumina

and feldspar, and are covered with a smooth glaze to shed water. Insulators made from porcelain rich in alumina are used where high mechanical strength is a criterion. Porcelain has a dielectric strength of about 4–10 kV/mm. Glass has a higher dielectric strength, but it attracts condensation and the thick irregular shapes needed for insulators are difficult to cast without internal strains. Some insulator manufacturers stopped making glass insulators in the late 1960s, switching to ceramic materials.

Recently, some electric utilities have begun converting to polymer composite materials for some types of insulators. These are typically composed of a central rod made of fibre reinforced plastic and an outer weathershed made of silicone rubber or EPDM. Composite insulators are less costly, lighter in weight, and have excellent hydrophobic capability. This combination makes them ideal for service in polluted areas. However, these materials do not yet have the long-term proven service life of glass and porcelain.

Design

High voltage ceramic bushing during manufacture, before glazing.

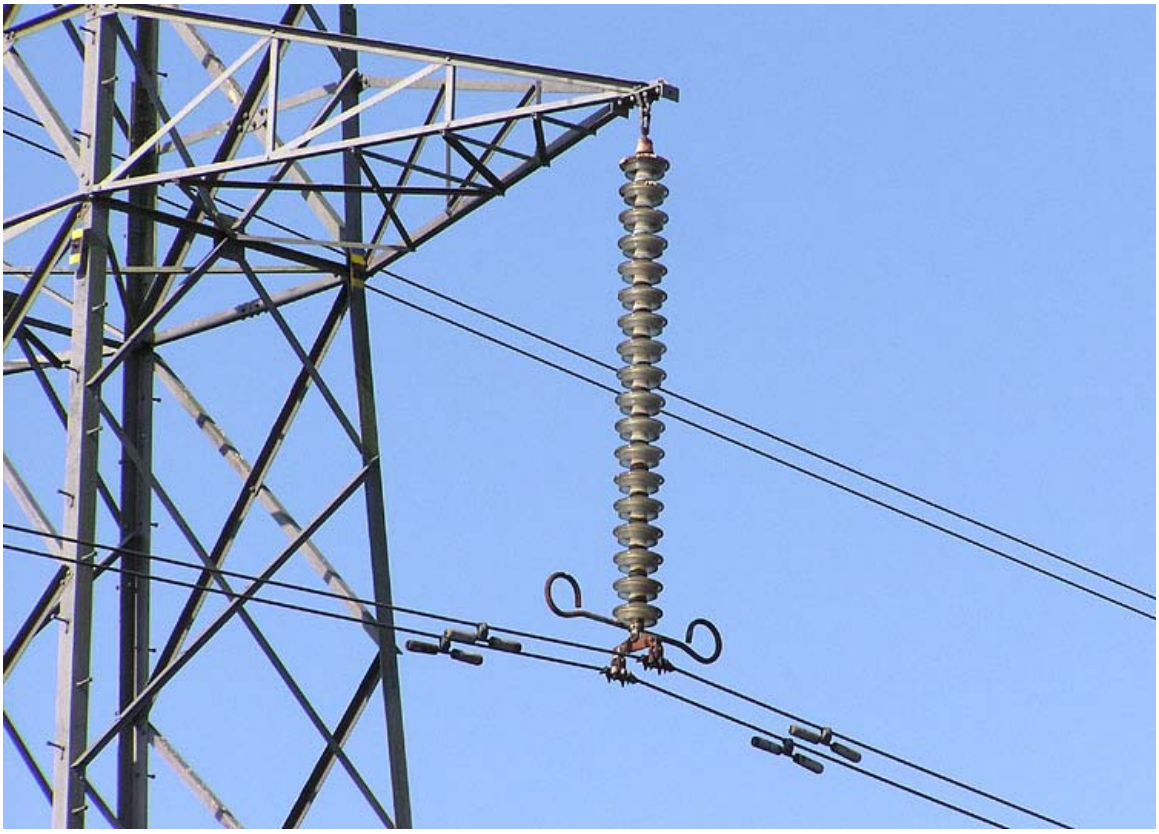
The electrical breakdown of an insulator due to excessive voltage can occur in one of two ways:

- *Puncture voltage* is the voltage across the insulator (when installed in its normal manner) which causes a breakdown and conduction through the interior of the insulator. The heat resulting from the puncture arc usually damages the insulator irreparably.
- *Flashover voltage* is the voltage which causes the air around or along the surface of the insulator to break down and conduct, causing a 'flashover' arc along the

outside of the insulator. They are usually designed to withstand this without damage.

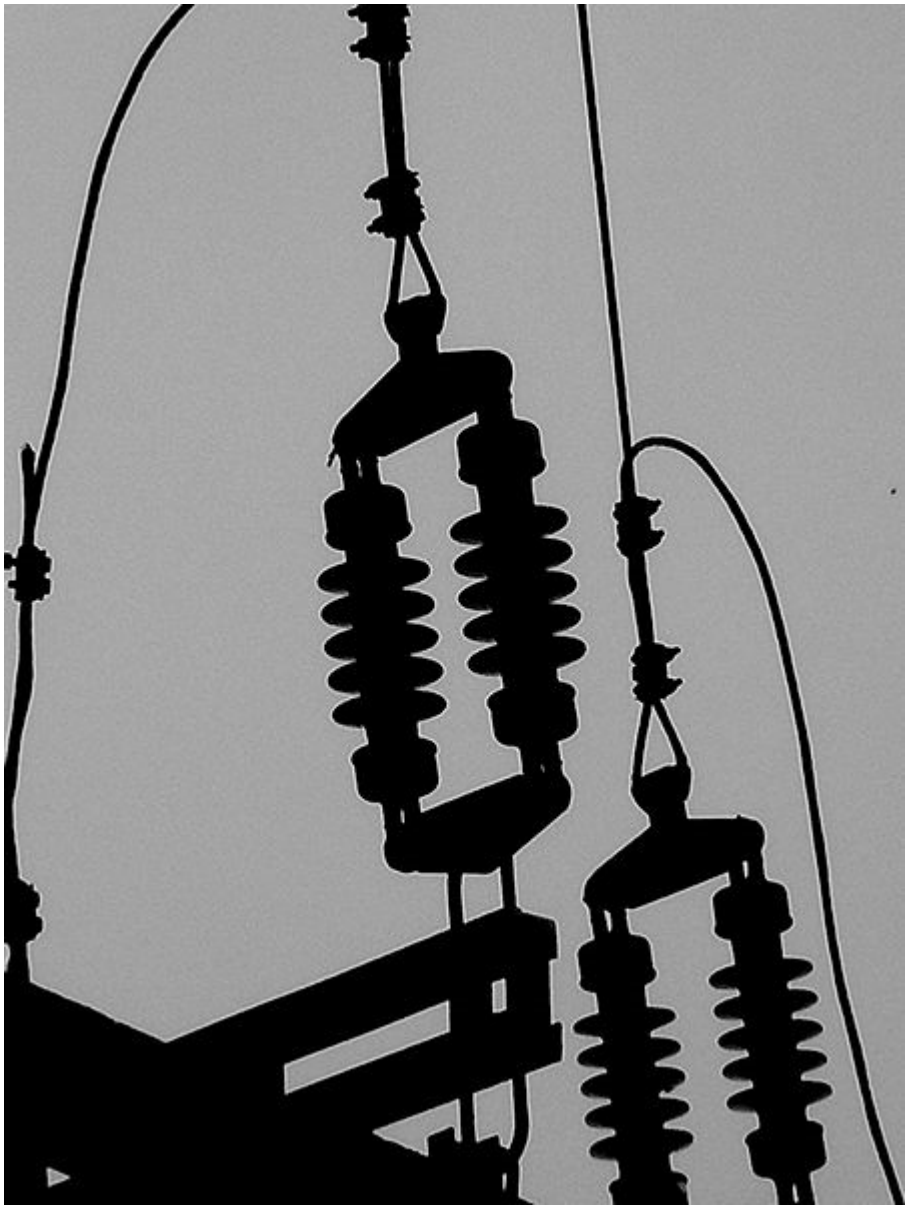
Most high voltage insulators are designed with a lower flashover voltage than puncture voltage, so they will flashover before they puncture, to avoid damage.

Dirt, pollution, salt, and particularly water on the surface of a high voltage insulator can create a conductive path across it, causing leakage currents and flashovers. The flashover voltage can be more than 50% lower when the insulator is wet. High voltage insulators for outdoor use are shaped to maximize the length of the leakage path along the surface from one end to the other, called the creepage length, to minimize these leakage currents. To accomplish this the surface is molded into a series of corrugations or concentric disk shapes. These usually include one or more *sheds*; downward facing cup-shaped surfaces that act as umbrellas to ensure that the part of the surface leakage path under the 'cup' stays dry in wet weather. Minimum creepage distances are 20–25 mm/kV, but must be increased in high pollution or airborne sea-salt areas.



Cap and pin insulator string (the vertical string of discs) on a 275 kV suspension pylon.

A recent photo of an open wire telegraph pole route with traditional porcelain insulators. Quidenham, Norfolk, United Kingdom.



Ceramic Insulators on a power line in Poland

Cap and pin insulators

Higher voltage transmission lines use modular *cap and pin* insulator designs. The wires are suspended from a 'string' of identical disk-shaped insulators which attach to each other with metal clevis pin or ball and socket links. The advantage of this design is that insulator strings with different breakdown voltages, for use with different line voltages, can be constructed by using different numbers of the basic units. Also, if one of the insulator units in the string breaks, it can be replaced without discarding the entire string.

Each unit is constructed of a ceramic or glass disk with a metal cap and pin cemented to opposite sides. In order to make defective units obvious, glass units are designed with

Class B construction, so that an overvoltage causes a puncture arc through the glass instead of a flashover. The glass is heat-treated so it will shatter, making the damaged unit visible. However the mechanical strength of the unit is unchanged, so the insulator string will stay together.

Standard disk insulator units are 10 inches (25 cm) in diameter and $5\frac{3}{4}$ in (15 cm) long, can support a load of 80-120 kN (18-27 klbf), have a dry flashover voltage of about 72 kV, and are rated at an operating voltage of 10-12 kV. However, the flashover voltage of a string is less than the sum of its component disks, because the electric field is not distributed evenly across the string but is strongest at the disk nearest to the conductor, which will flashover first. Metal *grading rings* are sometimes added around the lowest disk, to reduce the electric field across that disk and improve flashover voltage.

Typical number of disk insulator units for standard line voltages

Line voltage (kV)	Disks
34.5	3
46	4
69	5
92	7
115	8
138	9
161	11
196	13
230	15
287	19
345	22
360	23

History

The first electrical systems to make use of insulators were telegraph lines; direct attachment of wires to wooden poles was found to give very poor results, especially during damp weather.

The first glass insulators used in large quantities had an unthreaded pinhole. These pieces of glass were positioned on a tapered wooden pin, vertically extending upwards from the pole's crossarm (commonly only two insulators to a pole and maybe one on top of the pole itself). Natural contraction and expansion of the wires tied to these "threadless insulators" resulted in insulators unseating from their pins, requiring manual reseating.

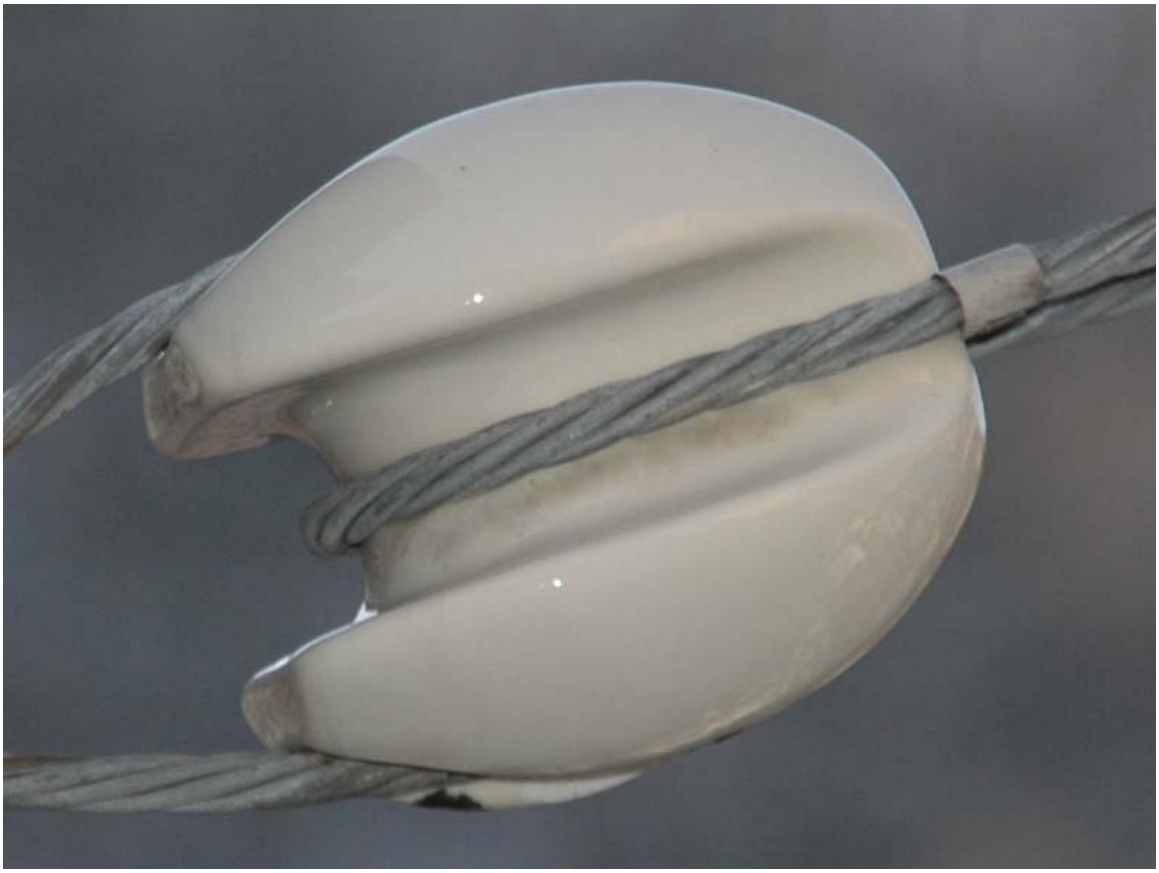
Amongst the first to produce ceramic insulators were companies in the United Kingdom, with Stiff and Doulton using stoneware from the mid 1840s, Joseph Bourne (later

renamed Denby) producing them from around 1860 and Bullers from 1868. Utility patent number 48,906 was granted to Louis A. Cauvet on July 25, 1865 for a process to produce insulators with a threaded pinhole. To this day, pin-type insulators still have threaded pinholes.

The invention of suspension-type insulators made high-voltage power transmission possible. Pin-type insulators were unsatisfactory over about 60,000 volts.

A large variety of telephone, telegraph and power insulators have been made; some people collect them, both for their historic interest and for the aesthetic quality of many insulator designs and finishes.

Insulation of antennas



Egg shaped strain insulator

Often a broadcasting radio antenna is built as a mast radiator, which means that the entire mast structure is energized with high voltage and must be insulated from the ground. Steatite mountings are used. They have to withstand not only the voltage of the mast radiator to ground, which can reach values up to 400 kV at some antennas, but also the weight of the mast construction and dynamic forces. Arcing horns and lightning arresters are necessary because lightning strikes to the mast are common.

Guy wires supporting antenna masts usually have strain insulators inserted in the cable run, to keep the high voltages on the antenna from short circuiting to ground or creating a shock hazard. Often guy cables have several insulators, placed to break up the cable into lengths that are not submultiples of the transmitting wavelength to avoid unwanted electrical resonances in the guy. These insulators are usually ceramic and cylindrical or egg-shaped. This construction has the advantage that the ceramic is under compression rather than tension, so it can withstand greater load, and that if the insulator breaks the cable ends will still be linked.

These insulators also have to be equipped with overvoltage protection equipment. For the dimensions of the guy insulation, static charges on guys have to be considered, at high masts these can be much higher than the voltage caused by the transmitter requiring guys divided by insulators in multiple sections on the highest masts. In this case, guys which are grounded at the anchor basements via a coil - or if possible, directly - are the better choice.

Feedlines attaching antennas to radio equipment, particularly twin lead type, often must be kept at a distance from metal structures. The insulated supports used for this purpose are called *standoff insulators*.

Insulation in electrical apparatus

The most important insulation material is air. A variety of solid, liquid, and gaseous insulators are also used in electrical apparatus. In smaller transformers, generators, and electric motors, insulation on the wire coils consists of up to four thin layers of polymer varnish film. Film insulated **magnet wire** permits a manufacturer to obtain the maximum number of turns within the available space. Windings that use thicker conductors are often wrapped with supplemental fiberglass insulating tape. Windings may also be impregnated with insulating varnishes to prevent electrical corona and reduce magnetically induced wire vibration. Large power transformer windings are still mostly insulated with paper, wood, varnish, and mineral oil; although these materials have been used for more than 100 years, they still provide a good balance of economy and adequate performance. Busbars and circuit breakers in switchgear may be insulated with glass-reinforced plastic insulation, treated to have low flame spread and to prevent tracking of current across the material.

In older apparatus made up to the early 1970s, boards made of compressed asbestos may be found; while this is an adequate insulator at power frequencies, handling or repairs to asbestos material will release dangerous fibers into the air and must be carried out with caution. Wire insulated with felted asbestos was used in high-temperature and rugged applications from the 1920s. Wire of this type was sold by General Electric under the trade name "Deltabeston"

Live-front switchboards up to the early part of the 20th century were made of slate or marble.

Some high voltage equipment is designed to operate within a high pressure insulating gas such as sulfur hexafluoride.

Insulation materials that perform well at power and low frequencies may be unsatisfactory at radio frequency, due to heating from excessive dielectric dissipation.

Electrical wires may be insulated with polyethylene, crosslinked polyethylene (either through electron beam processing or chemical crosslinking), PVC, Kapton, rubber-like polymers, oil impregnated paper, Teflon, silicone, or modified ethylene tetrafluoroethylene (ETFE). Larger power cables may use compressed inorganic powder, depending on the application.

Flexible insulating materials such as PVC (polyvinyl chloride) are used to insulate the circuit and prevent human contact with a 'live' wire – one having voltage of 600 volts or less. Alternative materials are likely to become increasingly used due to EU safety and environmental legislation making PVC less economic.

Class 1 and Class 2 insulation

All portable or hand-held electrical devices are insulated to protect their user from harmful shock.

Class 1 insulation requires that the metal body and other exposed metal parts of the device is connected to earth via a "grounding" wire which is earthed at the main service panel; but only basic insulation of the conductors is needed. This equipment is easily identified by a third pin on the power plug for the grounding connection.

Class 2 insulation means that the device is *double insulated*. This is used on some appliances such as electric shavers, hair dryers and portable power tools. Double insulation requires that the devices have both basic and supplementary insulation, each of which is sufficient to prevent electric shock. All internal electrically energized components are totally enclosed within an insulated body that prevents any contact with "live" parts. In the EU, double insulated appliances all are marked with a symbol of two squares, one inside the other.