Handbook of Electrical and Electromagnetic Components

Meri Marlow

Guillermo Dutton

First Edition, 2012

ISBN 978-81-323-0904-8

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

Table of Contents

Chapter- 1

Operational Amplifier

A Signetics μa741 operational amplifier, one of the most successful op-amps.

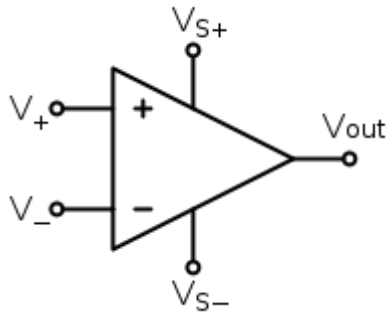
An **operational amplifier** ("op-amp") is a DC-coupled high-gain electronic voltage amplifier with a differential input and, usually, a single-ended output. An op-amp produces an output voltage that is typically hundreds of thousands times larger than the voltage *difference* between its input terminals.

Operational amplifiers are important building blocks for a wide range of electronic circuits. They had their origins in analog computers where they were used in many linear, non-linear and frequency-dependent circuits. Their popularity in circuit design largely stems from the fact the characteristics of the final elements (such as their gain) are set by external components with little dependence on temperature changes and manufacturing variations in the op-amp itself.

Op-amps are among the most widely used electronic devices today, being used in a vast array of consumer, industrial, and scientific devices. Many standard IC op-amps cost only a few cents in moderate production volume; however some integrated or hybrid operational amplifiers with special performance specifications may cost over $100 US in small quantities. Op-amps may be packaged as components, or used as elements of more complex integrated circuits.

The op-amp is one type of differential amplifier. Other types of differential amplifier include the fully differential amplifier (similar to the op-amp, but with two outputs), the instrumentation amplifier (usually built from three op-amps), the isolation amplifier (similar to the instrumentation amplifier, but with tolerance to common-mode voltages that would destroy an ordinary op-amp), and negative feedback amplifier (usually built from one or more op-amps and a resistive feedback network).

Circuit notation



Circuit diagram symbol for an op-amp

The circuit symbol for an op-amp is shown to the right, where:

- V_+: non-inverting input
- V_-: inverting input
- V_{out}: output
- $V_{\text{S}+}$: positive power supply
- $V_{\text{S}-}$: negative power supply

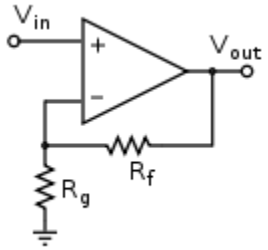
The power supply pins ($V_{\text{S}+}$ and $V_{\text{S}-}$) can be labeled in different ways. Despite different labeling, the function remains the same — to provide additional power for amplification of the signal. Often these pins are left out of the diagram for clarity, and the power configuration is described or assumed from the circuit.

Operation

The amplifier's differential inputs consist of a V_+ input and a V_- input, and ideally the op-amp amplifies only the difference in voltage between the two, which is called the *differential input voltage*. The output voltage of the op-amp is given by the equation,

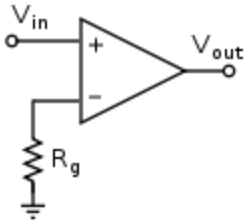
$$V_{\text{out}} = (V_+ - V_-)\, A_{OL}$$

where V_+ is the voltage at the non-inverting terminal, V_- is the voltage at the inverting terminal and A_{OL} is the open-loop gain of the amplifier. (The term "open-loop" refers to the absence of a feedback loop from the output to the input.)



Typically the op-amp's very large gain is controlled by negative feedback, which largely determines the magnitude of its output ("closed-loop") voltage gain in amplifier applications, or the transfer function required (in analog computers). Without negative feedback, and perhaps with positive feedback for regeneration, an op-amp acts as a comparator. High input impedance at the input terminals and low output impedance at the output terminal(s) are important typical characteristics.

With no negative feedback, the op-amp acts as a comparator. The inverting input is held at ground (0 V) by the resistor, so if the V_{in} applied to the non-inverting input is positive, the output will be maximum positive, and if V_{in} is negative, the output will be maximum negative. Since there is no feedback from the output to either input, this is an *open loop* circuit. The circuit's gain is just the G_{OL} of the op-amp.



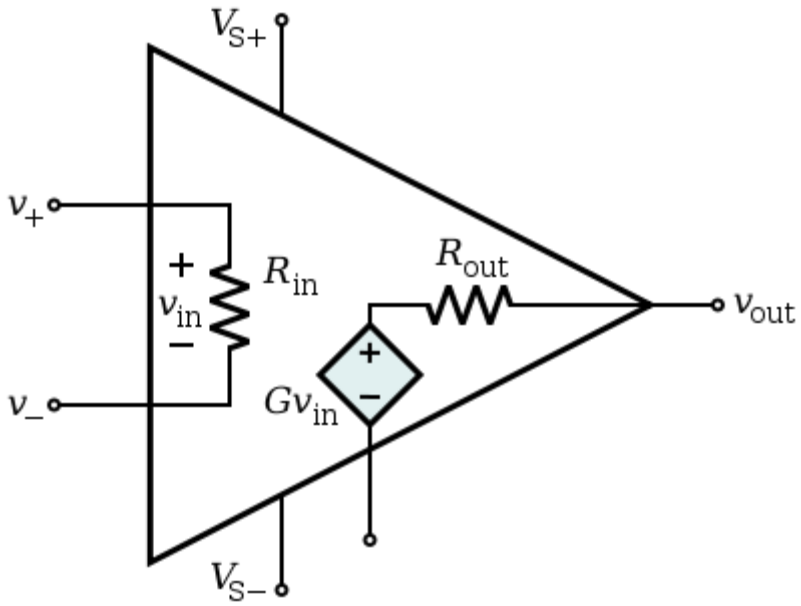
Adding negative feedback via the voltage divider R_f,R_g reduces the gain. Equilibrium will be established when V_{out} is just sufficient to reach around and "pull" the inverting input to the same voltage as V_{in}. As a simple example, if V_{in} = 1 V and $R_f = R_g$, V_{out} will be 2 V, the amount required to keep V_- at 1 V. Because of the feedback provided by R_f,R_g this is a *closed loop* circuit. Its over-all gain V_{out} / V_{in} is called the *closed-loop gain* A_{CL}. Because the feedback is negative, in this case A_{CL} is less than the A_{OL} of the op-amp.

The magnitude of A_{OL} is typically very large—10,000 or more for integrated circuit op-amps—and therefore even a quite small difference between V_+ and V_- drives the amplifier output nearly to the supply voltage. This is called *saturation* of the amplifier. The magnitude of A_{OL} is not well controlled by the manufacturing process, and so it is impractical to use an operational amplifier as a stand-alone differential amplifier. If predictable operation is desired, negative feedback is used, by applying a portion of the output voltage to the inverting input. The *closed loop* feedback greatly reduces the gain of the amplifier. If negative feedback is used, the circuit's overall gain and other parameters become determined more by the feedback network than by the op-amp itself. If the feedback network is made of components with relatively constant, stable values, the unpredictability and inconstancy of the op-amp's parameters do not seriously affect the circuit's performance.

If no negative feedback is used, the op-amp functions as a switch or comparator.

Positive feedback may be used to introduce hysteresis or oscillation.

Ideal and real op-amps



An equivalent circuit of an operational amplifier that models some resistive non-ideal parameters.

An ideal op-amp is usually considered to have the following properties, and they are considered to hold for all input voltages:

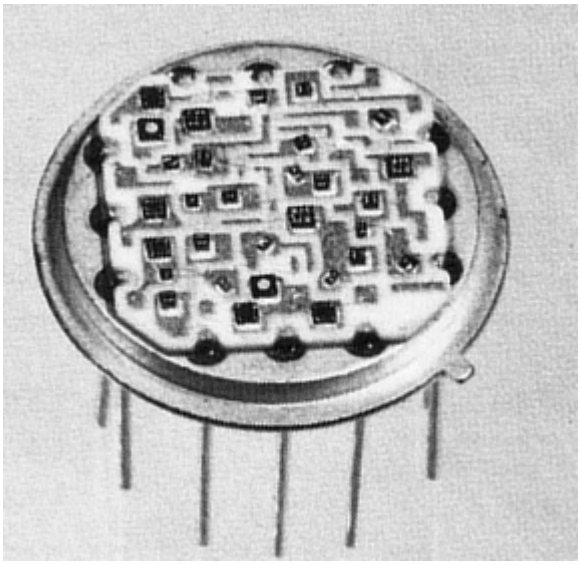
- Infinite open-loop gain (when doing theoretical analysis, a limit may be taken as open loop gain A_{OL} goes to infinity).

- Infinite voltage range available at the output (v_{out}) (in practice the voltages available from the output are limited by the supply voltages V_{S+} and V_{S-}). The power supply sources are called rails.
- Infinite bandwidth (i.e., the frequency magnitude response is considered to be flat everywhere with zero phase shift).
- Infinite input impedance (so, in the diagram, $R_{in} = \infty$, and zero current flows from v_+ to v_-).
- Zero input current (i.e., there is assumed to be no leakage or bias current into the device).
- Zero input offset voltage (i.e., when the input terminals are shorted so that $v_+ = v_-$, the output is a virtual ground or $v_{out} = 0$).
- Infinite slew rate (i.e., the rate of change of the output voltage is unbounded) and power bandwidth (full output voltage and current available at all frequencies).
- Zero output impedance (i.e., $R_{out} = 0$, so that output voltage does not vary with output current).
- Zero noise.
- Infinite Common-mode rejection ratio (CMRR).
- Infinite Power supply rejection ratio for both power supply rails.

In practice, none of these ideals can be realized, and various shortcomings and compromises have to be accepted. Depending on the parameters of interest, a real op-amp may be modeled to take account of some of the non-infinite or non-zero parameters using equivalent resistors and capacitors in the op-amp model. The designer can then include the effects of these undesirable, but real, effects into the overall performance of the final circuit. Some parameters may turn out to have negligible effect on the final design while others represent actual limitations of the final performance, that must be evaluated.

History

GAP/R's K2-W: a vacuum-tube op-amp (1953)



ADI's HOS-050: a high speed hybrid IC op-amp (1979)

An op-amp in a modern DIP

1941: First (vacuum tube) op-amp

An op-amp, defined as a general-purpose, DC-coupled, high gain, inverting feedback amplifier, is first found in U.S. Patent 2,401,779 "Summing Amplifier" filed by Karl D. Swartzel Jr. of Bell labs in 1941. This design used three vacuum tubes to achieve a gain of 90 dB and operated on voltage rails of ±350 V. It had a single inverting input rather than differential inverting and non-inverting inputs, as are common in today's op-amps. Throughout World War II, Swartzel's design proved its value by being liberally used in the M9 artillery director designed at Bell Labs. This artillery director worked with the SCR584 radar system to achieve extraordinary hit rates (near 90%) that would not have been possible otherwise.

1947: First op-amp with an explicit non-inverting input

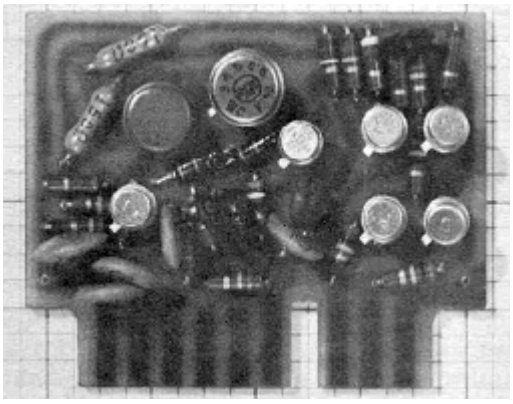
In 1947, the operational amplifier was first formally defined and named in a paper by Professor John R. Ragazzini of Columbia University. In this same paper a footnote mentioned an op-amp design by a student that would turn out to be quite significant. This op-amp, designed by Loebe Julie, was superior in a variety of ways. It had two major innovations. Its input stage used a long-tailed triode pair with loads matched to reduce drift in the output and, far more importantly, it was the first op-amp design to have two inputs (one inverting, the other non-inverting). The differential input made a whole range of new functionality possible, but it would not be used for a long time due to the rise of the chopper-stabilized amplifier.

1949: First chopper-stabilized op-amp

In 1949, Edwin A. Goldberg designed a chopper-stabilized op-amp. This set-up uses a normal op-amp with an additional AC amplifier that goes alongside the op-amp. The chopper gets an AC signal from DC by switching between the DC voltage and ground at a fast rate (60 Hz or 400 Hz). This signal is then amplified, rectified, filtered and fed into the op-amp's non-inverting input. This vastly improved the gain of the op-amp while significantly reducing the output drift and DC offset. Unfortunately, any design that used a chopper couldn't use their non-inverting input for any other purpose. Nevertheless, the much improved characteristics of the chopper-stabilized op-amp made it the dominant way to use op-amps. Techniques that used the non-inverting input regularly would not be very popular until the 1960s when op-amp ICs started to show up in the field.

In 1953, vacuum tube op-amps became commercially available with the release of the model K2-W from George A. Philbrick Researches, Incorporated. The designation on the devices shown, GAP/R, is a contraction for the complete company name. Two nine-pin 12AX7 vacuum tubes were mounted in an octal package and had a model K2-P chopper add-on available that would effectively "use up" the non-inverting input. This op-amp was based on a descendant of Loebe Julie's 1947 design and, along with its successors, would start the widespread use of op-amps in industry.

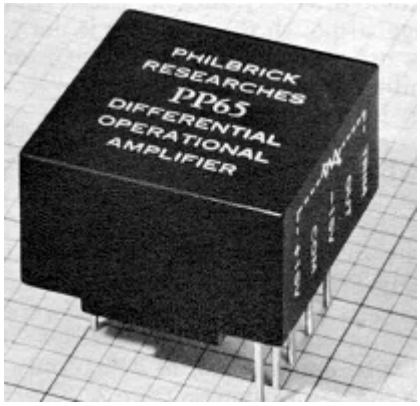
1961: First discrete IC op-amps



GAP/R's model P45: a solid-state, discrete op-amp (1961).

With the birth of the transistor in 1947, and the silicon transistor in 1954, the concept of ICs became a reality. The introduction of the planar process in 1959 made transistors and ICs stable enough to be commercially useful. By 1961, solid-state, discrete op-amps were being produced. These op-amps were effectively small circuit boards with packages such as edge-connectors. They usually had hand-selected resistors in order to improve things such as voltage offset and drift. The P45 (1961) had a gain of 94 dB and ran on ±15 V rails. It was intended to deal with signals in the range of ±10 V.

1962: First op-amps in potted modules



GAP/R's model PP65: a solid-state op-amp in a potted module (1962)

By 1962, several companies were producing modular potted packages that could be plugged into printed circuit boards. These packages were crucially important as they made the operational amplifier into a single black box which could be easily treated as a component in a larger circuit.

1963: First monolithic IC op-amp

In 1963, the first monolithic IC op-amp, the μA702 designed by Bob Widlar at Fairchild Semiconductor, was released. Monolithic ICs consist of a single chip as opposed to a chip and discrete parts (a discrete IC) or multiple chips bonded and connected on a circuit board (a hybrid IC). Almost all modern op-amps are monolithic ICs; however, this first IC did not meet with much success. Issues such as an uneven supply voltage, low gain and a small dynamic range held off the dominance of monolithic op-amps until 1965 when the μA709 (also designed by Bob Widlar) was released.

1966: First varactor bridge op-amps

Since the 741, there have been many different directions taken in op-amp design. Varactor bridge op-amps started to be produced in the late 1960s. They were designed to have extremely small input current and are still amongst the best op-amps available in terms of common-mode rejection with the ability to correctly deal with hundreds of volts at their inputs.

1968: Release of the μA741

The popularity of monolithic op-amps was further improved upon the release of the LM101 in 1967, which solved a variety of issues, and the subsequent release of the μA741 in 1968. The μA741 was extremely similar to the LM101 except that Fairchild's facilities allowed them to include a 30 pF compensation capacitor inside the chip instead of requiring external compensation. This simple difference has made the 741 *the* canonical op-amp and many modern amps base their pinout on the 741s. The μA741 is still in production, and has become ubiquitous in electronics—many manufacturers produce a version of this classic chip, recognizable by part numbers containing *741*.

1970: First high-speed, low-input current FET design

In the 1970s high speed, low-input current designs started to be made by using FETs. These would be largely replaced by op-amps made with MOSFETs in the 1980s. During the 1970s single sided supply op-amps also became available.

1972: Single sided supply op-amps being produced

A single sided supply op-amp is one where the input and output voltages can be as low as the negative power supply voltage instead of needing to be at least two volts above it. The result is that it can operate in many applications with the negative supply pin on the op-amp being connected to the signal ground, thus eliminating the need for a separate negative power supply.

The LM324 (released in 1972) was one such op-amp that came in a quad package (four separate op-amps in one package) and became an industry standard. In addition to packaging multiple op-amps in a single package, the 1970s also saw the birth of op-amps

in hybrid packages. These op-amps were generally improved versions of existing monolithic op-amps. As the properties of monolithic op-amps improved, the more complex hybrid ICs were quickly relegated to systems that are required to have extremely long service lives or other specialty systems.

Recent trends

Recently supply voltages in analog circuits have decreased (as they have in digital logic) and low-voltage op-amps have been introduced reflecting this. Supplies of ±5 V and increasingly 5 V are common. To maximize the signal range modern op-amps commonly have rail-to-rail outputs and sometimes rail-to-rail inputs (the input signals can range from the lowest supply voltage to the highest).

Classification

Op-amps may be classified by their construction:

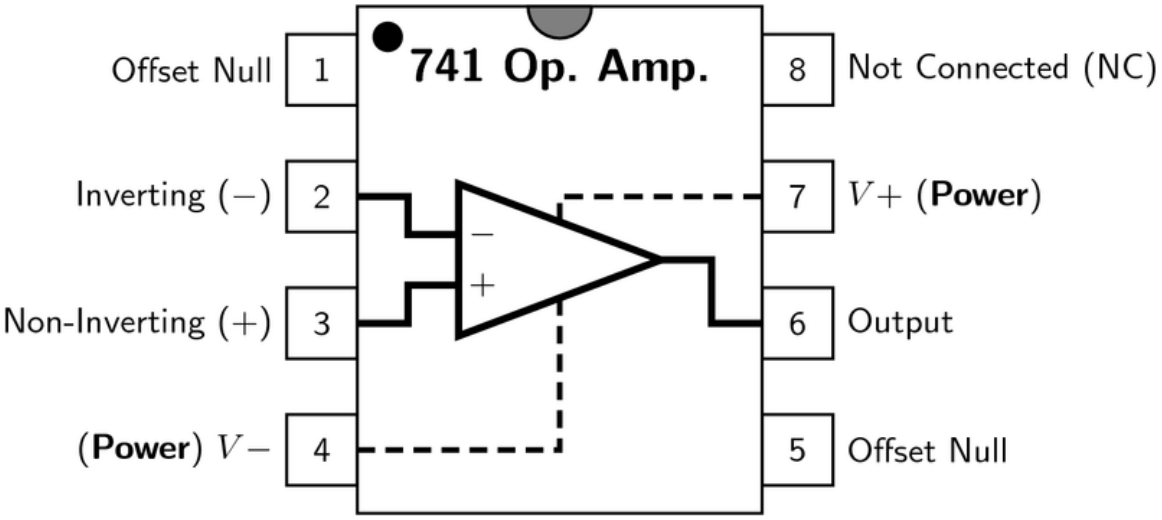
- discrete (built from individual transistors or tubes/valves)
- IC (fabricated in an Integrated circuit) - most common
- hybrid

IC op-amps may be classified in many ways, including:

- Military, Industrial, or Commercial grade (for example: the LM301 is the commercial grade version of the LM101, the LM201 is the industrial version). This may define operating temperature ranges and other environmental or quality factors.
- Classification by package type may also affect environmental hardiness, as well as manufacturing options; DIP, and other through-hole packages are tending to be replaced by Surface-mount devices.
- Classification by internal compensation: op-amps may suffer from high frequency instability in some negative feedback circuits unless a small compensation capacitor modifies the phase- and frequency- responses; op-amps with capacitor built in are termed "*compensated*", or perhaps compensated for closed-loop gains down to (say) 5, others: uncompensated.
- Single, dual and quad versions of many commercial op-amp IC are available, meaning 1, 2 or 4 operational amplifiers are included in the same package.
- Rail-to-rail input (and/or output) op-amps can work with input (and/or output) signals very close to the power supply rails.
- CMOS op-amps (such as the CA3140E) provide extremely high input resistances, higher than JFET-input op-amps, which are normally higher than bipolar-input op-amps.
- other varieties of op-amp include programmable op-amps (simply meaning the quiescent current, gain, bandwidth and so on can be adjusted slightly by an external resistor).

- manufacturers often tabulate their op-amps according to purpose, such as low-noise pre-amplifiers, wide bandwidth amplifiers, and so on.

Applications



DIP pinout for 741-type operational amplifier

Use in electronics system design

The use of op-amps as circuit blocks is much easier and clearer than specifying all their individual circuit elements (transistors, resistors, etc.), whether the amplifiers used are integrated or discrete. In the first approximation op-amps can be used as if they were ideal differential gain blocks; at a later stage limits can be placed on the acceptable range of parameters for each op-amp.

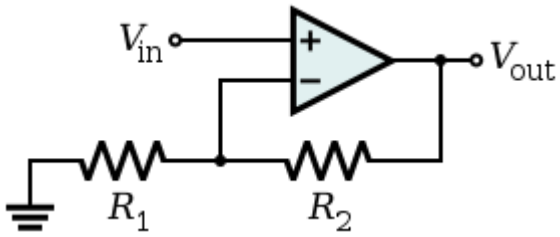
Circuit design follows the same lines for all electronic circuits. A specification is drawn up governing what the circuit is required to do, with allowable limits. For example, the gain may be required to be 100 times, with a tolerance of 5% but drift of less than 1% in a specified temperature range; the input impedance not less than one megohm; etc.

A basic circuit is designed, often with the help of circuit modeling (on a computer). Specific commercially available op-amps and other components are then chosen that meet the design criteria within the specified tolerances at acceptable cost. If not all criteria can be met, the specification may need to be modified.

A prototype is then built and tested; changes to meet or improve the specification, alter functionality, or reduce the cost, may be made.

Basic single stage amplifiers

Non-inverting amplifier



An op-amp connected in the non-inverting amplifier configuration

In a non-inverting amplifier, the output voltage changes in the same direction as the input voltage.

The gain equation for the op-amp is:

$$V_{\text{out}} = (V_+ - V_-)\, A_{OL}$$

However, in this circuit V_- is a function of V_{out} because of the negative feedback through the R_1R_2 network. R_1 and R_2 form a voltage divider, and as V_- is a high-impedance input, it does not load it appreciably. Consequently:

$$V_- = \beta \cdot V_{\text{out}}$$

where

$$\beta = \frac{R_1}{R_1 + R_2}$$

Substituting this into the gain equation, we obtain:

$$V_{\text{out}} = (V_{\text{in}} - \beta \cdot V_{\text{out}}) \cdot A_{OL}$$

Solving for V_{out}:

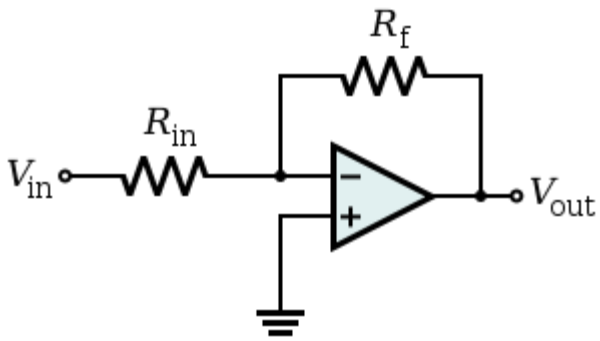
$$V_{\text{out}} = V_{\text{in}} \cdot \left(\frac{1}{\beta + 1/A_{OL}}\right)$$

If A_{OL} is very large, this simplifies to

$$V_{\text{out}} \approx \frac{V_{\text{in}}}{\beta} = \frac{V_{\text{in}}}{\frac{R_1}{R_1+R_2}} = V_{\text{in}}\left(1 + \frac{R_2}{R_1}\right)$$

.

Inverting amplifier



An op-amp connected in the inverting amplifier configuration

In an inverting amplifier, the output voltage changes in an opposite direction to the input voltage.

As for the non-inverting amplifier, we start with the gain equation of the op-amp:

$$V_{\text{out}} = (V_+ - V_-)\, A_{OL}$$

This time, V_- is a function of both V_{out} and V_{in} due to the voltage divider formed by R_{f} and R_{in}. Again, the op-amp input does not apply an appreciable load, so:

$$V_- = \frac{1}{R_{\text{f}} + R_{\text{in}}}\left(R_{\text{f}} V_{\text{in}} + R_{\text{in}} V_{\text{out}}\right)$$

Substituting this into the gain equation and solving for V_{out}:

$$V_{\text{out}} = -V_{\text{in}} \cdot \frac{A_{OL} R_{\text{f}}}{R_{\text{f}} + R_{\text{in}} + A_{OL} R_{\text{in}}}$$

If A_{OL} is very large, this simplifies to

$$V_{\text{out}} \approx -V_{\text{in}} \frac{R_{\text{f}}}{R_{\text{in}}}.$$

A resistor is often inserted between the non-inverting input and ground (so both inputs "see" similar resistances), reducing the input offset voltage due to different voltage drops due to bias current, and may reduce distortion in some op-amps.

A DC-blocking capacitor may be inserted in series with the input resistor when a frequency response down to DC is not needed and any DC voltage on the input is

unwanted. That is, the capacitive component of the input impedance inserts a DC zero and a low-frequency pole that gives the circuit a bandpass or high-pass characteristic.

Positive feedback configurations

Another typical configuration of op-amps is with positive feedback, which takes a fraction of the output signal back to the non-inverting input. An important application of it is the comparator with hysteresis, the Schmitt trigger.

Positive voltage level detector

A positive reference voltage V_{ref} is applied to one of the op-amp's inputs. This means that the op-amp is set up as a comparator to detect a positive voltage. If the voltage to be sensed, E_i, is applied to op amp's (+) input, the result is a noninverting positive-level detector. When E_i is above V_{ref}, V_O equals $+V_{sat}$. When E_i is below V_{ref}, V_O equals $-V_{sat}$.

If E_i, is applied to the inverting input, the circuit is an inverting positive-level detector: When E_i is above V_{ref}, V_O equals $-V_{sat}$.

Negative voltage level detector

A negative voltage detector is a circuit that detects when input signal E_i crosses the negative voltage $-V_{ref}$. When E_i is above $-V_{ref}$, V_O equals $+V_{sat}$. When $E_i$is below $-V_{ref}$, V_O equals $-V_{sat}$.When E_i is above $-V_{ref}$, V_O equals $-V_{sat}$, and when $E_i$is below $-V_{ref}$, V_O equals $+V_{sat}$.

Sine to square wave converter

The zero detector will convert the output of a sine-wave from a function generator into a variable-frequency square wave. If E_i is a sine wave, triangular wave, or wave of any other shape that is symmetrical around zero, the zero-crossing detector's output will be square.

Because of the wide slew-range and lack of positive feedback, the response of all the level detectors described above will be relatively slow. Using a general-purpose op-amp, for example, the frequency of E_i for the sine to square wave converter should probably be below 100 Hz.

Other applications

- audio- and video-frequency pre-amplifiers and buffers
- voltage comparators
- differential amplifiers
- differentiators and integrators
- filters
- precision rectifiers

- precision peak detectors
- voltage and current regulators
- analog calculators
- analog-to-digital converters
- digital-to-analog converter
- voltage clamps
- oscillators and waveform generators

Most single, dual and quad op-amps available have a standardized pin-out which permits one type to be substituted for another without wiring changes. A specific op-amp may be chosen for its open loop gain, bandwidth, noise performance, input impedance, power consumption, or a compromise between any of these factors.

Limitations of real op-amps

Real op-amps differ from the ideal model in various respects.

DC imperfections

Real operational amplifiers suffer from several non-ideal effects:

Finite gain
: Open-loop gain is infinite in the ideal operational amplifier but finite in real operational amplifiers. Typical devices exhibit open-loop DC gain ranging from 100,000 to over 1 million. So long as the loop gain (i.e., the product of open-loop and feedback gains) is very large, the circuit gain will be determined entirely by the amount of negative feedback (i.e., it will be independent of open-loop gain). In cases where closed-loop gain must be very high, the feedback gain will be very low, and the low feedback gain causes low loop gain; in these cases, the operational amplifier will cease to behave ideally.

Finite input impedances
: The *differential input impedance* of the operational amplifier is defined as the impedance *between* its two inputs; the *common-mode input impedance* is the impedance from each input to ground. MOSFET-input operational amplifiers often have protection circuits that effectively short circuit any input differences greater than a small threshold, so the input impedance can appear to be very low in some tests. However, as long as these operational amplifiers are used in a typical high-gain negative feedback application, these protection circuits will be inactive. The input bias and leakage currents described below are a more important design parameter for typical operational amplifier applications.

Non-zero output impedance
: Low output impedance is important for low-impedance loads; for these loads, the voltage drop across the output impedance of the amplifier will be significant. Hence, the output impedance of the amplifier limits the maximum power that can be provided. In a negative-feedback configuration, the output impedance of the amplifier is effectively lowered; thus, in linear applications, op-amps usually

exhibit a very low output impedance indeed. Negative feedback can not, however, reduce the limitations that R_{load} in conjunction with R_{out} place on the maximum and minimum possible output voltages; it can only reduce output errors *within* that range.

Low-impedance outputs typically require high quiescent (i.e., idle) current in the output stage and will dissipate more power, so low-power designs may purposely sacrifice low output impedance.

Input current

Due to biasing requirements or leakage, a small amount of current (typically ~10 nanoamperes for bipolar op-amps, tens of picoamperes for JFET input stages, and only a few pA for MOSFET input stages) flows into the inputs. When large resistors or sources with high output impedances are used in the circuit, these small currents can produce large unmodeled voltage drops. If the input currents are matched, *and* the impedance looking *out* of *both* inputs are matched, then the voltages produced at each input will be equal. Because the operational amplifier operates on the *difference* between its inputs, these matched voltages will have no effect (unless the operational amplifier has poor CMRR, which is described below). It is more common for the input currents (or the impedances looking out of each input) to be slightly mismatched, and so a small *offset voltage* can be produced. This offset voltage can create offsets or drifting in the operational amplifier. It can often be nulled externally; however, many operational amplifiers include *offset null* or *balance* pins and some procedure for using them to remove this offset. Some operational amplifiers attempt to nullify this offset automatically.

Input offset voltage

This voltage, which is what is required across the op-amp's input terminals to drive the output voltage to zero, is related to the mismatches in input bias current. In the perfect amplifier, there would be no input offset voltage. However, it exists in actual op-amps because of imperfections in the differential amplifier that constitutes the input stage of the vast majority of these devices. Input offset voltage creates two problems: First, due to the amplifier's high voltage gain, it virtually assures that the amplifier output will go into saturation if it is operated without negative feedback, even when the input terminals are wired together. Second, in a closed loop, negative feedback configuration, the input offset voltage is amplified along with the signal and this may pose a problem if high precision DC amplification is required or if the input signal is very small.

Common mode gain

A perfect operational amplifier amplifies only the voltage difference between its two inputs, completely rejecting all voltages that are common to both. However, the differential input stage of an operational amplifier is never perfect, leading to the amplification of these identical voltages to some degree. The standard measure of this defect is called the common-mode rejection ratio (denoted CMRR). Minimization of common mode gain is usually important in non-inverting amplifiers (described below) that operate at high amplification.

Temperature effects

All parameters change with temperature. Temperature drift of the input offset voltage is especially important.

Power-supply rejection

The output of a perfect operational amplifier will be completely independent from ripples that arrive on its power supply inputs. Every real operational amplifier has a specified power supply rejection ratio (PSRR) that reflects how well the op-amp can reject changes in its supply voltage. Copious use of bypass capacitors can improve the PSRR of many devices, including the operational amplifier.

Drift

Real op-amp parameters are subject to slow change over time and with changes in temperature, input conditions, etc.

Noise

Amplifiers generate random voltage at the output even when there is no signal applied. This can be due to thermal noise and flicker noise of the devices. For applications with high gain or high bandwidth, noise becomes a very important consideration.

AC imperfections

The op-amp gain calculated at DC does not apply at higher frequencies. To a first approximation, the gain of a typical op-amp is inversely proportional to frequency. This means that an op-amp is characterized by its gain-bandwidth product. For example, an op-amp with a gain bandwidth product of 1 MHz would have a gain of 5 at 200 kHz, and a gain of 1 at 1 MHz. This low-pass characteristic is introduced deliberately, because it tends to stabilize the circuit by introducing a dominant pole. This is known as frequency compensation.

Typical low cost, general purpose op-amps exhibit a gain bandwidth product of a few megahertz. Specialty and high speed op-amps can achieve gain bandwidth products of hundreds of megahertz. For very high-frequency circuits, a completely different form of op-amp called the current-feedback operational amplifier is often used.

Other imperfections include:

Finite bandwidth

All amplifiers have a finite bandwidth. This creates several problems for op amps. First, associated with the bandwidth limitation is a phase difference between the input signal and the amplifier output that can lead to oscillation in some feedback circuits. The internal frequency compensation used in some op amps to increase the gain or phase margin intentionally reduces the bandwidth even further to maintain output stability when using a wide variety of feedback networks. Second, reduced bandwidth results in lower amounts of feedback at higher frequencies, producing higher distortion, noise, and output impedance and also reduced output phase linearity as the frequency increases.

Input capacitance

Most important for high frequency operation because it further reduces the open loop bandwidth of the amplifier.

Non-linear imperfections

Saturation
: output voltage is limited to a minimum and maximum value close to the power supply voltages. Saturation occurs when the output of the amplifier reaches this value and is usually due to:

  - In the case of an op-amp using a bipolar power supply, a voltage gain that produces an output that is more positive or more negative than that maximum or minimum; or
  - In the case of an op-amp using a single supply voltage, either a voltage gain that produces an output that is more positive than that maximum, or a signal so close to ground that the amplifier's gain is not sufficient to raise it above the lower threshold.

Slewing
: the amplifier's output voltage reaches its maximum rate of change. Measured as the slew rate, it is usually specified in volts per microsecond. When slewing occurs, further increases in the input signal have no effect on the rate of change of the output. Slewing is usually caused by internal capacitances in the amplifier, especially those used to implement its frequency compensation.

Non-linear input-output relationship
: The output voltage may not be accurately proportional to the difference between the input voltages. It is commonly called distortion when the input signal is a waveform. This effect will be very small in a practical circuit if substantial negative feedback is used.

Power considerations

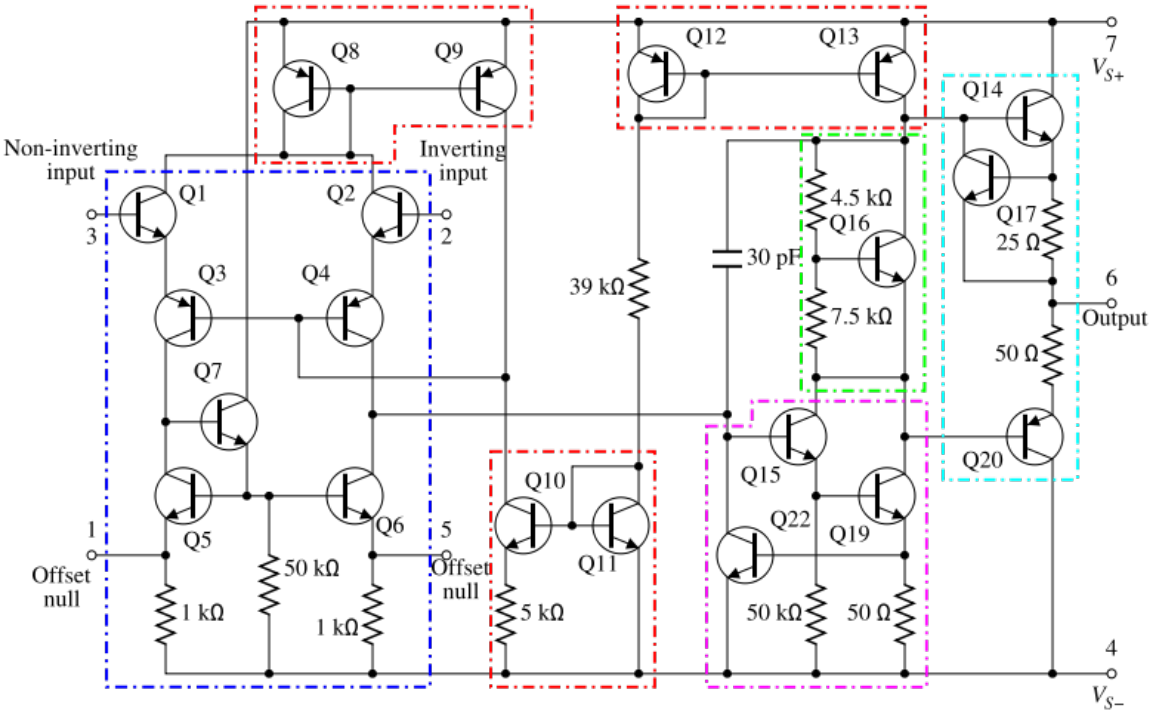
Limited output current
: The output current must be finite. In practice, most op-amps are designed to limit the output current so as not to exceed a specified level — around 25 mA for a type 741 IC op-amp — thus protecting the op-amp and associated circuitry from damage. Modern designs are electronically more rugged than earlier implementations and some can sustain direct short circuits on their outputs without damage.

Limited dissipated power
: The output current flows through the op-amp's internal output impedance, dissipating heat. If the op-amp dissipates too much power, then its temperature will increase above some safe limit. The op-amp may enter thermal shutdown, or it may be destroyed.

Modern integrated FET or MOSFET op-amps approximate more closely the ideal op-amp than bipolar ICs when it comes to input impedance and input bias and offset currents. Bipolars are generally better when it comes to input *voltage* offset, and often have lower noise. Generally, at room temperature, with a fairly large signal, and limited bandwidth, FET and MOSFET op-amps now offer better performance.

Internal circuitry of 741 type op-amp

Though designs vary between products and manufacturers, all op-amps have basically the same internal structure, which consists of three stages:



A component level diagram of the common 741 op-amp. Dotted lines outline: current mirrors (red); differential amplifier (blue); class A gain stage (magenta); voltage level shifter (green); output stage (cyan).

1. Differential amplifier – provides low noise amplification, high input impedance, usually a differential output.
2. Voltage amplifier – provides high voltage gain, a single-pole frequency roll-off, usually single-ended output.
3. Output amplifier – provides high current driving capability, low output impedance, current limiting and short circuit protection circuitry.

Input stage

Constant-current stabilization system

The input stage DC conditions are stabilized by a high-gain negative feedback system whose main parts are the two current mirrors on the left of the figure, outlined in red. The main purpose of this negative feedback system—to supply the differential input stage with a stable constant current—is realized as follows.

The current through the 39 kΩ resistor acts as a current reference for the other bias currents used in the chip. The voltage across the resistor is equal to the voltage across the supply rails ($V_{S+} - V_{S-}$) minus two transistor diode drops (i.e., from Q11 and Q12), and so the current has value $I_{\text{ref}} = (V_{S+} - V_{S-} - 2V_{be})/(39\ \text{k}\Omega)$. The Widlar current mirror built by Q10, Q11, and the 5 kΩ resistor produces a very small fraction of I_{ref} at the Q10 collector. This small constant current through Q10's collector supplies the base currents for Q3 and Q4 as well as the Q9 collector current. The Q8/Q9 current mirror tries to make Q9's collector current the same as the Q3 and Q4 *collector* currents. Thus Q3 and Q4's combined base currents (which are of the same order as the overall chip's input currents) will be a small fraction of the already small Q10 current.

So, if the input stage current increases for any reason, the Q8/Q9 current mirror will draw current away from the bases of Q3 and Q4, which reduces the input stage current, and vice versa. The feedback loop also isolates the rest of the circuit from common-mode signals by making the base voltage of Q3/Q4 follow tightly $2V_{be}$ below the higher of the two input voltages.

Differential amplifier

The blue outlined section is a differential amplifier. Q1 and Q2 are input emitter followers and together with the common base pair Q3 and Q4 form the differential input stage. In addition, Q3 and Q4 also act as level shifters and provide voltage gain to drive the class A amplifier. They also help to increase the reverse V_{be} rating on the input transistors (the emitter-base junctions of the NPN transistors Q1 and Q2 break down at around 7 V but the PNP transistors Q3 and Q4 have breakdown voltages around 50 V).

The differential amplifier formed by Q1–Q4 drives a current mirror active load formed by transistors Q5–Q7 (actually, Q6 is the very active load). Q7 increases the accuracy of the current mirror by decreasing the amount of signal current required from Q3 to drive the bases of Q5 and Q6. This configuration provides differential to single ended conversion as follows:

The signal current of Q3 is the input to the current mirror while the output of the mirror (the collector of Q6) is connected to the collector of Q4. Here, the signal currents of Q3 and Q4 are summed. For differential input signals, the signal currents of Q3 and Q4 are equal and opposite. Thus, the sum is twice the individual signal currents. This completes the differential to single ended conversion.

The open circuit signal voltage appearing at this point is given by the product of the summed signal currents and the paralleled collector resistances of Q4 and Q6. Since the collectors of Q4 and Q6 appear as high resistances to the signal current, the open circuit voltage gain of this stage is very high.

The base current at the inputs is not zero and the effective (differential) input impedance of a 741 is about 2 MΩ. The "offset null" pins may be used to place external resistors in parallel with the two 1 kΩ resistors (typically in the form of the two ends of a potentiometer) to adjust the balancing of the Q5/Q6 current mirror and thus indirectly control the output of the op-amp when zero signal is applied between the inputs.

Class A gain stage

The section outlined in magenta is the class A gain stage. The top-right current mirror Q12/Q13 supplies this stage by a constant current load, via the collector of Q13, that is largely independent of the output voltage. The stage consists of two NPN transistors in a Darlington configuration and uses the output side of a current mirror as its collector load to achieve high gain. The 30 pF capacitor provides frequency selective negative feedback around the class A gain stage as a means of frequency compensation to stabilise the amplifier in feedback configurations. This technique is called Miller compensation and functions in a similar manner to an op-amp integrator circuit. It is also known as 'dominant pole compensation' because it introduces a dominant pole (one which masks the effects of other poles) into the open loop frequency response. This pole can be as low as 10 Hz in a 741 amplifier and it introduces a −3 dB loss into the open loop response at this frequency. This internal compensation is provided to achieve unconditional stability of the amplifier in negative feedback configurations where the feedback network is non-reactive and the closed loop gain is unity or higher. Hence, the use of the operational amplifier is simplified because no external compensation is required for unity gain stability; amplifiers without this internal compensation may require external compensation or closed loop gains significantly higher than unity.

Output bias circuitry

The green outlined section (based on Q16) is a voltage level shifter or rubber diode (i.e., a V_{BE} multiplier); a type of voltage source. In the circuit as shown, Q16 provides a constant voltage drop between its collector and emitter regardless of the current through the circuit. If the base current to the transistor is assumed to be zero, and the voltage between base and emitter (and across the 7.5 kΩ resistor) is 0.625 V (a typical value for a BJT in the active region), then the current through the 4.5 kΩ resistor will be the same as that through the 7.5 kΩ, and will produce a voltage of 0.375 V across it. This keeps the voltage across the transistor, and the two resistors at 0.625 + 0.375 = 1 V. This serves to bias the two output transistors slightly into conduction reducing crossover distortion. In some discrete component amplifiers this function is achieved with (usually two) silicon diodes.

Output stage

The output stage (outlined in cyan) is a Class AB push-pull emitter follower (Q14, Q20) amplifier with the bias set by the V_{be} multiplier voltage source Q16 and its base resistors. This stage is effectively driven by the collectors of Q13 and Q19. Variations in the bias with temperature, or between parts with the same type number, are common so crossover distortion and quiescent current may be subject to significant variation. The output range of the amplifier is about one volt less than the supply voltage, owing in part to V_{be} of the output transistors Q14 and Q20.

The 25 Ω resistor in the output stage acts as a current sense to provide the output current-limiting function which limits the current in the emitter follower Q14 to about 25 mA for the 741. Current limiting for the negative output is done by sensing the voltage across Q19's emitter resistor and using this to reduce the drive into Q15's base. Later versions of this amplifier schematic may show a slightly different method of output current limiting. The output resistance is not zero, as it would be in an ideal op-amp, but with negative feedback it approaches zero at low frequencies.

Note: while the 741 was historically used in audio and other sensitive equipment, such use is now rare because of the improved noise performance of more modern op-amps. Apart from generating noticeable hiss, 741s and other older op-amps may have poor common-mode rejection ratios and so will often introduce cable-borne mains hum and other common-mode interference, such as switch 'clicks', into sensitive equipment.

The "741" has come to often mean a generic op-amp IC (such as uA741, LM301, 558, LM324, TBA221 - or a more modern replacement such as the TL071). The description of the 741 output stage is qualitatively similar for many other designs (that may have quite different input stages), except:

- Some devices (uA748, LM301, LM308) are not internally compensated (require an external capacitor from output to some point within the operational amplifier, if used in low closed-loop gain applications).
- Some modern devices have rail-to-rail output capability (output can be taken to positive or negative power supply rail within a few millivolts).

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.

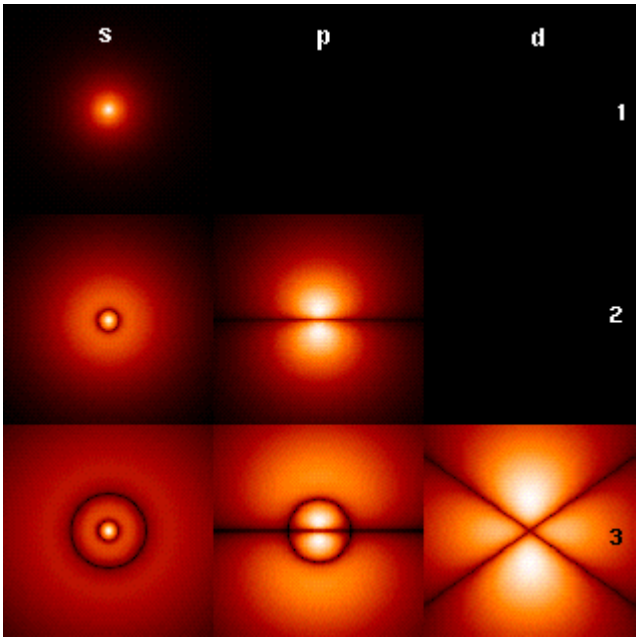
Silicon is used to create most semiconductors commercially. Dozens of other materials are used, including germanium, gallium arsenide, and silicon carbide. A pure semiconductor is often called an “intrinsic” semiconductor. The electronic properties and the conductivity of a semiconductor can be changed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

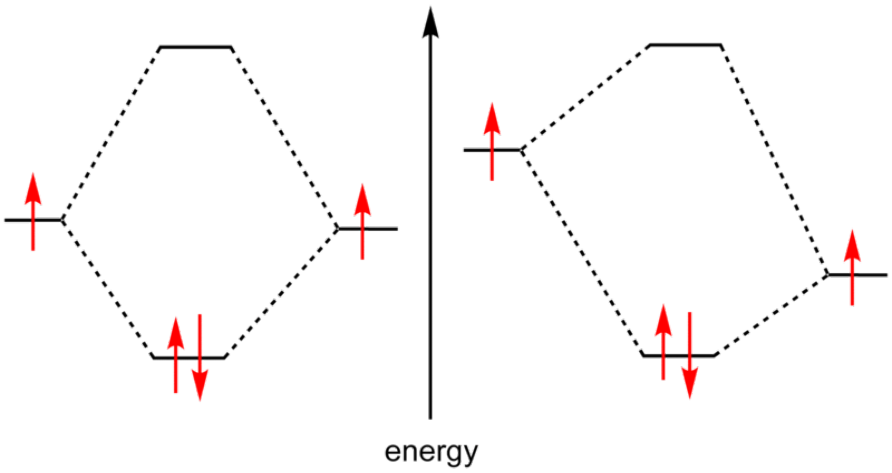
There are three popular ways to classify the electronic structure of a crystal.

- Band structure

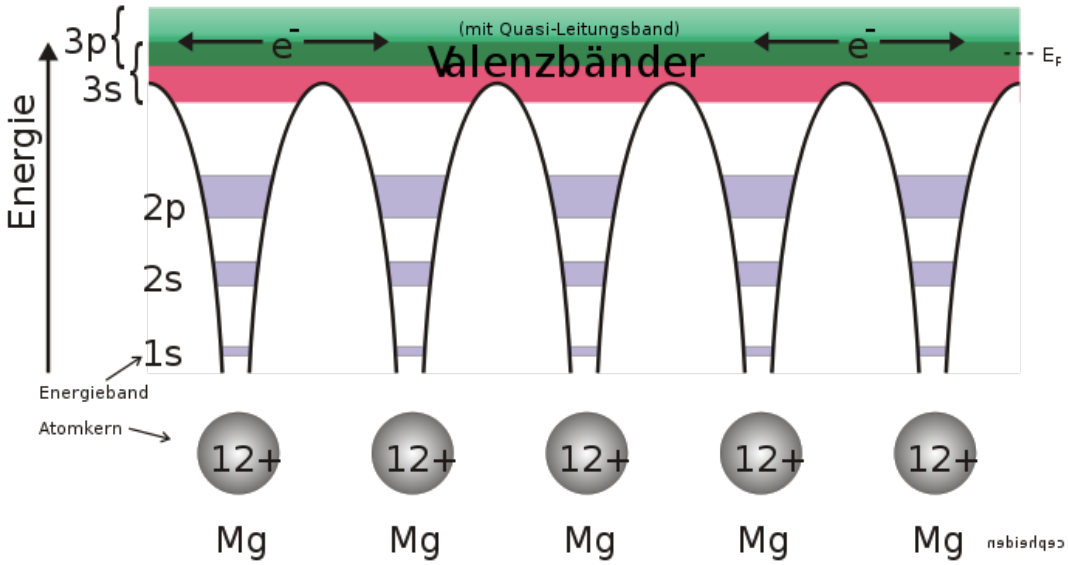
atoms – crystal – vacuum



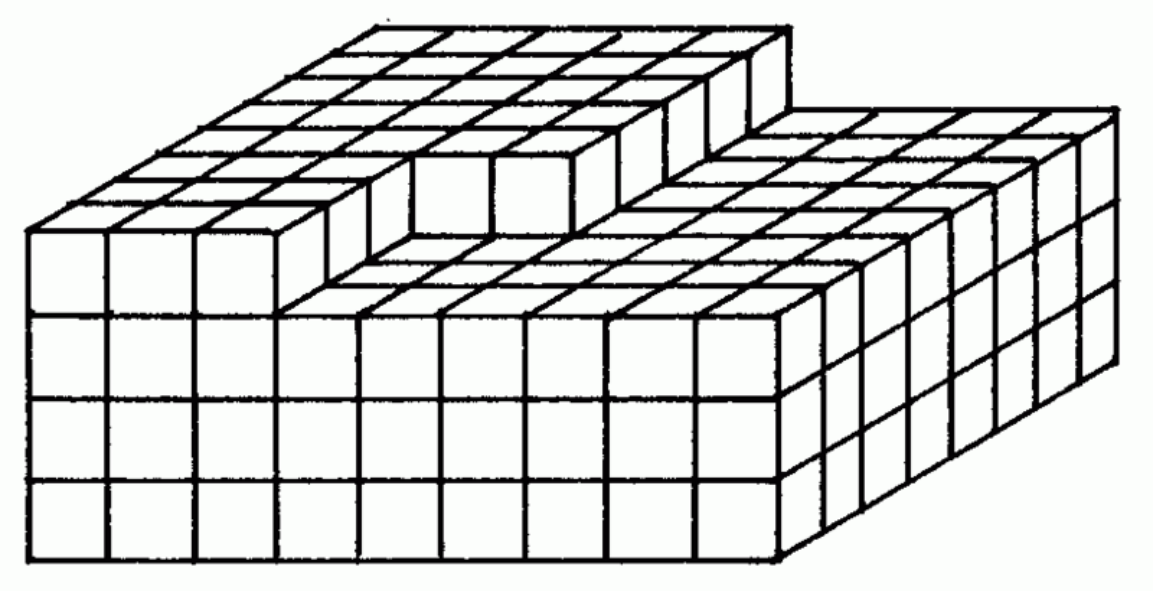
In a single H-atom an electron resides in well known orbitals. Note that the orbitals are called s,p,d in order of increasing circular current.



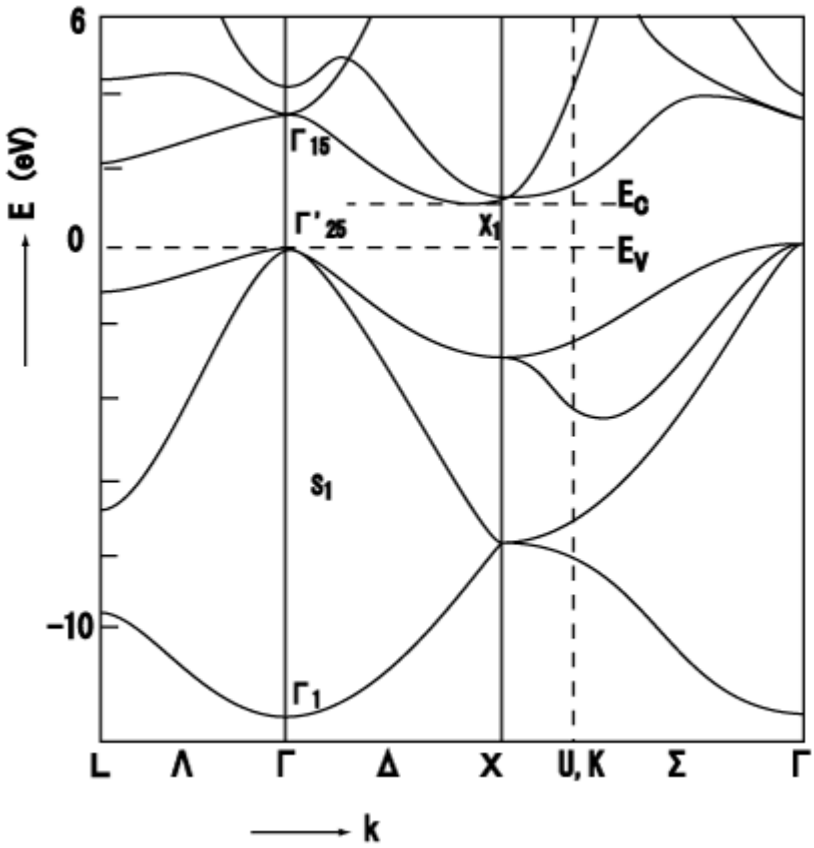
Putting two atoms together leads to delocalized orbitals across two atoms, yielding a covalent bond. Due to the Pauli exclusion principle, every state can contain only one electron.



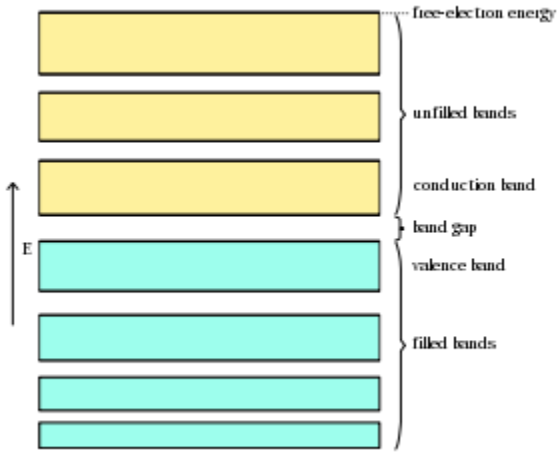
This can be continued with more atoms. Note: This picture shows a metal, not an actual semiconductor.



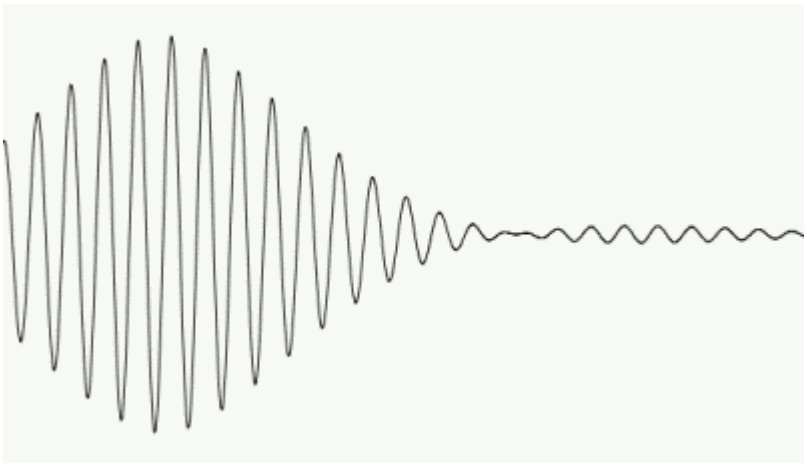
Continuing to add creates a crystal, which may then be cut into a tape and fused together at the ends to allow circular currents.



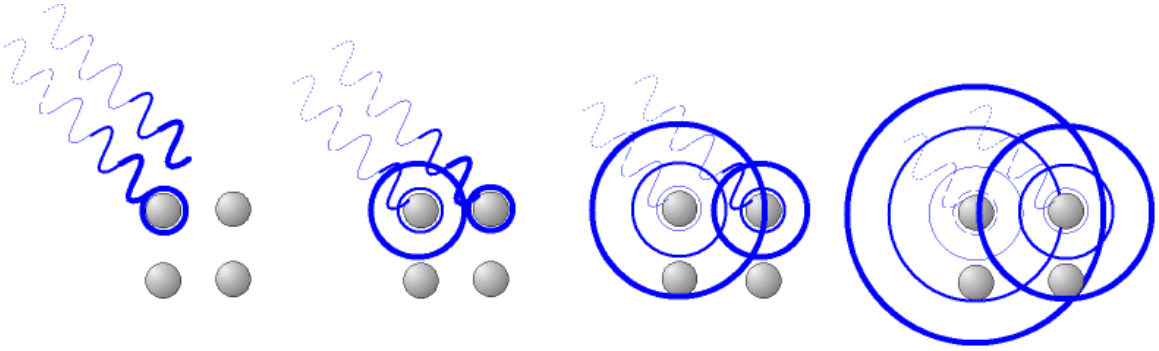
For this regular solid the band structure can be calculated or measured.



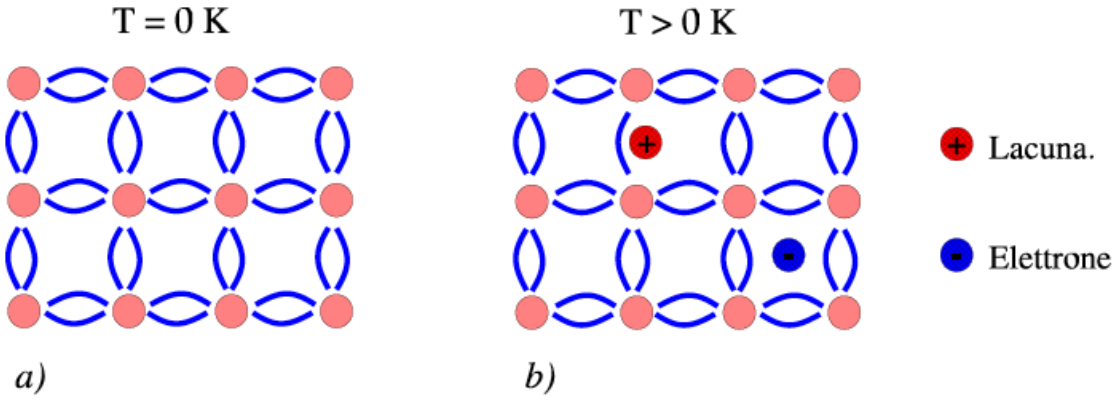
Integrating over the k axis gives the bands of a semiconductor showing a full valence band and an empty conduction band. Generally stopping at the vacuum level is undesirable, because some people want to calculate: photoemission, inverse photoemission



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.

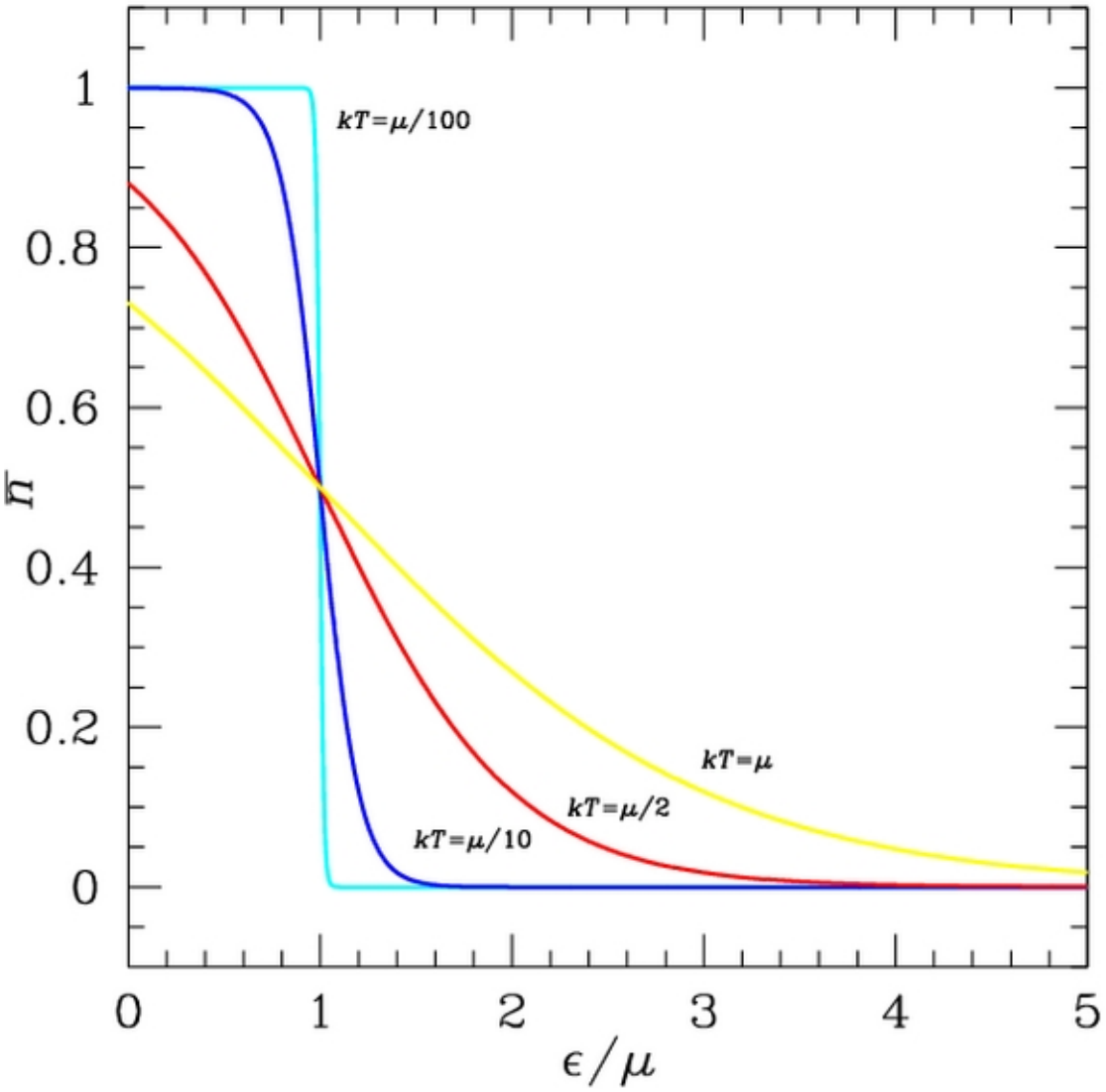
With covalent bonds, an electron moves by hopping to a neighboring bond. The Pauli exclusion principle requires the electron to be lifted into the higher anti-bonding state of that bond. For delocalized states, for example in one dimension – that is in a nanowire, for every energy there is a state with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

$$n = p = n_i.$$

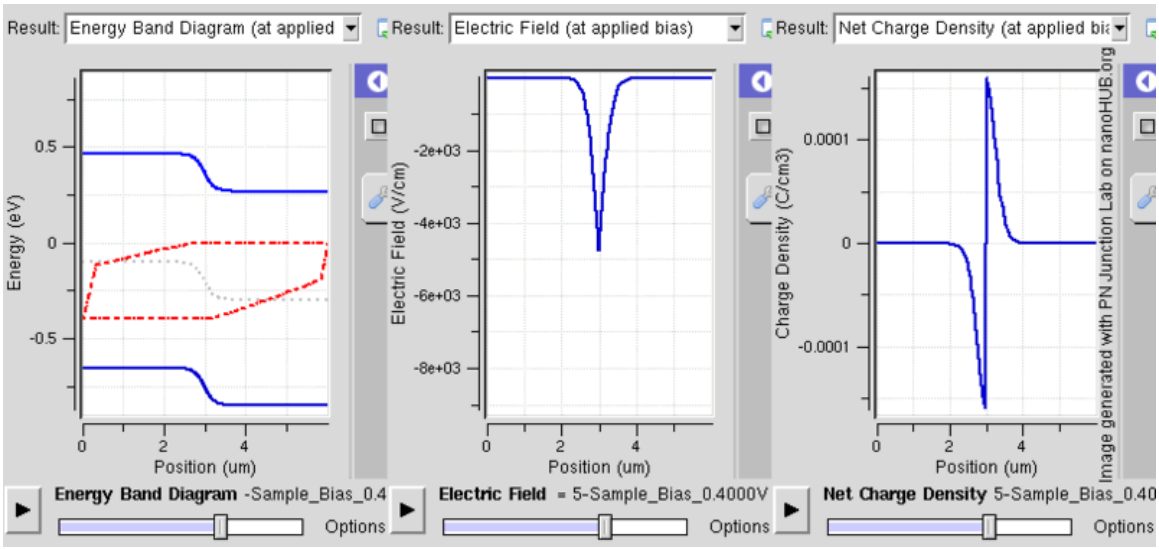
If we have a non-intrinsic semiconductor in thermal equilibrium the relation becomes:

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

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

Diode

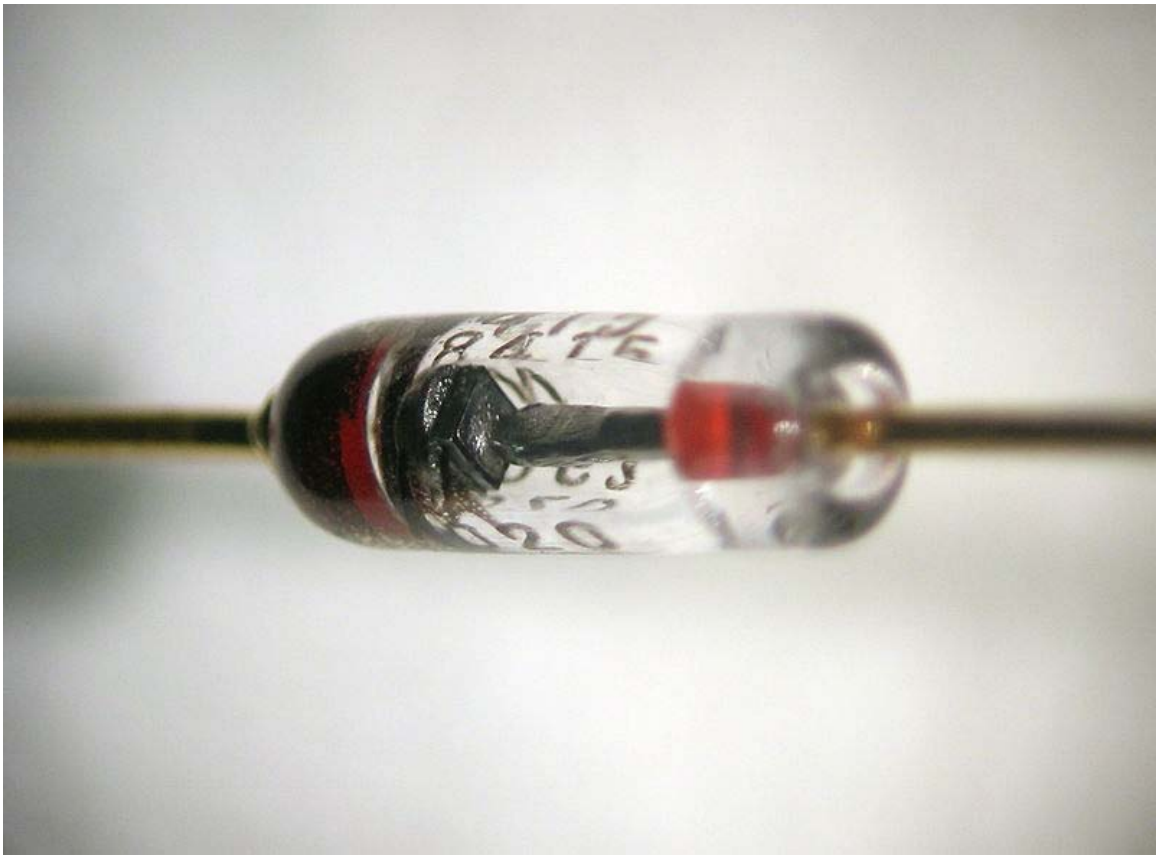


Figure 1: Closeup of a diode, showing the square shaped semiconductor crystal *(black object on left)*.

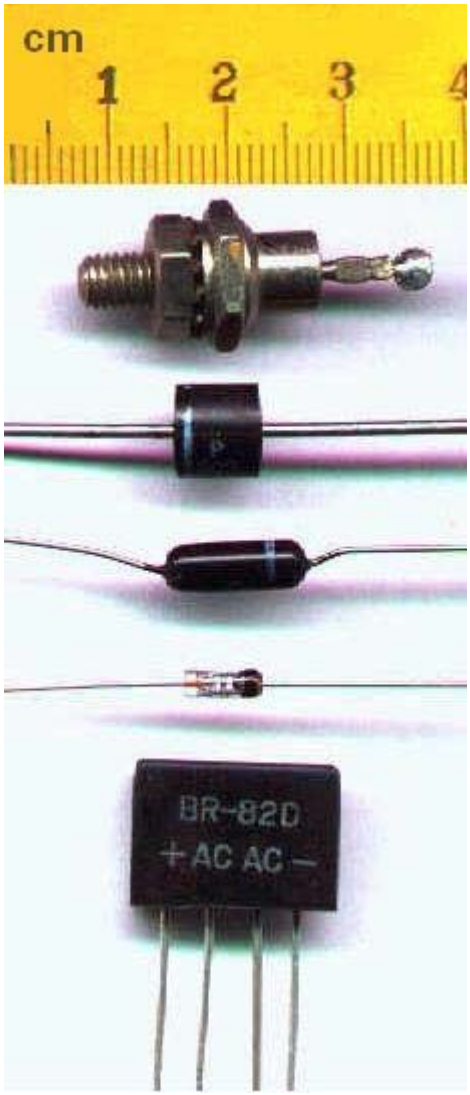


Figure 2: Various semiconductor diodes. Bottom: A bridge rectifier. In most diodes, a white or black painted band identifies the cathode terminal, that is, the terminal which conventional current flows out of when the diode is conducting.

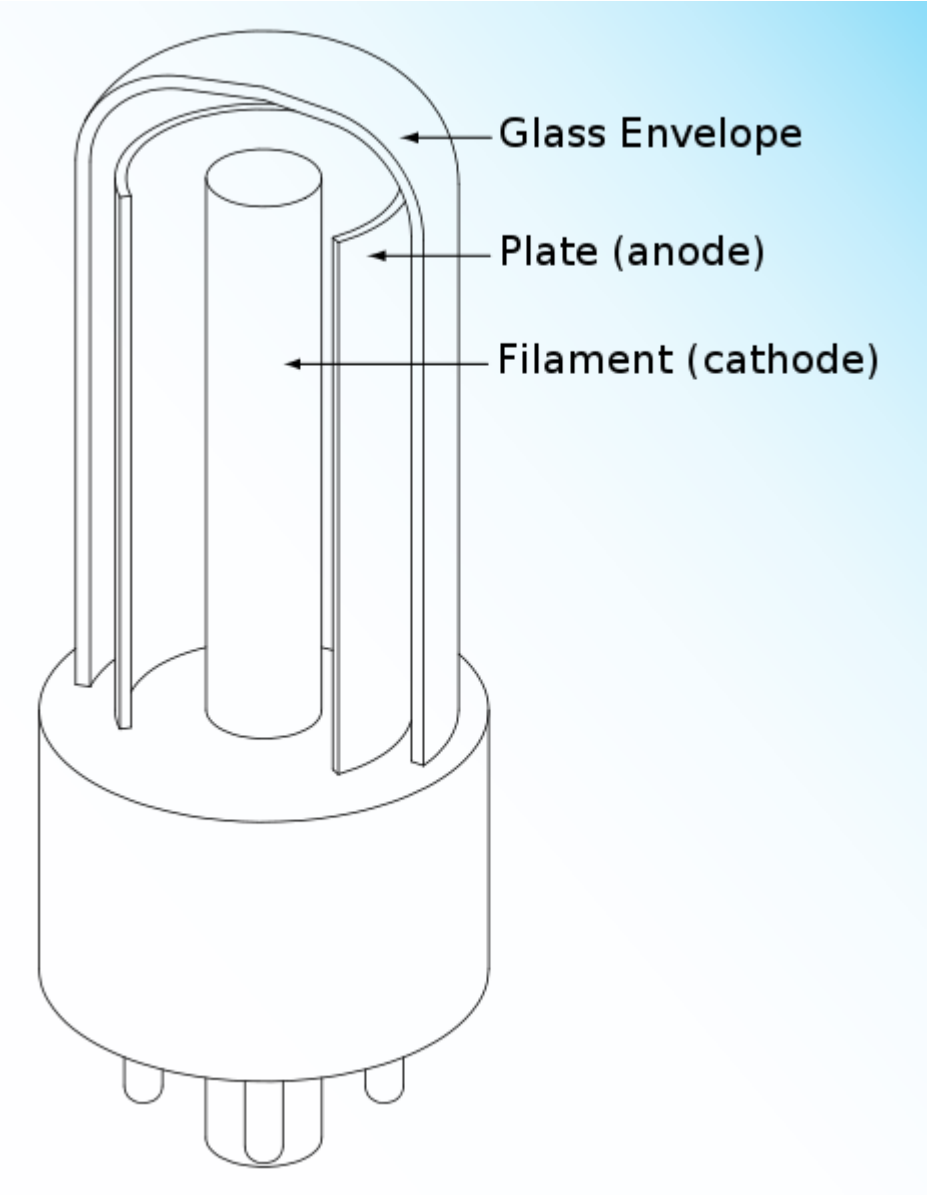


Figure 3: Structure of a vacuum tube diode. The filament may be bare, or more commonly (as shown here), embedded within and insulated from an enclosing cathode

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

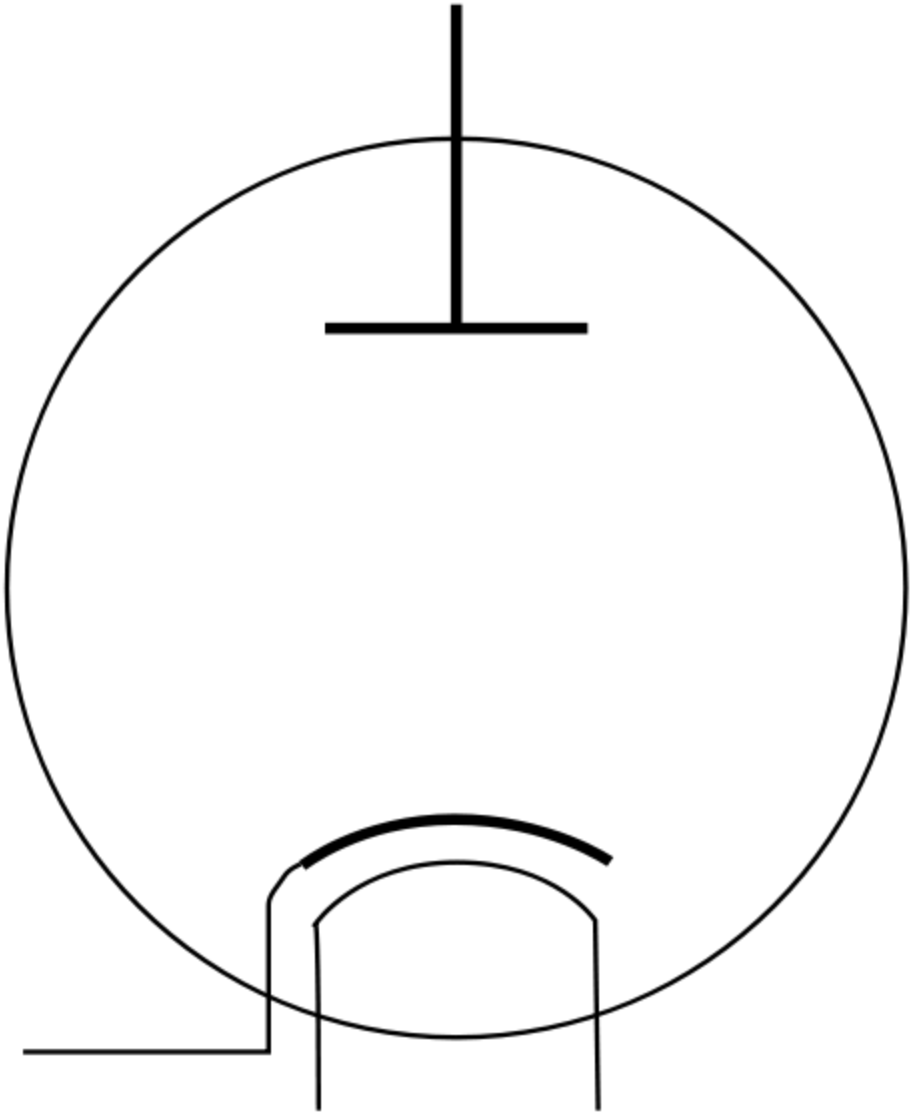


Figure 4: The symbol for an indirect heated vacuum tube diode. From top to bottom, the components are the anode, the cathode, and the heater filament.

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

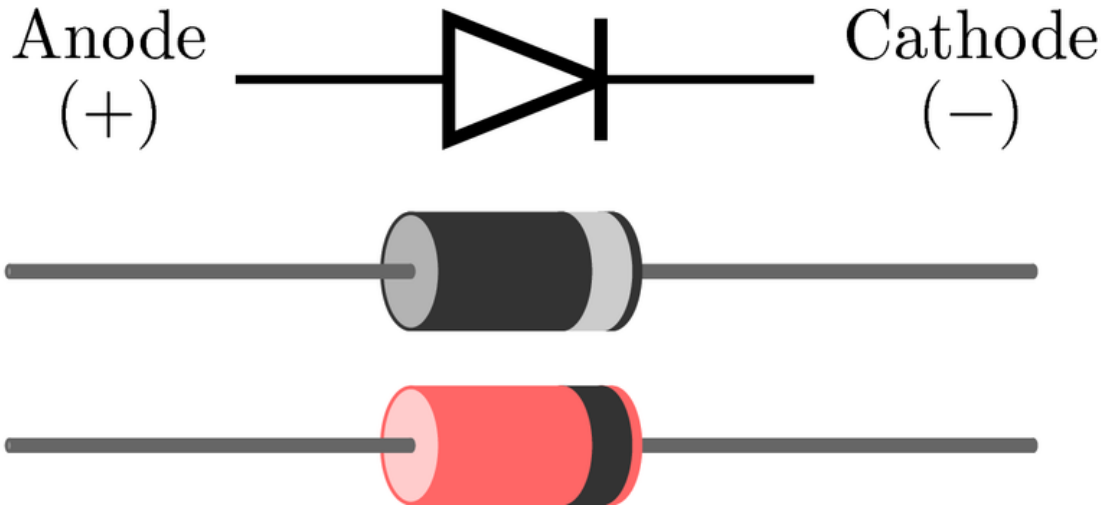


Figure 7: Typical diode packages in same alignment as diode symbol. Thin bar depicts the cathode.

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. 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.

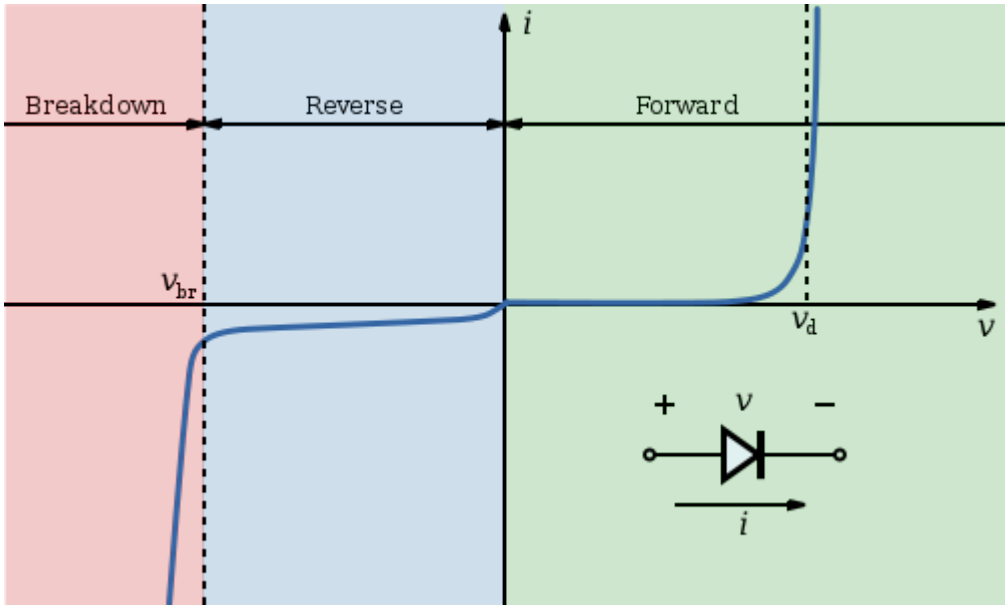


Figure 5: I–V characteristics of a P-N junction diode (not to scale).

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 (see Figure 5) 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 (see Figure 5) 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

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

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! That is to say, current will effectively flow from the cathode to the anode! 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.

Types of semiconductor diode

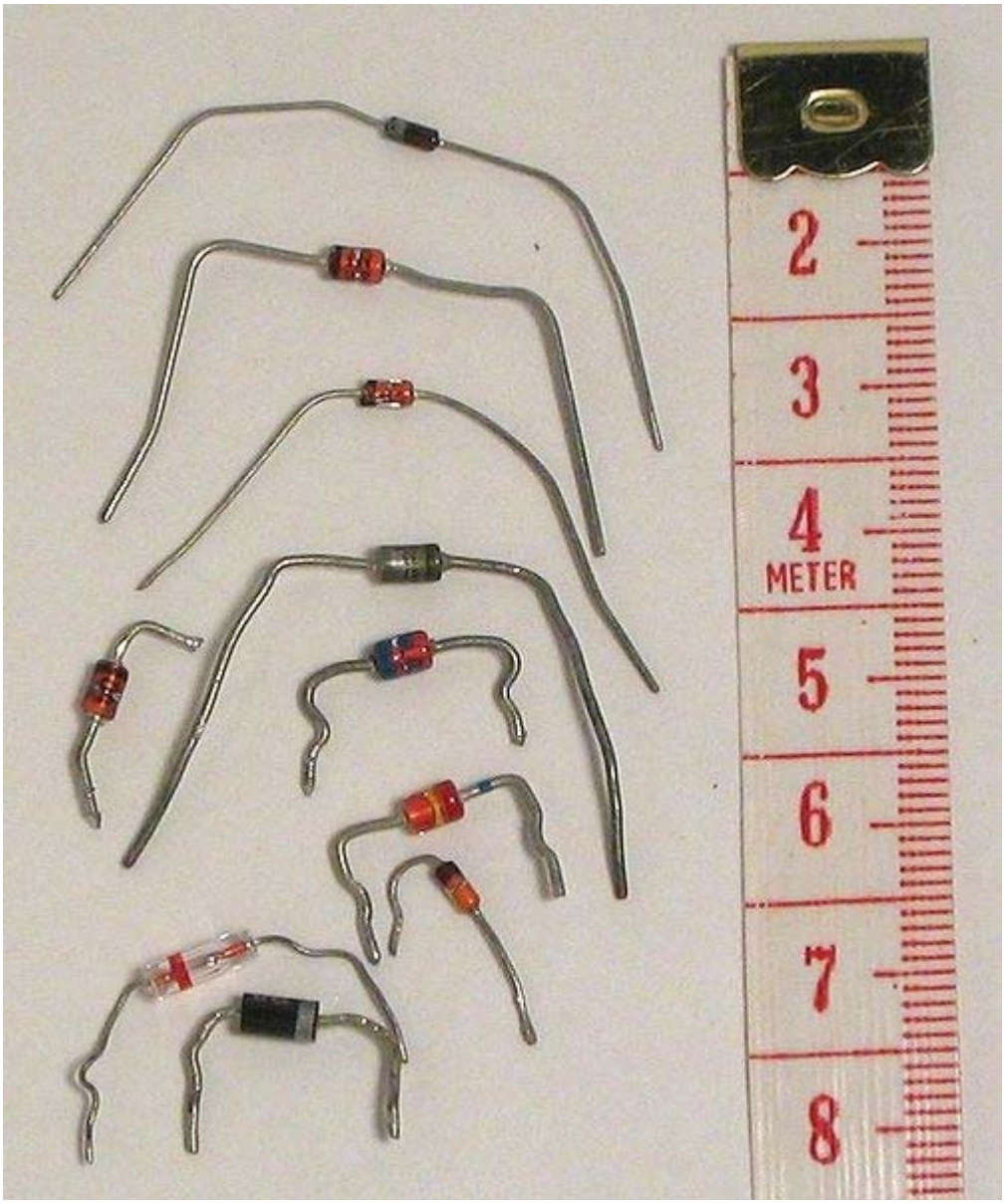


Figure 8: Several types of diodes. The scale is centimeters.

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 (see above). 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.

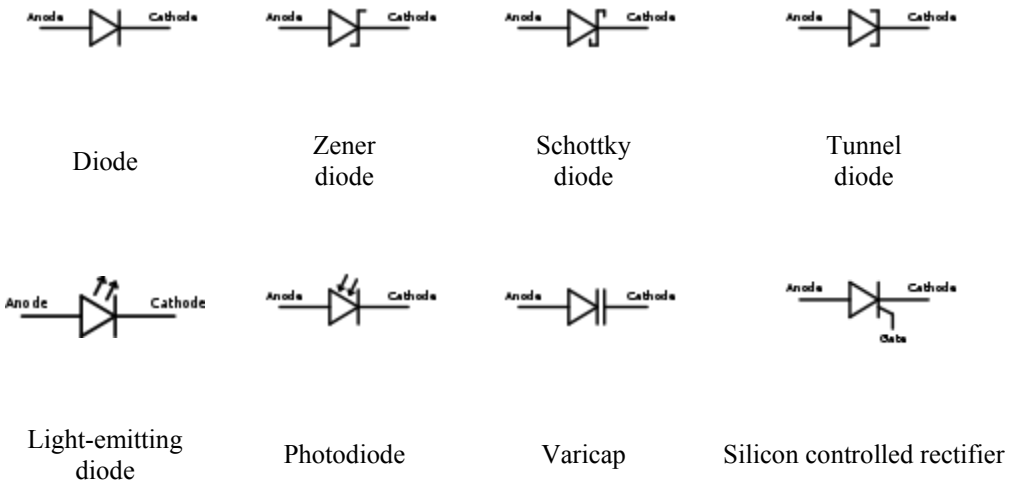


Figure 6: Some diode symbols.

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.

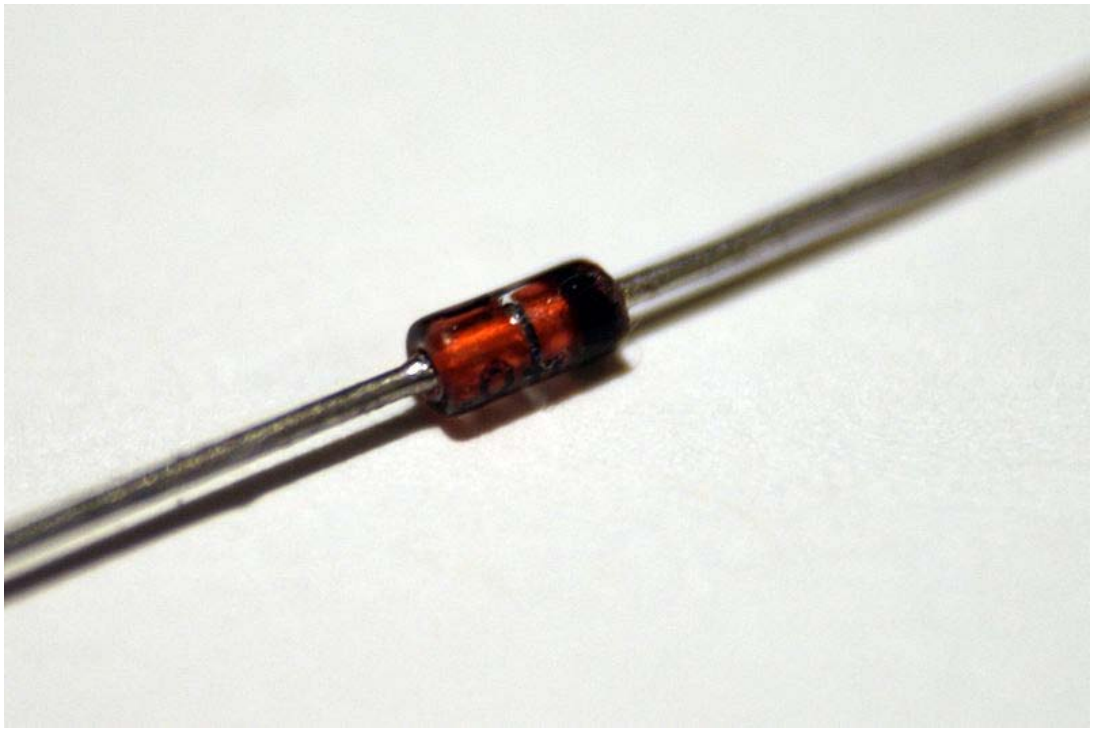
Diodes are also used in electronic musical keyboards. To reduce the amount of wiring needed in electronic musical keyboards, these instruments often use keyboard matrix circuits. The keyboard controller scans the rows and columns to determine which note the player has pressed. The problem with matrix 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

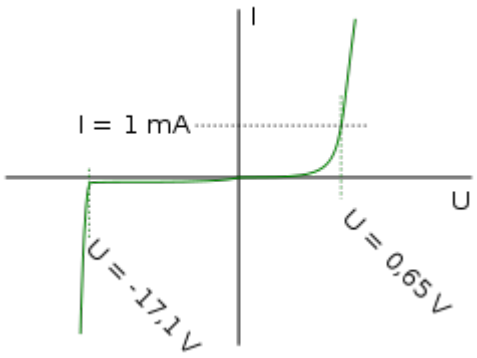
Diodes are usually referred to as *D* for diode on PCBs. Sometimes the abbreviation *CR* for **crystal rectifier** is used.

Chapter- 4

Zener Diode



Zener diode



Current-voltage characteristic of a Zener diode with a breakdown voltage of 17 volts. Notice the change of voltage scale between the forward biased (positive) direction and the reverse biased (negative) direction.

A **Zener diode** is a type of diode that permits current not only in the forward direction like a normal diode, but also in the reverse direction if the voltage is larger than the breakdown voltage known as "Zener knee voltage" or "Zener voltage". The device was named after Clarence Zener, who discovered this electrical property.

A conventional solid-state diode will not allow significant current if it is reverse-biased below its reverse breakdown voltage. When the reverse bias breakdown voltage is exceeded, a conventional diode is subject to high current due to avalanche breakdown. Unless this current is limited by circuitry, the diode will be permanently damaged due to overheating. In case of large forward bias (current in the direction of the arrow), the diode exhibits a voltage drop due to its junction built-in voltage and internal resistance. The amount of the voltage drop depends on the semiconductor material and the doping concentrations.

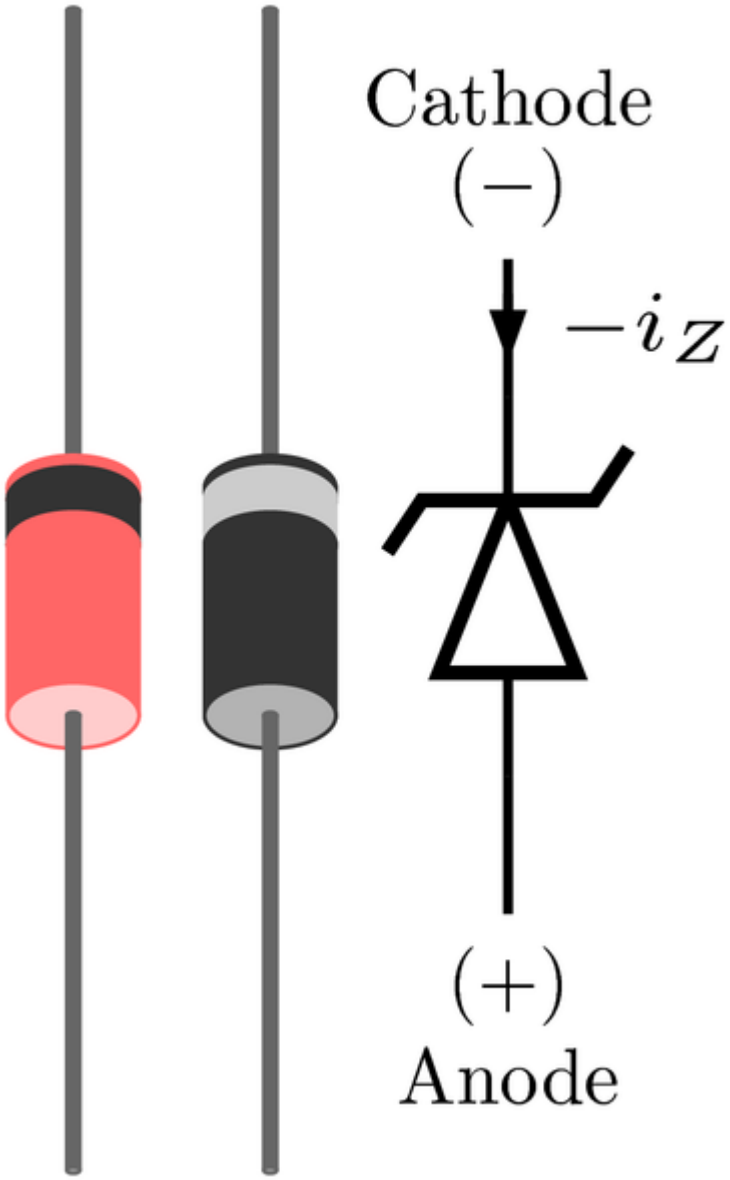
A Zener diode exhibits almost the same properties, except the device is specially designed so as to have a greatly reduced breakdown voltage, the so-called Zener voltage. By contrast with the conventional device, a reverse-biased Zener diode will exhibit a controlled breakdown and allow the current to keep the voltage across the Zener diode close to the Zener voltage. For example, a diode with a Zener breakdown voltage of 3.2 V will exhibit a voltage drop of very nearly 3.2 V across a wide range of reverse currents. The Zener diode is therefore ideal for applications such as the generation of a reference voltage (e.g. for an amplifier stage), or as a voltage stabilizer for low-current applications.

The Zener diode's operation depends on the heavy doping of its 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. In the atomic scale, this tunneling corresponds to the transport of valence band electrons into the empty conduction band states; as a result of the reduced barrier between these bands and high electric fields that are induced due to the relatively high levels of dopings on both sides. The breakdown voltage can be controlled quite accurately in the doping process. While tolerances within 0.05% are available, the most widely used tolerances are 5% and 10%. Breakdown voltage for commonly available zener diodes can vary widely from 1.2 volts to 200 volts.

Another mechanism that produces a similar effect is the avalanche effect as in the avalanche diode. The two types of diode are in fact constructed the same way and both effects are present in diodes of this type. In silicon diodes up to about 5.6 volts, the Zener effect is the predominant effect and shows a marked negative temperature coefficient. Above 5.6 volts, the avalanche effect becomes predominant and exhibits a positive temperature coefficient. In a 5.6 V diode, the two effects occur together and their temperature coefficients neatly cancel each other out, thus the 5.6 V diode is the component of choice in temperature-critical applications. Modern manufacturing techniques have produced devices with voltages lower than 5.6 V with negligible temperature coefficients, but as higher voltage devices are encountered, the temperature coefficient rises dramatically. A 75 V diode has 10 times the coefficient of a 12 V diode.

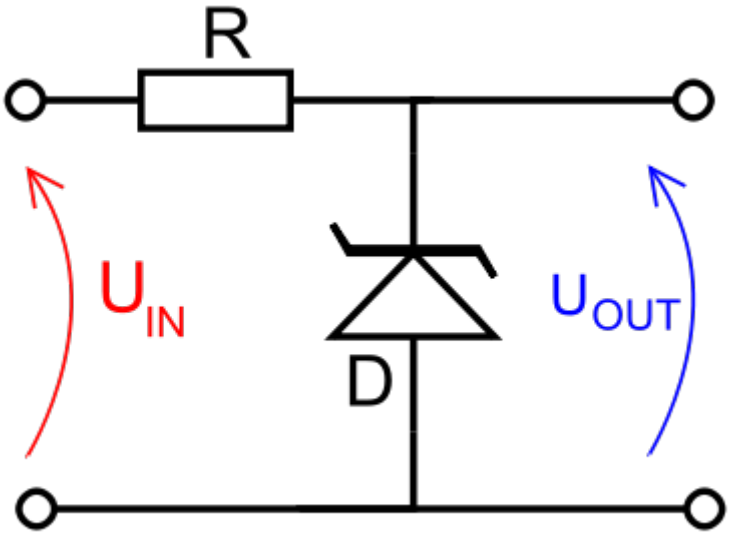
All such diodes, regardless of breakdown voltage, are usually marketed under the umbrella term of "Zener diode".

Uses



Zener diode shown with typical packages. *Reverse* current − i_Z is shown.

Zener diodes are widely used as voltage references and as shunt regulators to regulate the voltage across small circuits. When connected in parallel with a variable voltage source so that it is reverse biased, a Zener diode conducts when the voltage reaches the diode's reverse breakdown voltage. From that point on, the relatively low impedance of the diode keeps the voltage across the diode at that value.



In this circuit, a typical voltage reference or regulator, an input voltage, U_{IN}, is regulated down to a stable output voltage U_{OUT}. The intrinsic voltage drop of diode D is stable over a wide current range and holds U_{OUT} relatively constant even though the input voltage may fluctuate over a fairly wide range. Because of the low impedance of the diode when operated like this, Resistor R is used to limit current through the circuit.

In the case of this simple reference, the current flowing in the diode is determined using Ohms law and the known voltage drop across the resistor R. $I_{Diode} = (U_{IN} - U_{OUT}) / R_{\Omega}$

The value of *R* must satisfy two conditions:

1. *R* must be small enough that the current through D keeps D in reverse breakdown. The value of this current is given in the data sheet for D. For example, the common BZX79C5V6 device, a 5.6 V 0.5 W Zener diode, has a recommended reverse current of 5 mA. If insufficient current exists through D, then U_{OUT} will be unregulated, and less than the nominal breakdown voltage (this differs to voltage regulator tubes where the output voltage will be higher than nominal and could rise as high as U_{IN}). When calculating *R*, allowance must be made for any current through the external load, not shown in this diagram, connected across U_{OUT}.
2. *R* must be large enough that the current through D does not destroy the device. If the current through D is I_D, its breakdown voltage V_B and its maximum power dissipation P_{MAX}, then $I_D V_B < P_{MAX}$.

A load may be placed across the diode in this reference circuit, and as long as the zener stays in reverse breakdown, the diode will provide a stable voltage source to the load.

Shunt regulators are simple, but the requirements that the ballast resistor be small enough to avoid excessive voltage drop during worst-case operation (low input voltage concurrent with high load current) tends to leave a lot of current flowing in the diode much of the time, making for a fairly wasteful regulator with high quiescent power dissipation, only suitable for smaller loads.

Zener diodes in this configuration are often used as stable references for more advanced voltage regulator circuits.

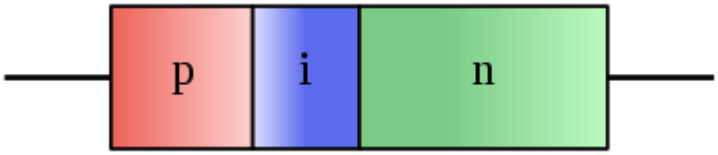
These devices are also encountered, typically in series with a base-emitter junction, in transistor stages where selective choice of a device centered around the avalanche/Zener point can be used to introduce compensating temperature co-efficient balancing of the transistor PN junction. An example of this kind of use would be a DC error amplifier used in a regulated power supply circuit feedback loop system.

Zener diodes are also used in surge protectors to limit transient voltage spikes.

Another notable application of the zener diode is the use of noise caused by its avalanche breakdown in a random number generator that never repeats.

Chapter- 5

PIN Diode



Layers of a PIN diode

A **PIN diode** is a diode with a wide, lightly doped 'near' intrinsic semiconductor region between a p-type semiconductor and an n-type semiconductor region. The p-type and n-type regions are typically heavily doped because they are used for ohmic contacts.

The wide intrinsic region is in contrast to an ordinary PN diode. The wide intrinsic region makes the PIN diode an inferior rectifier (one typical function of a diode), but it makes the PIN diode suitable for attenuators, fast switches, photodetectors, and high voltage power electronics applications.

Operation

A PIN diode operates under what is known as high-level injection. In other words, the intrinsic "i" region is flooded with charge carriers from the "p" and "n" regions. Its function can be likened to filling up a water bucket with a hole on the side. Once the water reaches the hole's level it will begin to pour out. Similarly, the diode will conduct current once the flooded electrons and holes reach an equilibrium point, where the number of electrons is equal to the number of holes in the intrinsic region. When the diode is forward biased, the injected carrier concentration is typically several orders of magnitude higher than the intrinsic level carrier concentration. Due to this high level injection, which in turn is due to the depletion process, the electric field extends deeply (almost the entire length) into the region. This electric field helps in speeding up of the transport of charge carriers from P to N region, which results in faster operation of the diode, making it a suitable device for high frequency operations.

Characteristics

A PIN diode obeys the standard diode equation for low frequency signals. At higher frequencies, the diode looks like an almost perfect (very linear, even for large signals) resistor. There is a lot of stored charge in the intrinsic region. At low frequencies, the charge can be removed and the diode turns off. At higher frequencies, there is not enough time to remove the charge, so the diode never turns off. The PIN diode has a poor reverse recovery time.

The high-frequency resistance is inversely proportional to the DC bias current through the diode. A PIN diode, suitably biased, therefore acts as a variable resistor. This high-frequency resistance may vary over a wide range (from 0.1 ohm to 10 kΩ in some cases; the useful range is smaller, though).

The wide intrinsic region also means the diode will have a low capacitance when reverse biased.

In a PIN diode, the depletion region exists almost completely within the intrinsic region. This depletion region is much larger than in a PN diode, and almost constant-size, independent of the reverse bias applied to the diode. This increases the volume where electron-hole pairs can be generated by an incident photon. Some photodetector devices, such as PIN photodiodes and phototransistors (in which the base-collector junction is a PIN diode), use a PIN junction in their construction.

The diode design has some design tradeoffs. Increasing the dimensions of the intrinsic region (and its stored charge) allows the diode to look like a resistor at lower frequencies. It adversely affects the time needed to turn off the diode and its shunt capacitance. PIN diodes will be tailored for a particular use.

Applications

PIN diodes are useful as RF switches, attenuators, and photodetectors.

RF and Microwave Switches

A PIN Diode RF Microwave Switch. Picture courtesy of Herley

Under zero or reverse bias, a PIN diode has a low capacitance. The low capacitance will not pass much of an RF signal. Under a forward bias of 1 mA, a typical PIN diode will have an RF resistance of about 1 ohm, making it a good RF conductor. Consequently, the PIN diode makes a good RF switch.

Although RF relays can be used as switches, they switch very slowly (on the order of 10 milliseconds). A PIN diode switch can switch much more quickly (e.g., 1 microsecond).

The capacitance of an off discrete PIN diode might be 1pF. At 320MHz, the reactance of 1pF is about 500 ohms. In a 50 ohm system, the off state attenuation would be about 20dB -- which may not be enough attenuation. In applications that need higher isolation, switches are cascaded to improve the isolation. Cascading three of the above switches would give 60dB of attenuation.

PIN diode switches are used not only for signal selection, but they are also used for component selection. For example, some low phase noise oscillators use PIN diodes to range switch inductors.

RF and Microwave Variable Attenuators

A RF Microwave PIN diode Attenuator. Picture courtesy of Herley

By changing the bias current through a PIN diode, it's possible to quickly change the RF resistance.

At high frequencies, the PIN diode appears as a resistor whose resistance is an inverse function of its forward current. Consequently, PIN diode can be used in some variable attenuator designs as amplitude modulators or output leveling circuits.

PIN diodes might be used, for example, as the bridge and shunt resistors in a bridged-T attenuator.

Limiters

PIN diodes are sometimes used as input protection devices for high frequency test probes. If the input signal is within range, the PIN diode has little impact as a small capacitance. If the signal is large, then the PIN diode starts to conduct and becomes a resistor that shunts most of the signal to ground.

Photodetector and photovoltaic cell

The PIN photodiode was invented by Jun-ichi Nishizawa and his colleagues in 1950.

PIN photodiodes are used in fibre optic network cards and switches. As a photodetector, the PIN diode is reverse biased. Under reverse bias, the diode ordinarily does not conduct (save a small dark current or I_s leakage). A photon entering the intrinsic region frees a carrier. The reverse bias field sweeps the carrier out of the region and creates a current. Some detectors can use avalanche multiplication.

The PIN photovoltaic cell works in the same mechanism. In this case, the advantage of using a PIN structure over conventional semiconductor junction is the better long wavelength response of the former. In case of long wavelength irradiation, photons penetrate deep into the cell. But only those electron-hole pairs generated in and near the depletion region contribute to current generation. The depletion region of a PIN structure extends across the intrinsic region, deep into the device. This wider depletion width enables electron-hole pair generation deep within the device. This increases the quantum efficiency of the cell.

Typically, amorphous silicon thin-film cells use PIN structures. On the other hand, CdTe cells use NIP structure, a variation of the PIN structure. In a NIP structure, an intrinsic CdTe layer is sandwiched by n-doped CdS and p-doped ZnTe. The photons are incident on the n-doped layer unlike a PIN diode.

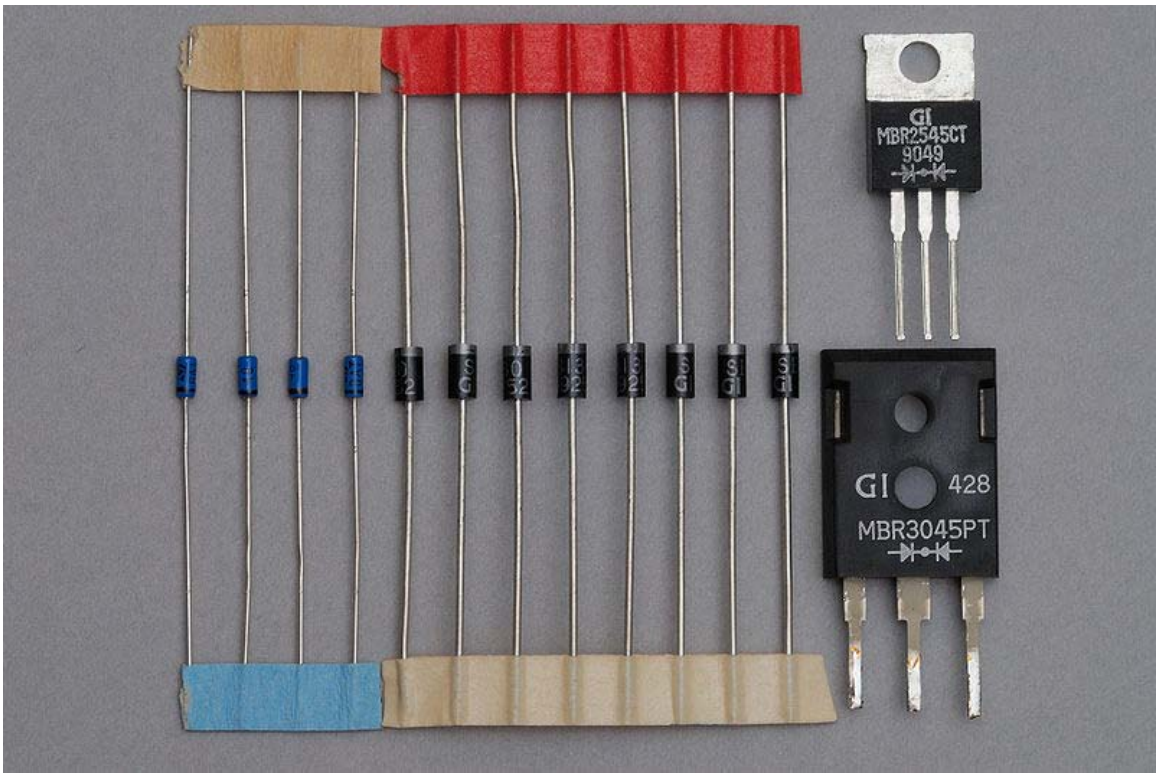
Example Diodes

SFH203 or BPW43 are cheap general purpose PIN diodes in 5 mm clear plastic case with bandwidth over 100 MHz. They are used in RONJA telecommunication systems and other circuitry applications.

Chapter- 6

Schottky Diode

Schottky diode schematic symbol



Various Schottky barrier diodes: Small signal rf devices (left), medium and high power Schottky rectifying diodes (middle and right).

The **Schottky diode** (named after German physicist Walter H. Schottky; also known as **hot carrier diode**) is a semiconductor diode with a low forward voltage drop and a very

fast switching action. The cat's-whisker detectors used in the early days of wireless can be considered as primitive Schottky diodes.

A Schottky diode is a special type of diode with a very low forward-voltage drop. When current flows through a diode there is a small voltage drop across the diode terminals. A normal silicon diode has a voltage drop between 0.6–1.7 volts, while a Schottky diode voltage drop is between approximately 0.15–0.45 volts. This lower voltage drop can provide higher switching speed and better system efficiency.

Construction

A Schottky diode uses a metal–semiconductor junction as a Schottky barrier (instead of a semiconductor–semiconductor junction as in conventional diodes). This Schottky barrier results in both very fast switching and low forward voltage drop.

Reverse recovery time

The most important difference between p-n and Schottky diode is reverse recovery time, when the diode switches from non-conducting to conducting state and vice versa. Where in a p-n diode the reverse recovery time can be in the order of hundreds of nanoseconds and less than 100 ns for fast diodes, Schottky diodes do not have a recovery time, as there is nothing to recover from. The switching time is ~100 ps for the small signal diodes, and up to tens of nanoseconds for special high-capacity power diodes. With p-n junction switching, there is also a reverse recovery current, which in high-power semiconductors brings increased EMI noise. With Schottky diodes switching essentially instantly with only slight capacitive loading, this is much less of a concern.

It is often said that the Schottky diode is a "majority carrier" semiconductor device. This means that if the semiconductor body is doped n-type, only the n-type carriers (mobile electrons) play a significant role in normal operation of the device. The majority carriers are quickly injected into the conduction band of the metal contact on the other side of the diode to become free moving electrons. Therefore no slow, random recombination of n- and p- type carriers is involved, so that this diode can cease conduction faster than an ordinary p-n rectifier diode. This property in turn allows a smaller device area, which also makes for a faster transition. This is another reason why Schottky diodes are useful in switch-mode power converters; the high speed of the diode means that the circuit can operate at frequencies in the range 200 kHz to 2 MHz, allowing the use of small inductors and capacitors with greater efficiency than would be possible with other diode types. Small-area Schottky diodes are the heart of RF detectors and mixers, which often operate up to 50 GHz.

Limitations

The most evident limitations of Schottky diodes are the relatively low reverse voltage rating for silicon-metal Schottky diodes, 50 V and below, and a relatively high reverse

leakage current. Diode designs have been improving over time. Voltage ratings now can reach 200 V. Reverse leakage current, because it increases with temperature, leads to a thermal instability issue. This often limits the useful reverse voltage to well below the actual rating.

Silicon carbide Schottky diode

Since 2001 another important invention was presented by Siemens Semiconductor (now Infineon): a silicon carbide (SiC) Schottky diode. SiC Schottky diodes have about 40 times lower reverse leakage current compared to silicon Schottky diodes and are available in 300 V and 600 V variants. As of 2007 a new 1200 volt 7.5 A variant is sold as 2x2 mm chip for power inverter manufacturers.

Silicon carbide has a high thermal conductivity and temperature has little influence on its switching and thermal characteristics. With special packaging it is possible to have operating junction temperatures of over 500 K, which allows passive radiation cooling in aerospace applications.

Applications

Voltage clamping

While standard silicon diodes have a forward voltage drop of about 0.6 volts and germanium diodes 0.3 volts, Schottky diodes' voltage drop at forward biases of around 1 mA is in the range 0.15 V to 0.46 V, which makes them useful in voltage clamping applications and prevention of transistor saturation. This is due to the higher current density in the Schottky diode.

Discharge protection

A typical application of power Schottky diodes is discharge-protection for solar cells connected to lead-acid batteries.

Power supply

They are also used as rectifiers in switched-mode power supplies; the low forward voltage and fast recovery time leads to increased efficiency.

Schottky diodes can be used in power supply "OR"ing circuits in products that have both an internal battery and a mains adapter input, or similar. However, the high reverse leakage current presents a problem in this case, as any high-impedance voltage sensing circuit (e.g. monitoring the battery voltage or detecting whether a mains adaptor is present) will see the voltage from the other power source through the diode leakage.

Designation

Commonly encountered Schottky diodes include the 1N5817 series (1 Ampere) rectifiers. Schottky metal-semiconductor junctions are featured in the successors to the 7400 TTL family of logic devices, the 74S, 74LS and 74ALS series, where they are employed as clamps in parallel with the collector-base junctions of the bipolar transistors to prevent their saturation, thereby greatly reducing their turn-off delays.

Small signal Schottky diodes like the 1N5711, 1N6263, 1SS106, 1SS108 or the BAT41–43, 45–49 series are widely used in high frequency applications as detectors, mixers and nonlinear elements, and have replaced germanium diodes, rendering them obsolete. They are also suitable for ESD protection of ESD sensitive devices like III-V-semiconductor devices, laser diodes and, to a lesser extent, exposed lines of CMOS circuitry.

Alternatives

When less power dissipation is desired a MOSFET and a control circuit can be used instead, in an operation mode known as Active rectification.

A super diode consisting of a pn-diode or Schottky diode and an operational amplifier provides an almost perfect diode characteristic due to the effect of negative feedback, although its use is restricted to frequencies the operational amplifier used can handle.

Chapter- 7

Avalanche Diode & DIAC

Avalanche diode

An **avalanche diode** is a diode (usually made from silicon, but can be made from another semiconductor) that is designed to go through avalanche breakdown at a specified reverse bias voltage and conduct as a type of voltage reference.

The Zener diode exhibits an apparently similar effect in addition to Zener breakdown. Both effects are actually present in any such diode, but one usually dominates the other. Avalanche diodes are optimized for avalanche effect so they exhibit small but significant voltage drop under breakdown conditions, unlike zener diodes that keep voltage always higher than breakdown. This feature provides better surge protection than simple zener diode and acts more like Gas discharge tube replacement.

Uses

Protection

A common application is protecting electronic circuits against damaging high voltages. The avalanche diode is connected to the circuit so that it is reverse-biased. In other words, its cathode is positive with respect to its anode. In this configuration, the diode is non-conducting and does not interfere with the circuit. If the voltage increases beyond the design limit, the diode suffers avalanche breakdown, causing the harmful voltage to be conducted to earth. When used in this fashion they are often referred to as clamper diodes or Transient voltage suppression diode because they "clamp" the maximum voltage to a predetermined level. Avalanche diodes are normally specified for this role by their clamping voltage V_{BR} and the maximum size of transient they can absorb, specified by either energy (in joules) or i^2t. Avalanche breakdown is not destructive, as long as the diode is not allowed to overheat.

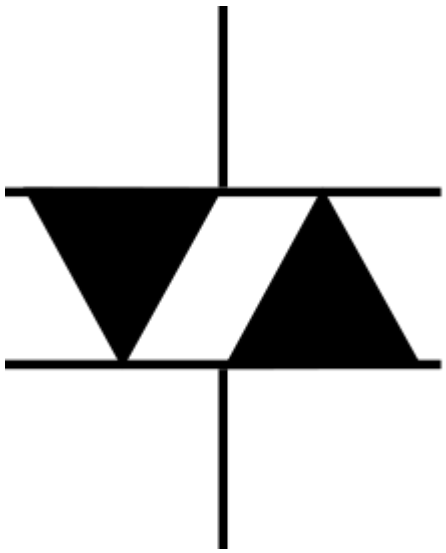
RF noise generation

Avalanche diodes generate radio frequency noise; they are commonly used as noise sources in radio equipment and hardware random number generators. For instance, they are often used as a source of RF for antenna analyzer bridges. Avalanche diodes can also be used as white noise generators.

Microwave frequency generation

If placed into a resonant circuit, avalanche diodes can act as negative resistance devices. The IMPATT diode is an avalanche diode optimized for frequency generation.

DIAC



DIAC

Three-layer DIAC

The **DIAC**, or 'diode for alternating current', is a diode that conducts current only after its breakdown voltage has been reached momentarily.

When this occurs, diode enters the region of negative dynamic resistance, leading to a decrease in the voltage drop across the diode and, usually, a sharp increase in current through the diode. The diode remains "in conduction" until the current through it drops below a value characteristic for the device, called the holding current. Below this value, the diode switches back to its high-resistance (non-conducting) state. This behavior is bidirectional, meaning typically the same for both directions of current.

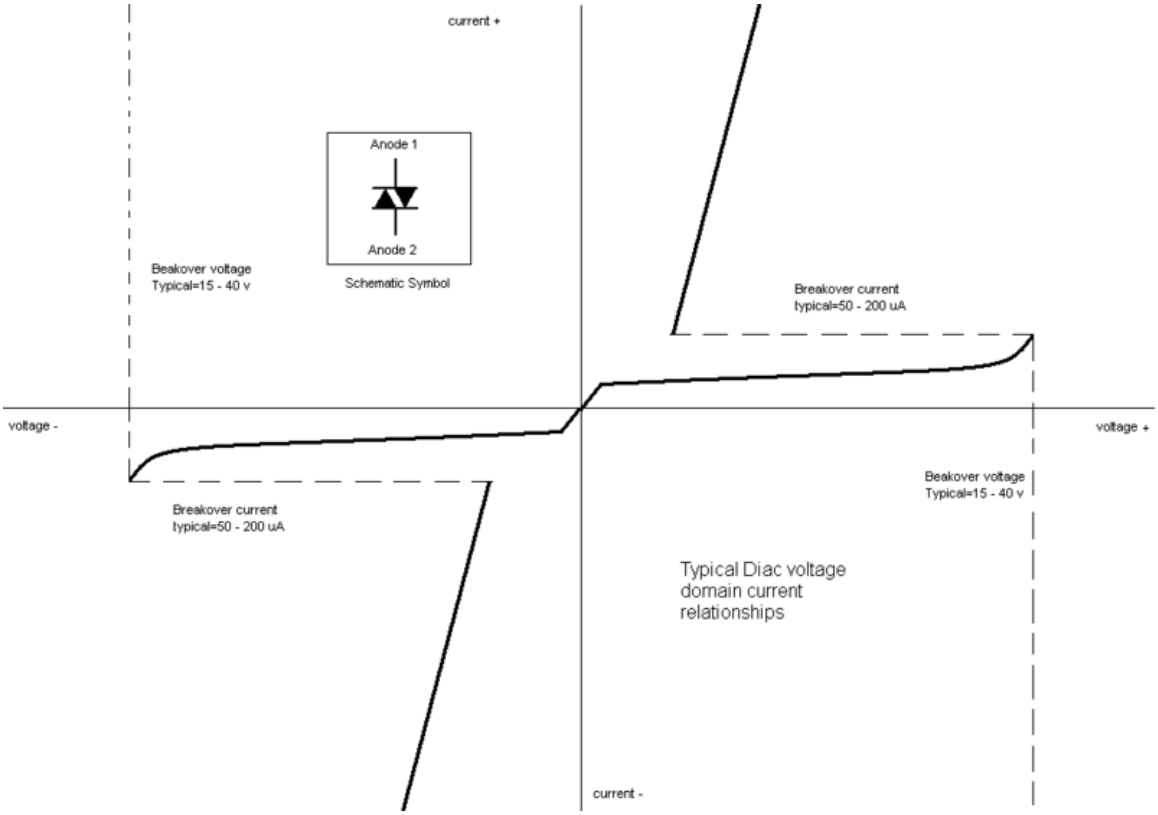
Most DIACs have a three-layer structure with breakdown voltage around 30 V. In this way, their behavior is somewhat similar to (but much more precisely controlled and taking place at lower voltages than) a neon lamp.

DIACs have no gate electrode, unlike some other thyristors that they are commonly used to trigger, such as TRIACs. Some TRIACs contain a built-in DIAC in series with the TRIAC's "gate" terminal for this purpose.

DIACs are also called *symmetrical trigger diodes* due to the symmetry of their characteristic curve. Because DIACs are bidirectional devices, their terminals are not labeled as *anode* and *cathode* but as A1 and A2 or MT1 ("Main Terminal") and MT2.

SIDAC

SIDAC



Idealized breakover diode voltage and current relationships. Once the voltage exceeds the turn-on threshold, the device turns on and the voltage rapidly falls while the current increases.

The **SIDAC** is a less common electrically similar device, the difference in naming being determined by the manufacturer. In general, SIDACs have higher breakover voltages and current handling.

The SIDAC, or *Silicon Diode for Alternating Current*, is another member of the thyristor family. Also referred to as a SYDAC (Silicon thYristor for Alternating Current), bi-directional thyristor breakover diode, or more simply a bi-directional thyristor diode, it is technically specified as a bilateral voltage triggered switch. Its operation is similar to that of the DIAC, but SIDAC is always a five-layer device with low-voltage drop in latched conducting state, more like a voltage triggered TRIAC without a gate. In general, SIDACs have higher breakover voltages and current handling capacities than DIACs, so

they can be directly used for switching and not just for triggering of another switching device.

The operation of the SIDAC is functionally similar to that of a spark gap. The SIDAC remains nonconducting until the applied voltage meets or exceeds its rated breakover voltage. Once entering this conductive state going through the negative dynamic resistance region, the SIDAC continues to conduct, regardless of voltage, until the applied current falls below its rated holding current. At this point, the SIDAC returns to its initial nonconductive state to begin the cycle once again.

Somewhat uncommon in most electronics, the SIDAC is relegated to the status of a special purpose device. However, where part-counts are to be kept low, simple relaxation oscillators are needed, and when the voltages are too low for practical operation of a spark gap, the SIDAC is an indispensable component.

Similar devices, though usually not functionally interchangeable with SIDACs, are the Thyristor Surge Protection Devices (TSPD), Trisil, SIDACtor, or the now-obsolete Surgector. These are designed to tolerate large surge currents for the suppression of overvoltage transients.

Chapter- 8

Capacitor

Capacitor

Modern capacitors, by a cm rule

Type Passive

Invented Ewald Georg von Kleist (October 1745)

Electronic symbol

A **capacitor** (formerly known as **condenser**) 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.

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 typical electrolytic capacitor

The effect 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.

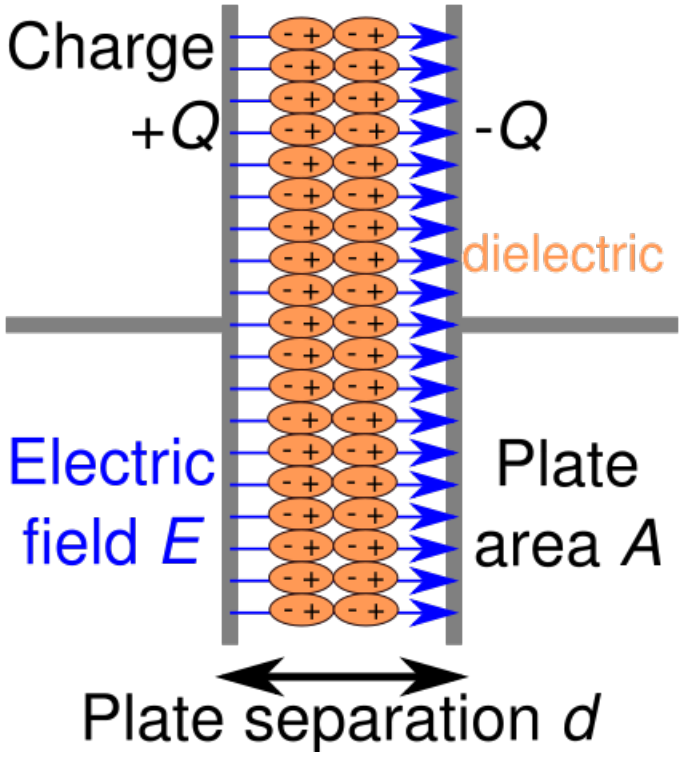
Daniel Gralath was the first to combine several jars in parallel into a "battery" to increase the charge storage capacity. Benjamin Franklin investigated the Leyden jar and "proved"

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.

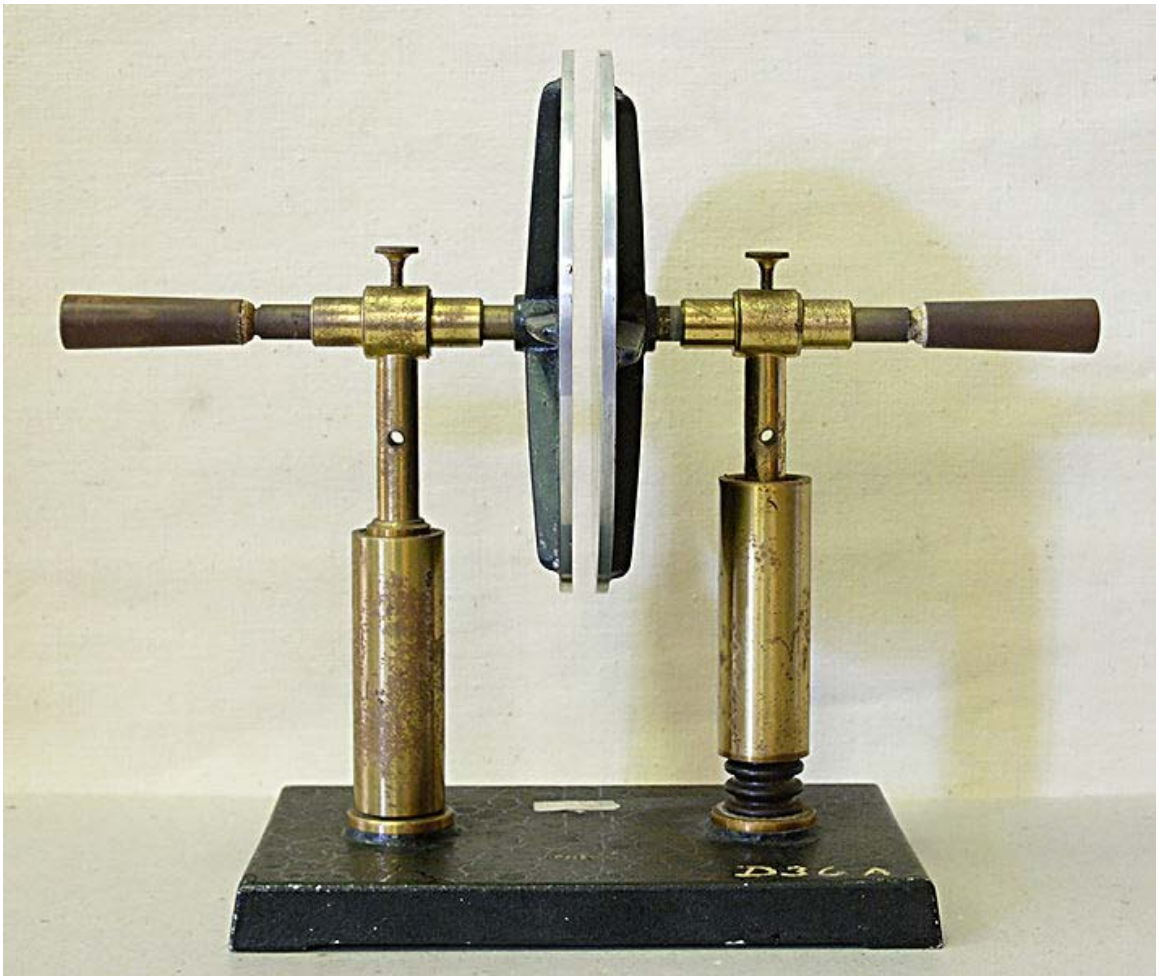
Leyden jars or more powerful devices employing flat glass plates alternating with foil conductors were used exclusively up until about 1900, when the invention of wireless (radio) created a demand for standard capacitors, and the steady move to higher frequencies required capacitors with lower inductance.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

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

.

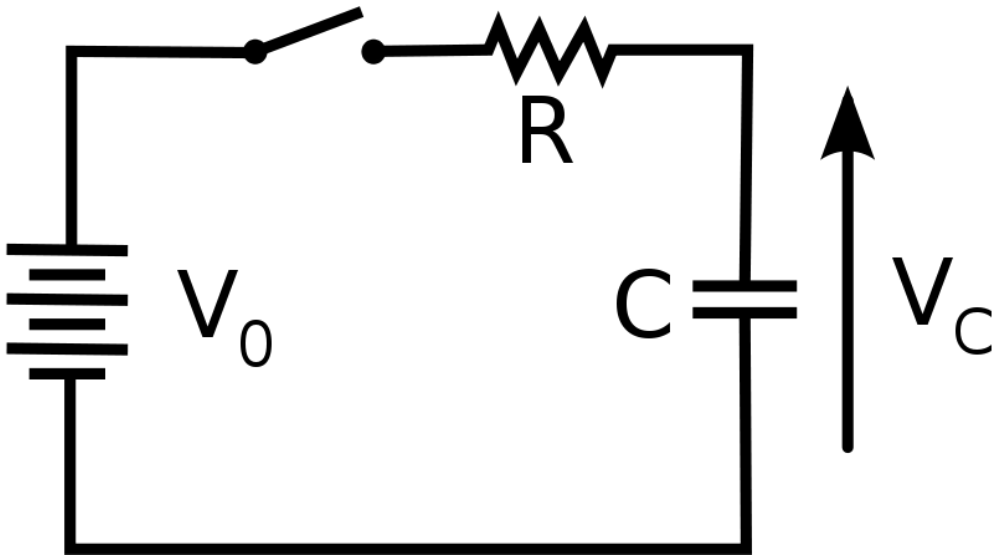
Taking the derivative of this, and multiplying by C, yields the derivative form,

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

.

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

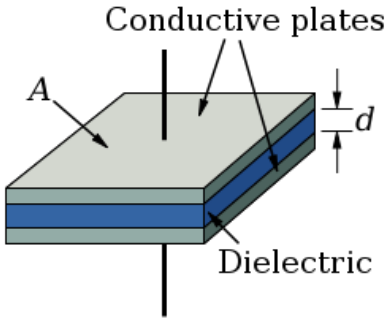
$$X = -\frac{1}{\omega C} = -\frac{1}{2\pi f C}$$
$$Z = \frac{1}{j\omega C} = -\frac{j}{\omega C} = -\frac{j}{2\pi f C}$$

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.

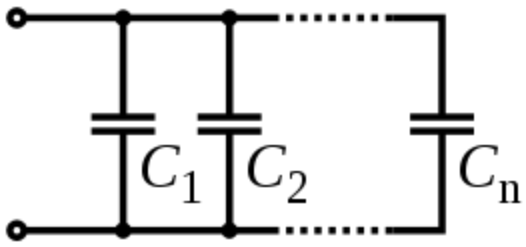
The simplest capacitor consists of two parallel conductive plates separated by a dielectric with permittivity ε (such as air). The model may also be used to make qualitative predictions for other device geometries. The plates are considered to extend uniformly over an area A and a charge density $\pm\rho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

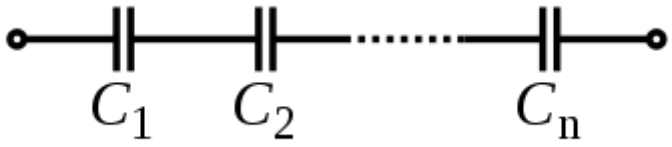
Networks

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,

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

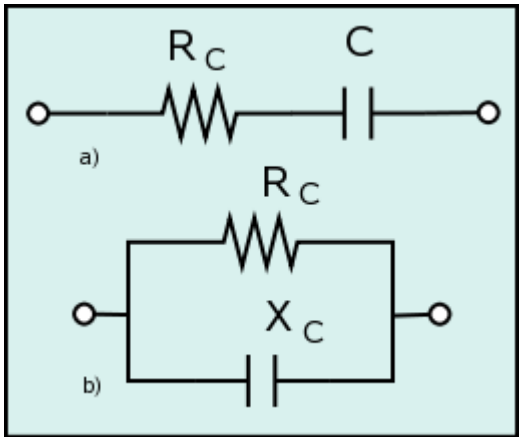
The maximum energy that can be stored safely in a capacitor is limited by the breakdown voltage. Due to the scaling of capacitance and breakdown voltage with dielectric

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.

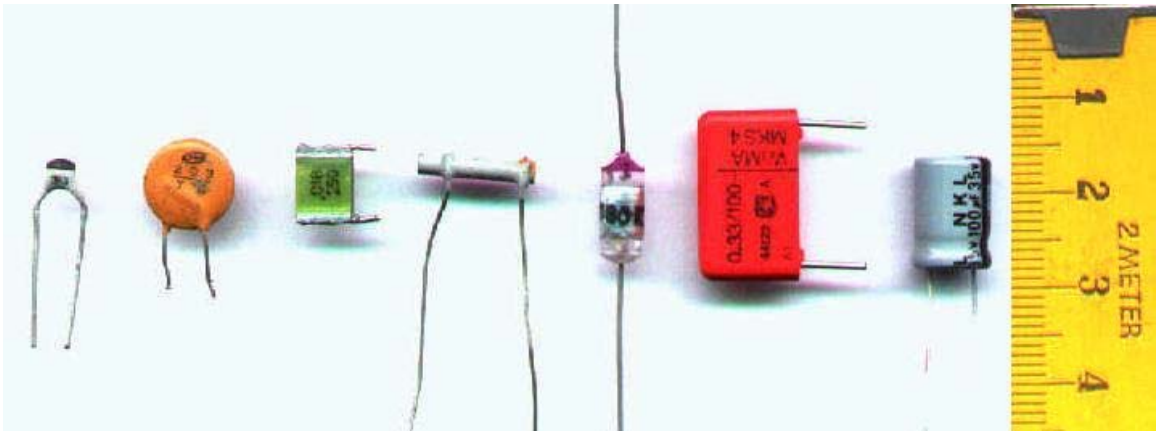
Capacitor types

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.

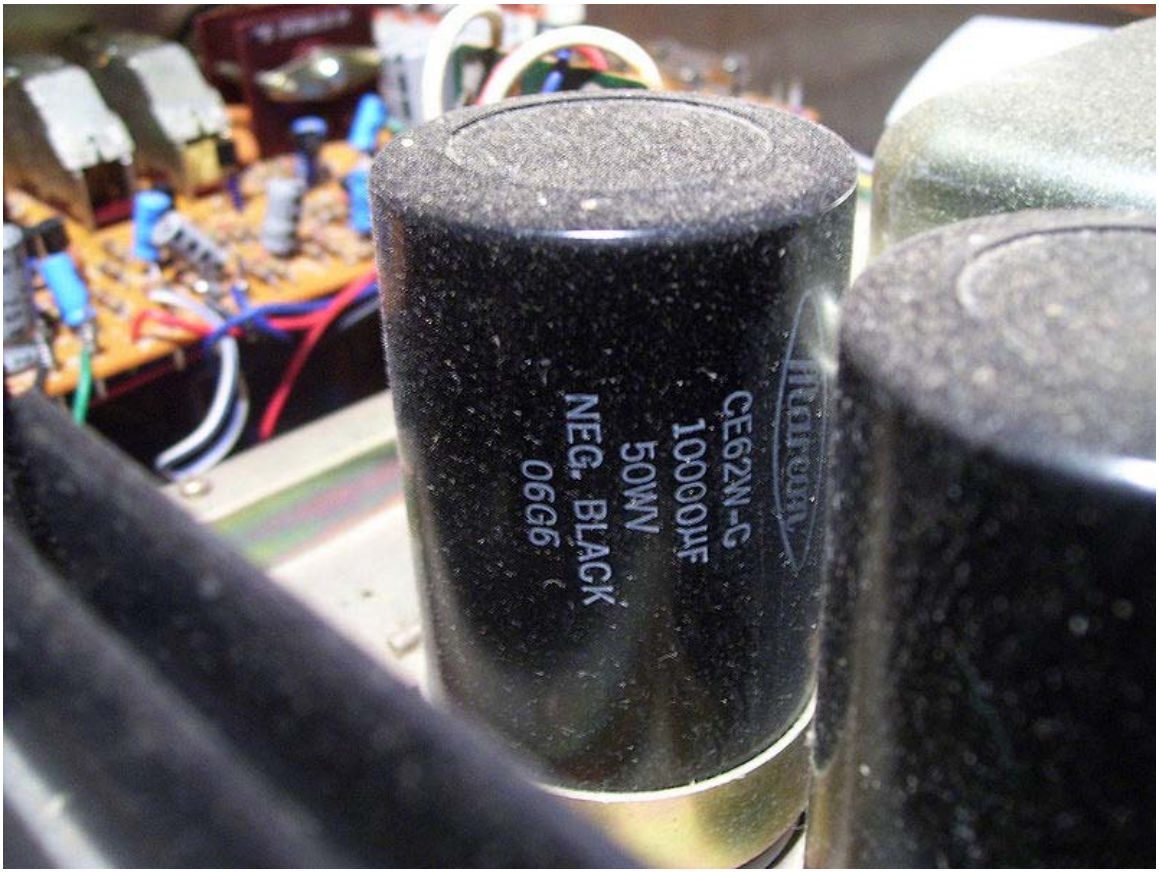
In car audio systems, large capacitors store energy for the amplifier to use on demand. Also for a flash tube a capacitor is used to hold the high voltage.

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.

High-voltage 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 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.

Chapter- 9

Transformer

Pole-mounted power distribution transformer with center-tapped secondary winding (note use of grounded conductor, right, as one leg of the primary feeder). It transforms the high voltage of the overhead distribution wires to the lower voltage used in house wiring.

A **transformer** is a static device that transfers electrical energy from one circuit to another through inductively coupled conductors—the transformer's coils. A varying current in the first or *primary* winding creates a varying magnetic flux in the

transformer's core and thus a varying magnetic field through the *secondary* winding. This varying magnetic field induces a varying electromotive force (EMF) or "voltage" in the secondary winding. This effect is called mutual induction.

If a load is connected to the secondary, an electric current will flow in the secondary winding and electrical energy will be transferred from the primary circuit through the transformer to the load. In an ideal transformer, the induced voltage in the secondary winding (V_s) is in proportion to the primary voltage (V_p), and is given by the ratio of the number of turns in the secondary (N_s) to the number of turns in the primary (N_p) as follows:

$$\frac{V_\mathrm{s}}{V_\mathrm{p}} = \frac{N_\mathrm{s}}{N_\mathrm{p}}$$

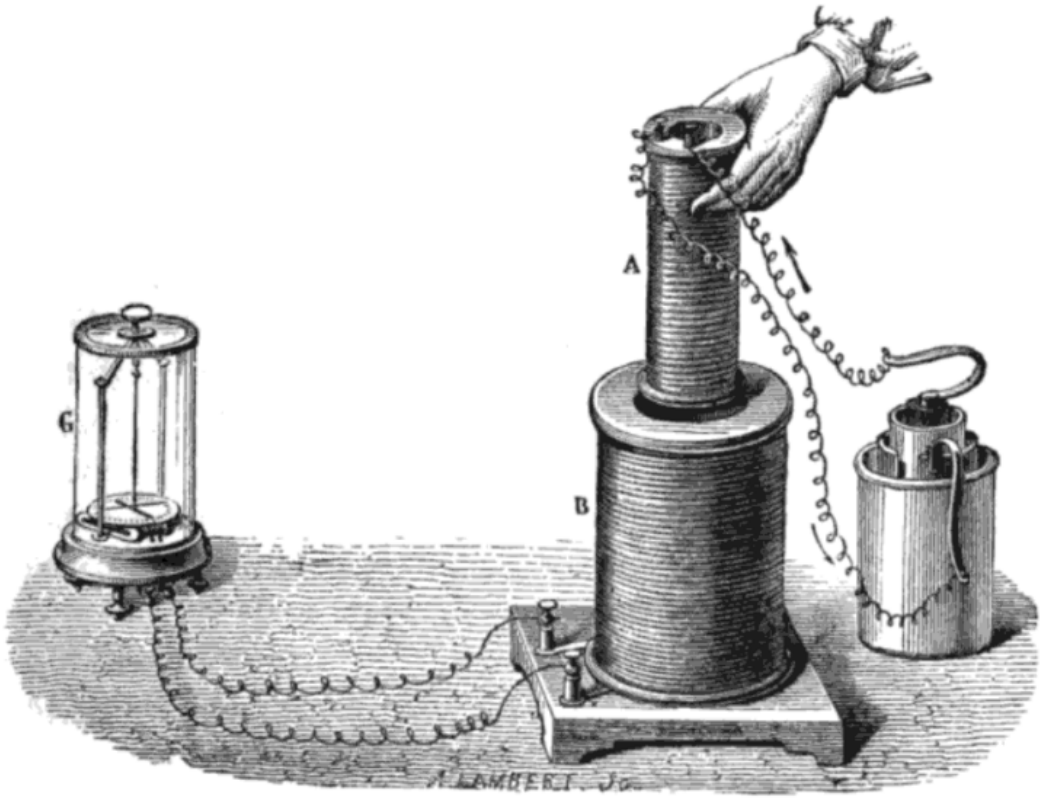
By appropriate selection of the ratio of turns, a transformer thus allows an alternating current (AC) voltage to be "stepped up" by making N_s greater than N_p, or "stepped down" by making N_s less than N_p.

In the vast majority of transformers, the windings are coils wound around a ferromagnetic core, air-core transformers being a notable exception.

Transformers range in size from a thumbnail-sized coupling transformer hidden inside a stage microphone to huge units weighing hundreds of tons used to interconnect portions of power grids. All operate with the same basic principles, although the range of designs is wide. While new technologies have eliminated the need for transformers in some electronic circuits, transformers are still found in nearly all electronic devices designed for household ("mains") voltage. Transformers are essential for high-voltage electric power transmission, which makes long-distance transmission economically practical.

History

Discovery



Faraday's experiment with induction between coils of wire

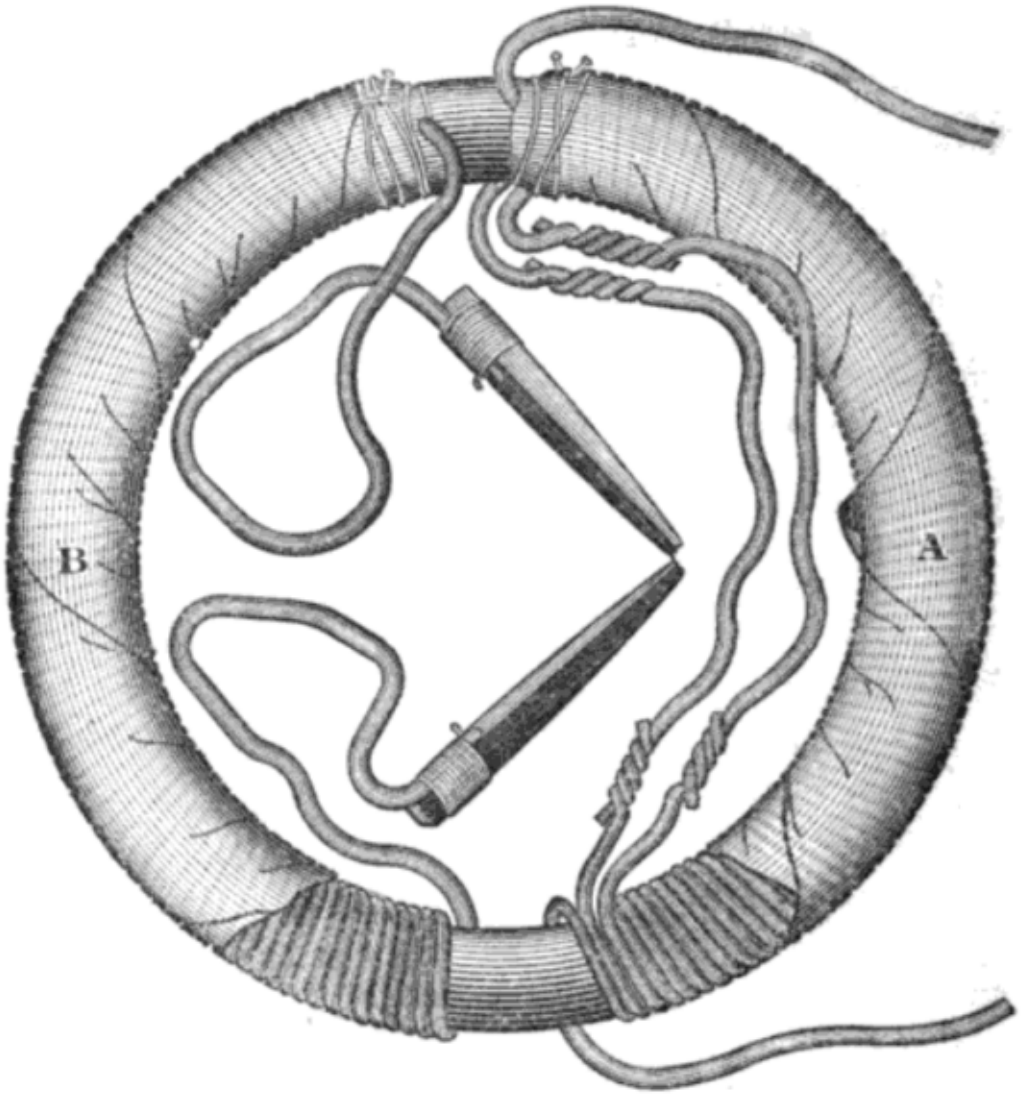
The phenomenon of electromagnetic induction was discovered independently by Michael Faraday and Joseph Henry in 1831. However, Faraday was the first to publish the results of his experiments and thus receive credit for the discovery. The relationship between electromotive force (EMF) or "voltage" and magnetic flux was formalized in an equation now referred to as "Faraday's law of induction":

$$|\mathcal{E}| = \left|\frac{d\Phi_B}{dt}\right|$$

where $|\mathcal{E}|$ is the magnitude of the EMF in volts and Φ_B is the magnetic flux through the circuit (in webers).

Faraday performed the first experiments on induction between coils of wire, including winding a pair of coils around an iron ring, thus creating the first toroidal closed-core transformer.

Induction coils



Faraday's ring transformer

The first type of transformer to see wide use was the induction coil, invented by Rev. Nicholas Callan of Maynooth College, Ireland in 1836. He was one of the first researchers to realize that the more turns the secondary winding has in relation to the primary winding, the larger is the increase in EMF. Induction coils evolved from scientists' and inventors' efforts to get higher voltages from batteries. Since batteries produce direct current (DC) rather than alternating current (AC), induction coils relied

upon vibrating electrical contacts that regularly interrupted the current in the primary to create the flux changes necessary for induction. Between the 1830s and the 1870s, efforts to build better induction coils, mostly by trial and error, slowly revealed the basic principles of transformers.

In 1876, Russian engineer Pavel Yablochkov invented a lighting system based on a set of induction coils where the primary windings were connected to a source of alternating current and the secondary windings could be connected to several "electric candles" (arc lamps) of his own design. The coils Yablochkov employed functioned essentially as transformers.

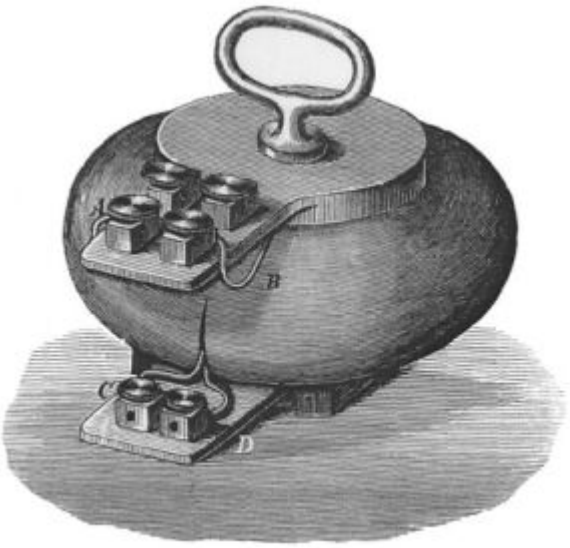
In 1878, the Ganz Company in Hungary began manufacturing equipment for electric lighting and, by 1883, had installed over fifty systems in Austria-Hungary. Their systems used alternating current exclusively and included those comprising both arc and incandescent lamps, along with generators and other equipment.

Lucien Gaulard and John Dixon Gibbs first exhibited a device with an open iron core called a "secondary generator" in London in 1882, then sold the idea to the Westinghouse company in the United States. They also exhibited the invention in Turin, Italy in 1884, where it was adopted for an electric lighting system. However, the efficiency of their open-core bipolar apparatus remained very low.

Induction coils with open magnetic circuits are inefficient for transfer of power to loads. Until about 1880, the paradigm for AC power transmission from a high voltage supply to a low voltage load was a series circuit. Open-core transformers with a ratio near 1:1 were connected with their primaries in series to allow use of a high voltage for transmission while presenting a low voltage to the lamps. The inherent flaw in this method was that turning off a single lamp affected the voltage supplied to all others on the same circuit. Many adjustable transformer designs were introduced to compensate for this problematic characteristic of the series circuit, including those employing methods of adjusting the core or bypassing the magnetic flux around part of a coil.

Efficient, practical transformer designs did not appear until the 1880s, but within a decade the transformer would be instrumental in the "War of Currents", and in seeing AC distribution systems triumph over their DC counterparts, a position in which they have remained dominant ever since.

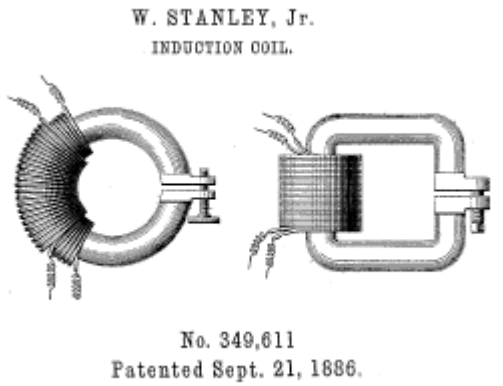
Closed-core lighting transformers



Drawing of Ganz Company's 1885 prototype. Capacity: 1400 VA, frequency: 40 Hz, voltage ratio: 120/72 V

Prototypes of the world's first high-efficiency transformers. They were built by the Z.B.D. team on 16th September 1884.

In the autumn of 1884, Ganz Company engineers Károly Zipernowsky, Ottó Bláthy and Miksa Déri had determined that open-core devices were impracticable, as they were incapable of reliably regulating voltage. In their joint patent application for the "Z.B.D." transformers, they described two designs with closed magnetic circuits: the "closed-core" and "shell-core" transformers. In the closed-core, the primary and secondary windings were wound around a closed iron ring; in the shell-core, the windings were passed *through* the iron core. In both designs, the magnetic flux linking the primary and secondary windings traveled almost entirely within the iron core, with no intentional path through air. The new Z.B.D. transformers reached 98 percent efficiency, which was 3.4 times higher than the open core bipolar devices of Gaulard and Gibs. When they employed it in parallel connected electric distribution systems, closed-core transformers finally made it technically and economically feasible to provide electric power for lighting in homes, businesses and public spaces. Bláthy had suggested the use of closed-cores, Zipernowsky the use of shunt connections, and Déri had performed the experiments; Bláthy also discovered the transformer formula, Vs/Vp = Ns/Np. The vast majority of transformers in use today rely on the basic principles discovered by the three engineers. They also reportedly popularized the word "transformer" to describe a device for altering the EMF of an electric current, although the term had already been in use by 1882. In 1886, the Ganz Company installed the world's first power station that used AC generators to power a parallel-connected common electrical network, the steam-powered Rome-Cerchi power plant.


W. STANLEY, Jr.
INDUCTION COIL.

No. 349,611
Patented Sept. 21, 1886.

Stanley's 1886 design for adjustable gap open-core induction coils

Although George Westinghouse had bought Gaulard and Gibbs' patents in 1885, the Edison Electric Light Company held an option on the U.S. rights for the Z.B.D. transformers, requiring Westinghouse to pursue alternative designs on the same principles. He assigned to William Stanley the task of developing a device for commercial use in United States. Stanley's first patented design was for induction coils with single cores of soft iron and adjustable gaps to regulate the EMF present in the secondary winding. This design was first used commercially in the U.S. in 1886. But Westinghouse soon had his team working on a design whose core comprised a stack of thin "E-shaped" iron plates, separated individually or in pairs by thin sheets of paper or other insulating material. Prewound copper coils could then be slid into place, and

straight iron plates laid in to create a closed magnetic circuit. Westinghouse applied for a patent for the new design in December 1886; it was granted in July 1887.

Other early transformers

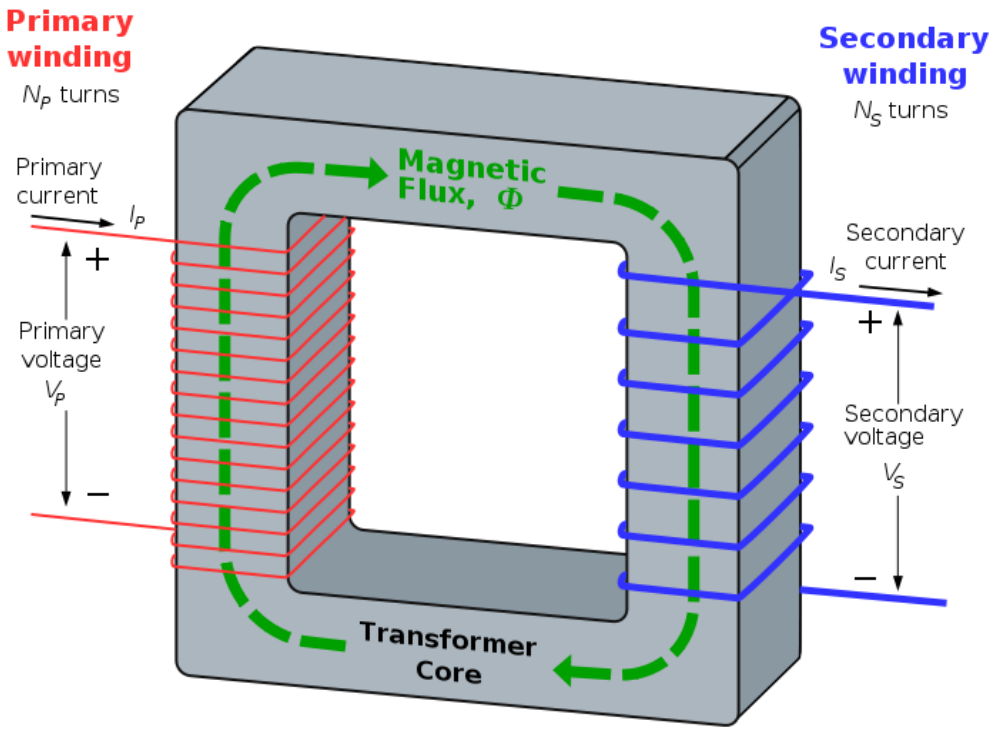
In 1889, Russian-born engineer Mikhail Dolivo-Dobrovolsky developed the first three-phase transformer at the Allgemeine Elektricitäts-Gesellschaft ("General Electricity Company") in Germany.

In 1891, Nikola Tesla invented the Tesla coil, an air-cored, dual-tuned resonant transformer for generating very high voltages at high frequency.

Audio frequency transformers ("repeating coils") were used by early experimenters in the development of the telephone.

Basic principles

The transformer is based on two principles: first, that an electric current can produce a magnetic field (electromagnetism), and, second that a changing magnetic field within a coil of wire induces a voltage across the ends of the coil (electromagnetic induction). Changing the current in the primary coil changes the magnetic flux that is developed. The changing magnetic flux induces a voltage in the secondary coil.



An ideal transformer

An ideal transformer is shown in the adjacent figure. Current passing through the primary coil creates a magnetic field. The primary and secondary coils are wrapped around a core of very high magnetic permeability, such as iron, so that most of the magnetic flux passes through both the primary and secondary coils.

Induction law

The voltage induced across the secondary coil may be calculated from Faraday's law of induction, which states that:

$$V_\mathrm{s} = N_\mathrm{s}\frac{\mathrm{d}\Phi}{\mathrm{d}t},$$

where V_s is the instantaneous voltage, N_s is the number of turns in the secondary coil and Φ is the magnetic flux through one turn of the coil. If the turns of the coil are oriented perpendicular to the magnetic field lines, the flux is the product of the magnetic flux density B and the area A through which it cuts. The area is constant, being equal to the cross-sectional area of the transformer core, whereas the magnetic field varies with time according to the excitation of the primary. Since the same magnetic flux passes through both the primary and secondary coils in an ideal transformer, the instantaneous voltage across the primary winding equals

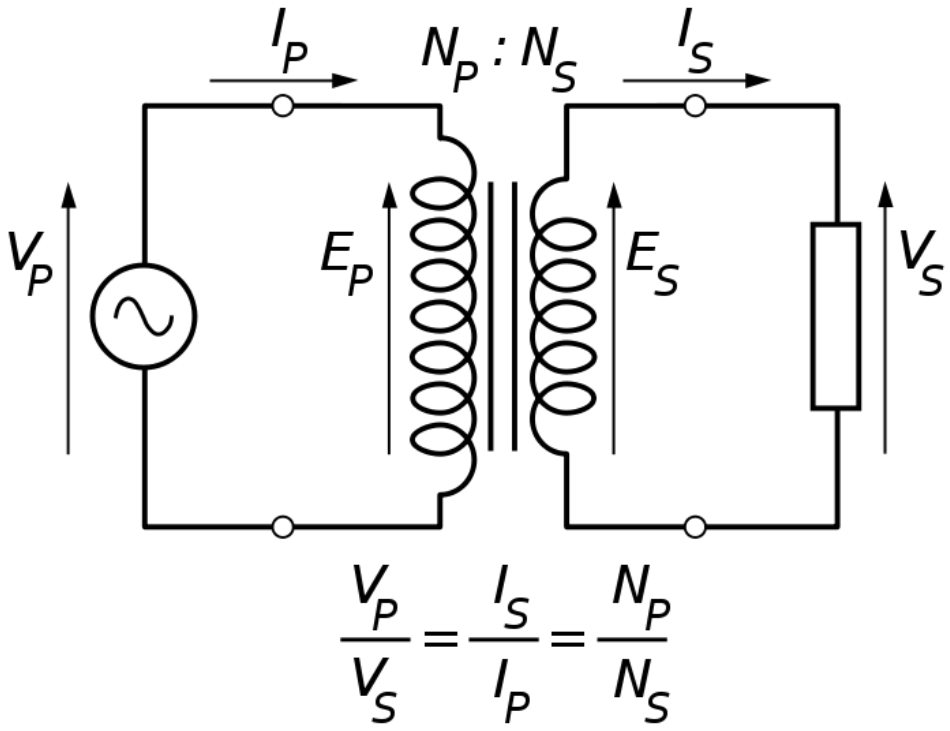
$$V_\mathrm{p} = N_\mathrm{p}\frac{\mathrm{d}\Phi}{\mathrm{d}t}.$$

Taking the ratio of the two equations for V_s and V_p gives the basic equation for stepping up or stepping down the voltage

$$\frac{V_\mathrm{s}}{V_\mathrm{p}} = \frac{N_\mathrm{s}}{N_\mathrm{p}}.$$

$N_\mathrm{p}/N_\mathrm{s}$ is known as the *turns ratio*, and is the primary functional characteristic of any transformer. In the case of step-up transformers, this may sometimes be stated as the reciprocal, $N_\mathrm{s}/N_\mathrm{p}$. *Turns ratio* is commonly expressed as an irreducible fraction or ratio: for example, a transformer with primary and secondary windings of, respectively, 100 and 150 turns is said to have a turns ratio of 2:3 rather than 0.667 or 100:150.

Ideal power equation



The ideal transformer as a circuit element

If the secondary coil is attached to a load that allows current to flow, electrical power is transmitted from the primary circuit to the secondary circuit. Ideally, the transformer is perfectly efficient; all the incoming energy is transformed from the primary circuit to the magnetic field and into the secondary circuit. If this condition is met, the incoming electric power must equal the outgoing power:

$$P_{\text{incoming}} = I_{\text{p}} V_{\text{p}} = P_{\text{outgoing}} = I_{\text{s}} V_{\text{s}},$$

giving the ideal transformer equation

$$\frac{V_{\text{s}}}{V_{\text{p}}} = \frac{N_{\text{s}}}{N_{\text{p}}} = \frac{I_{\text{p}}}{I_{\text{s}}}.$$

Transformers normally have high efficiency, so this formula is a reasonable approximation.

If the voltage is increased, then the current is decreased by the same factor. The impedance in one circuit is transformed by the *square* of the turns ratio. For example, if

an impedance Z_s is attached across the terminals of the secondary coil, it appears to the primary circuit to have an impedance of $(N_p/N_s)^2 Z_s$. This relationship is reciprocal, so that the impedance Z_p of the primary circuit appears to the secondary to be $(N_s/N_p)^2 Z_p$.

Detailed operation

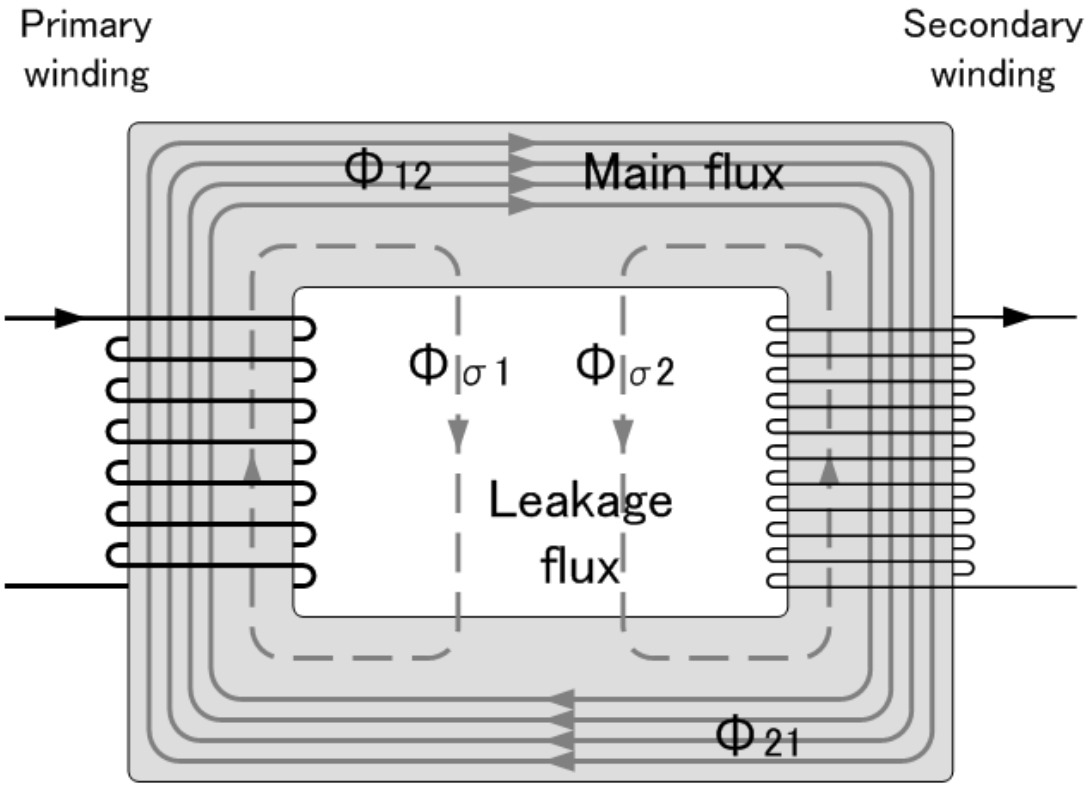
The simplified description above neglects several practical factors, in particular the primary current required to establish a magnetic field in the core, and the contribution to the field due to current in the secondary circuit.

Models of an ideal transformer typically assume a core of negligible reluctance with two windings of zero resistance. When a voltage is applied to the primary winding, a small current flows, driving flux around the magnetic circuit of the core. The current required to create the flux is termed the *magnetizing current*; since the ideal core has been assumed to have near-zero reluctance, the magnetizing current is negligible, although still required to create the magnetic field.

The changing magnetic field induces an electromotive force (EMF) across each winding. Since the ideal windings have no impedance, they have no associated voltage drop, and so the voltages V_P and V_S measured at the terminals of the transformer, are equal to the corresponding EMFs. The primary EMF, acting as it does in opposition to the primary voltage, is sometimes termed the "back EMF". This is due to Lenz's law which states that the induction of EMF would always be such that it will oppose development of any such change in magnetic field.

Practical considerations

Leakage flux



Leakage flux of a transformer

The ideal transformer model assumes that all flux generated by the primary winding links all the turns of every winding, including itself. In practice, some flux traverses paths that take it outside the windings. Such flux is termed *leakage flux*, and results in leakage inductance in series with the mutually coupled transformer windings. Leakage results in energy being alternately stored in and discharged from the magnetic fields with each cycle of the power supply. It is not directly a power loss, but results in inferior voltage regulation, causing the secondary voltage to fail to be directly proportional to the primary, particularly under heavy load. Transformers are therefore normally designed to have very low leakage inductance.

However, in some applications, leakage can be a desirable property, and long magnetic paths, air gaps, or magnetic bypass shunts may be deliberately introduced to a transformer's design to limit the short-circuit current it will supply. Leaky transformers may be used to supply loads that exhibit negative resistance, such as electric arcs, mercury vapor lamps, and neon signs; or for safely handling loads that become periodically short-circuited such as electric arc welders.

Air gaps are also used to keep a transformer from saturating, especially audio-frequency transformers in circuits that have a direct current flowing through the windings.

Leakage inductance is also helpful when transformers are operated in parallel. It can be shown that if the "per-unit" inductance of two transformers is the same (a typical value is 5%), they will automatically split power "correctly" (e.g. 500 kVA unit in parallel with 1,000 kVA unit, the larger one will carry twice the current).

Effect of frequency

Transformer universal EMF equation

If the flux in the core is purely sinusoidal, the relationship for either winding between its **rms voltage** E_{rms} of the winding , and the supply frequency f, number of turns N, core cross-sectional area a and peak magnetic flux density B is given by the universal EMF equation:

$$E_{\mathrm{rms}} = \frac{2\pi f N a B_{\mathrm{peak}}}{\sqrt{2}} \approx 4.44 f N a B$$

If the flux does not contain even harmonics the following equation can be used for **half-cycle average voltage** E_{avg} of any waveshape:

$$E_{\mathrm{avg}} = 4 f N a B_{\mathrm{peak}}$$

The time-derivative term in Faraday's Law shows that the flux in the core is the integral with respect to time of the applied voltage. Hypothetically an ideal transformer would work with direct-current excitation, with the core flux increasing linearly with time. In practice, the flux would rise to the point where magnetic saturation of the core occurs, causing a huge increase in the magnetizing current and overheating the transformer. All practical transformers must therefore operate with alternating (or pulsed) current.

The EMF of a transformer at a given flux density increases with frequency. By operating at higher frequencies, transformers can be physically more compact because a given core is able to transfer more power without reaching saturation and fewer turns are needed to achieve the same impedance. However, properties such as core loss and conductor skin effect also increase with frequency. Aircraft and military equipment employ 400 Hz power supplies which reduce core and winding weight. Conversely, frequencies used for some railway electrification systems were much lower (e.g. 16.7 Hz and 25 Hz) than normal utility frequencies (50 – 60 Hz) for historical reasons concerned mainly with the limitations of early electric traction motors. As such, the transformers used to step down the high over-head line voltages (e.g. 15 kV) are much heavier for the same power rating than those designed only for the higher frequencies.

Operation of a transformer at its designed voltage but at a higher frequency than intended will lead to reduced magnetizing current; at lower frequency, the magnetizing current

will increase. Operation of a transformer at other than its design frequency may require assessment of voltages, losses, and cooling to establish if safe operation is practical. For example, transformers may need to be equipped with "volts per hertz" over-excitation relays to protect the transformer from overvoltage at higher than rated frequency.

One example of state-of-the-art design is those transformers used for electric multiple unit high speed trains, particularly those required to operate across the borders of countries using different standards of electrification. The position of such transformers is restricted to being hung below the passenger compartment. They have to function at different frequencies (down to 16.7 Hz) and voltages (up to 25 kV) whilst handling the enhanced power requirements needed for operating the trains at high speed.

Knowledge of natural frequencies of transformer windings is of importance for the determination of the transient response of the windings to impulse and switching surge voltages.

Energy losses

An ideal transformer would have no energy losses, and would be 100% efficient. In practical transformers energy is dissipated in the windings, core, and surrounding structures. Larger transformers are generally more efficient, and those rated for electricity distribution usually perform better than 98%.

Experimental transformers using superconducting windings achieve efficiencies of 99.85%. The increase in efficiency can save considerable energy, and hence money, in a large heavily-loaded transformer; the trade-off is in the additional initial and running cost of the superconducting design.

Losses in transformers (excluding associated circuitry) vary with load current, and may be expressed as "no-load" or "full-load" loss. Winding resistance dominates load losses, whereas hysteresis and eddy currents losses contribute to over 99% of the no-load loss. The no-load loss can be significant, so that even an idle transformer constitutes a drain on the electrical supply and a running cost; designing transformers for lower loss requires a larger core, good-quality silicon steel, or even amorphous steel, for the core, and thicker wire, increasing initial cost, so that there is a trade-off between initial cost and running cost.

Transformer losses are divided into losses in the windings, termed copper loss, and those in the magnetic circuit, termed iron loss. Losses in the transformer arise from:

Winding resistance

Current flowing through the windings causes resistive heating of the conductors. At higher frequencies, skin effect and proximity effect create additional winding resistance and losses.

Hysteresis losses

Each time the magnetic field is reversed, a small amount of energy is lost due to hysteresis within the core. For a given core material, the loss is proportional to the frequency, and is a function of the peak flux density to which it is subjected.

Eddy currents

Ferromagnetic materials are also good conductors, and a core made from such a material also constitutes a single short-circuited turn throughout its entire length. Eddy currents therefore circulate within the core in a plane normal to the flux, and are responsible for resistive heating of the core material. The eddy current loss is a complex function of the square of supply frequency and inverse square of the material thickness. Eddy current losses can be reduced by making the core of a stack of plates electrically insulated from each other, rather than a solid block; all transformers operating at low frequencies use laminated or similar cores.

Magnetostriction

Magnetic flux in a ferromagnetic material, such as the core, causes it to physically expand and contract slightly with each cycle of the magnetic field, an effect known as magnetostriction. This produces the buzzing sound commonly associated with transformers, and can cause losses due to frictional heating.

Mechanical losses

In addition to magnetostriction, the alternating magnetic field causes fluctuating forces between the primary and secondary windings. These incite vibrations within nearby metalwork, adding to the buzzing noise, and consuming a small amount of power.

Stray losses

Leakage inductance is by itself largely lossless, since energy supplied to its magnetic fields is returned to the supply with the next half-cycle. However, any leakage flux that intercepts nearby conductive materials such as the transformer's support structure will give rise to eddy currents and be converted to heat. There are also radiative losses due to the oscillating magnetic field, but these are usually small.

Dot convention

It is common in transformer schematic symbols for there to be a dot at the end of each coil within a transformer, particularly for transformers with multiple primary and secondary windings. The dots indicate the direction of each winding relative to the others. Voltages at the dot end of each winding are in phase; current flowing into the dot end of a primary coil will result in current flowing out of the dot end of a secondary coil.

Equivalent circuit

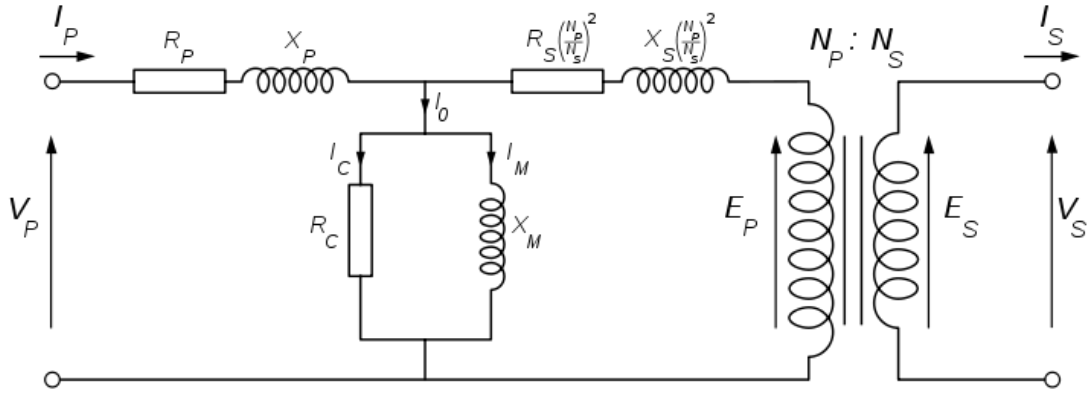
The physical limitations of the practical transformer may be brought together as an equivalent circuit model (shown below) built around an ideal lossless transformer. Power loss in the windings is current-dependent and is represented as in-series resistances R_p and R_s. Flux leakage results in a fraction of the applied voltage dropped without

contributing to the mutual coupling, and thus can be modeled as reactances of each leakage inductance X_p and X_s in series with the perfectly coupled region.

Iron losses are caused mostly by hysteresis and eddy current effects in the core, and are proportional to the square of the core flux for operation at a given frequency. Since the core flux is proportional to the applied voltage, the iron loss can be represented by a resistance R_C in parallel with the ideal transformer.

A core with finite permeability requires a magnetizing current I_m to maintain the mutual flux in the core. The magnetizing current is in phase with the flux; saturation effects cause the relationship between the two to be non-linear, but for simplicity this effect tends to be ignored in most circuit equivalents. With a sinusoidal supply, the core flux lags the induced EMF by 90° and this effect can be modeled as a magnetizing reactance (reactance of an effective inductance) X_m in parallel with the core loss component. R_c and X_m are sometimes together termed the *magnetizing branch* of the model. If the secondary winding is made open-circuit, the current I_0 taken by the magnetizing branch represents the transformer's no-load current.

The secondary impedance R_s and X_s is frequently moved (or "referred") to the primary side after multiplying the components by the impedance scaling factor $(N_p/N_s)^2$.



Transformer equivalent circuit, with secondary impedances referred to the primary side

The resulting model is sometimes termed the "exact equivalent circuit", though it retains a number of approximations, such as an assumption of linearity. Analysis may be simplified by moving the magnetizing branch to the left of the primary impedance, an implicit assumption that the magnetizing current is low, and then summing primary and referred secondary impedances, resulting in so-called equivalent impedance.

The parameters of equivalent circuit of a transformer can be calculated from the results of two transformer tests: open-circuit test and short-circuit test.

Types

A wide variety of transformer designs are used for different applications, though they share several common features. Important common transformer types include:

Autotransformer

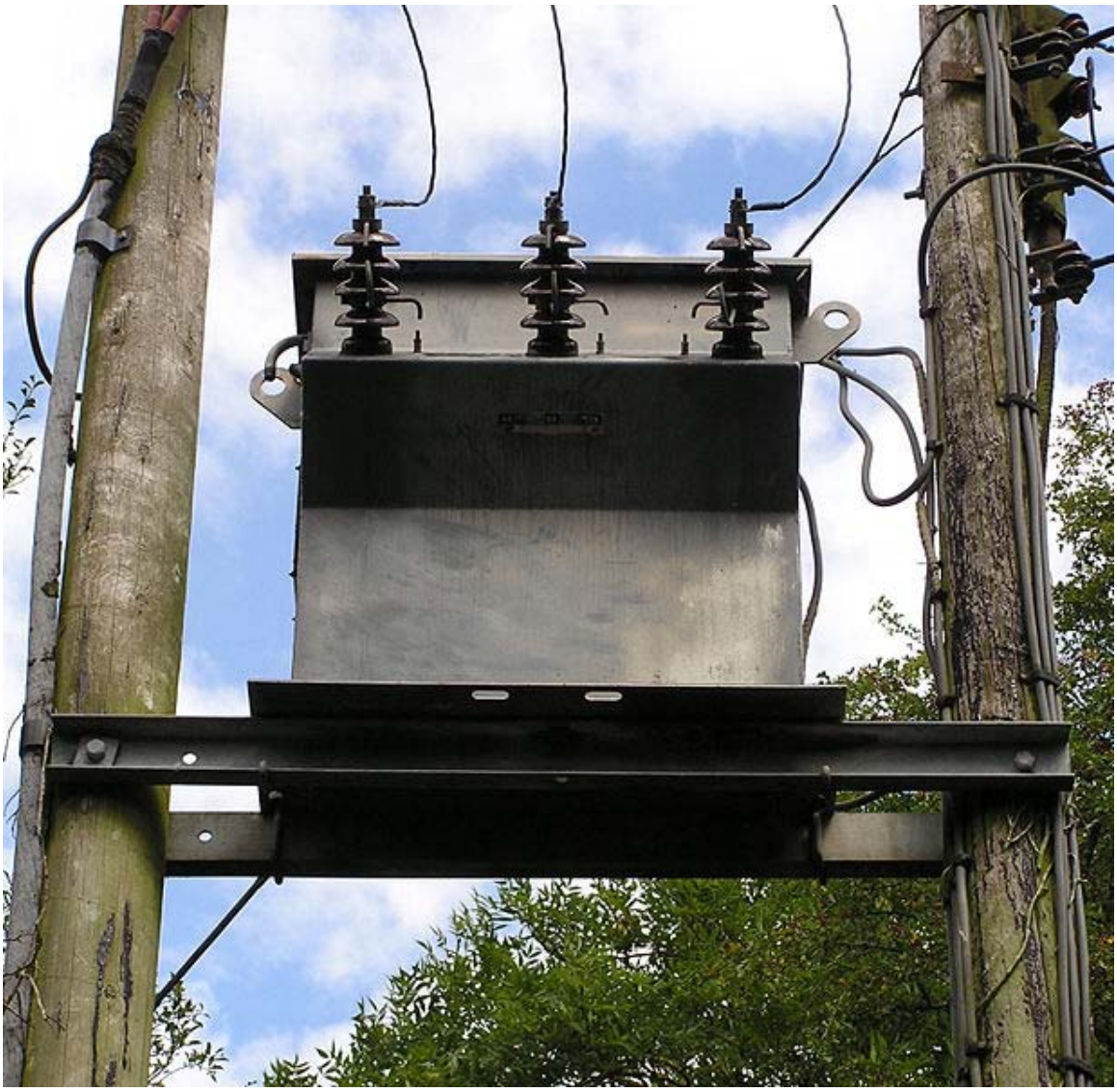
A variable autotransformer

In an autotransformer portions of the same winding act as both the primary and secondary. The winding has at least three taps where electrical connections are made. An autotransformer can be smaller, lighter and cheaper than a standard dual-winding transformer however the autotransformer does not provide electrical isolation.

Autotransformers are often used to step up or down between voltages in the 110-117-120 volt range and voltages in the 220-230-240 volt range, e.g., to output either 110 or 120V (with taps) from 230V input, allowing equipment from a 100 or 120V region to be used in a 230V region.

A variable autotransformer is made by exposing part of the winding coils and making the secondary connection through a sliding brush, giving a variable turns ratio. Such a device is often referred to by the trademark name *variac*.

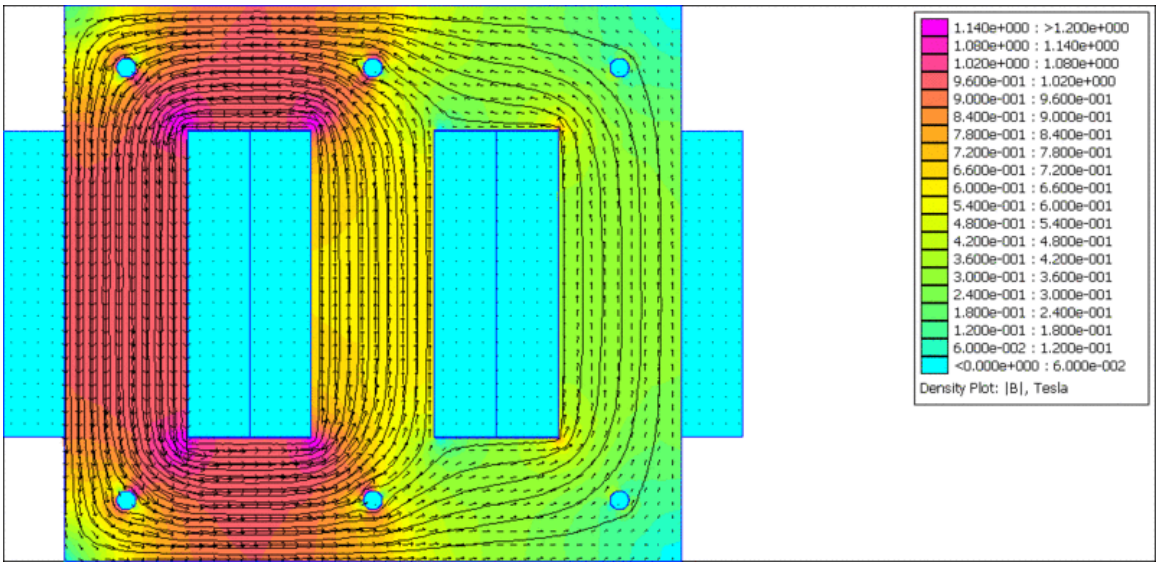
Polyphase transformers



Three-phase step-down transformer mounted between two utility poles

For three-phase supplies, a bank of three individual single-phase transformers can be used, or all three phases can be incorporated as a single three-phase transformer. In this

case, the magnetic circuits are connected together, the core thus containing a three-phase flow of flux. A number of winding configurations are possible, giving rise to different attributes and phase shifts. One particular polyphase configuration is the zigzag transformer, used for grounding and in the suppression of harmonic currents.



Screenshot of a FEM simulation of the magnetic flux inside a three-phase power transformer.

Leakage transformers

Leakage transformer

A leakage transformer, also called a stray-field transformer, has a significantly higher leakage inductance than other transformers, sometimes increased by a magnetic bypass or shunt in its core between primary and secondary, which is sometimes adjustable with a set screw. This provides a transformer with an inherent current limitation due to the loose coupling between its primary and the secondary windings. The output and input currents are low enough to prevent thermal overload under all load conditions—even if the secondary is shorted.

Leakage transformers are used for arc welding and high voltage discharge lamps (neon lights and cold cathode fluorescent lamps, which are series-connected up to 7.5 kV AC). It acts then both as a voltage transformer and as a magnetic ballast.

Other applications are short-circuit-proof extra-low voltage transformers for toys or doorbell installations.

Resonant transformers

A resonant transformer is a kind of leakage transformer. It uses the leakage inductance of its secondary windings in combination with external capacitors, to create one or more resonant circuits. Resonant transformers such as the Tesla coil can generate very high voltages, and are able to provide much higher current than electrostatic high-voltage generation machines such as the Van de Graaff generator. One of the applications of the resonant transformer is for the CCFL inverter. Another application of the resonant transformer is to couple between stages of a superheterodyne receiver, where the selectivity of the receiver is provided by tuned transformers in the intermediate-frequency amplifiers.

Audio transformers

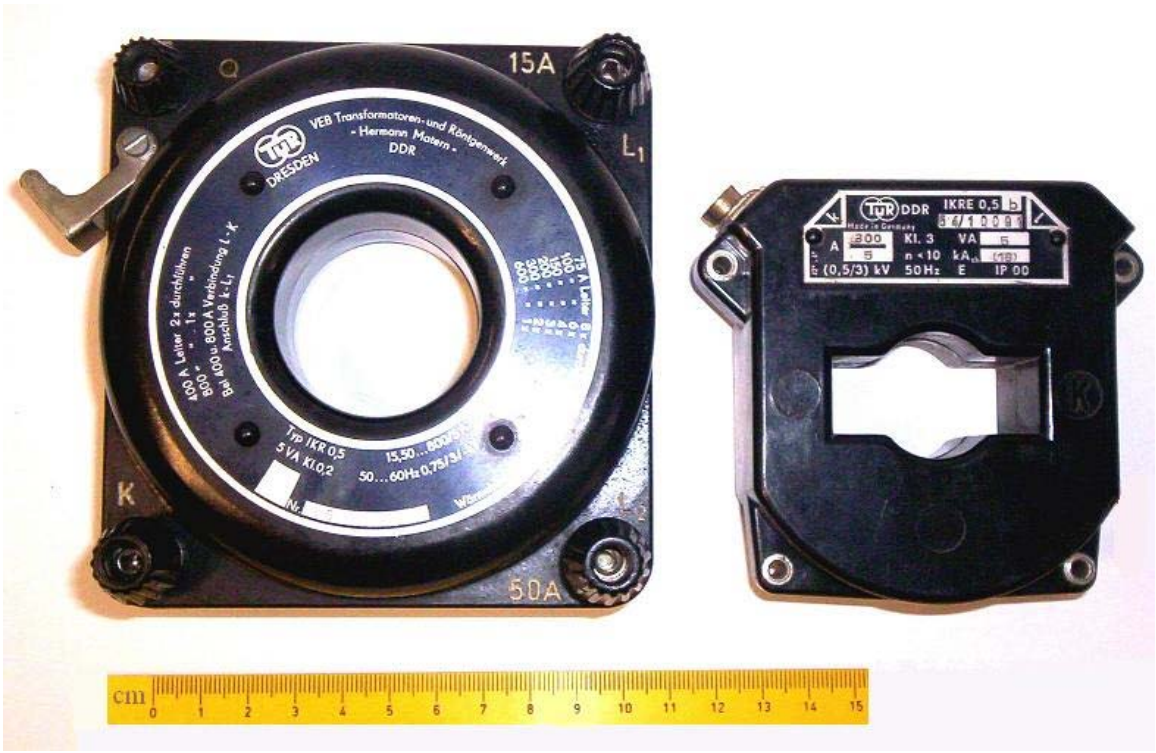
Audio transformers are those specifically designed for use in audio circuits. They can be used to block radio frequency interference or the DC component of an audio signal, to split or combine audio signals, or to provide impedance matching between high and low impedance circuits, such as between a high impedance tube (valve) amplifier output and a low impedance loudspeaker, or between a high impedance instrument output and the low impedance input of a mixing console.

Such transformers were originally designed to connect different telephone systems to one another while keeping their respective power supplies isolated, and are still commonly used to interconnect professional audio systems or system components.

Being magnetic devices, audio transformers are susceptible to external magnetic fields such as those generated by AC current-carrying conductors. "Hum" is a term commonly used to describe unwanted signals originating from the "mains" power supply (typically 50 or 60 Hz). Audio transformers used for low-level signals, such as those from microphones, often include shielding to protect against extraneous magnetically coupled signals.

Instrument transformers

Instrument transformers are used for measuring voltage and current in electrical power systems, and for power system protection and control. Where a voltage or current is too large to be conveniently used by an instrument, it can be scaled down to a standardized, low value. Instrument transformers isolate measurement, protection and control circuitry from the high currents or voltages present on the circuits being measured or controlled.



Current transformers, designed for placing around conductors

A current transformer is a transformer designed to provide a current in its secondary coil proportional to the current flowing in its primary coil.

Voltage transformers (VTs), also referred to as "potential transformers" (PTs), are designed to have an accurately known transformation ratio in both magnitude and phase, over a range of measuring circuit impedances. A voltage transformer is intended to present a negligible load to the supply being measured. The low secondary voltage allows protective relay equipment and measuring instruments to be operated at a lower voltages.

Both current and voltage instrument transformers are designed to have predictable characteristics on overloads. Proper operation of over-current protective relays requires that current transformers provide a predictable transformation ratio even during a short-circuit.

Classification

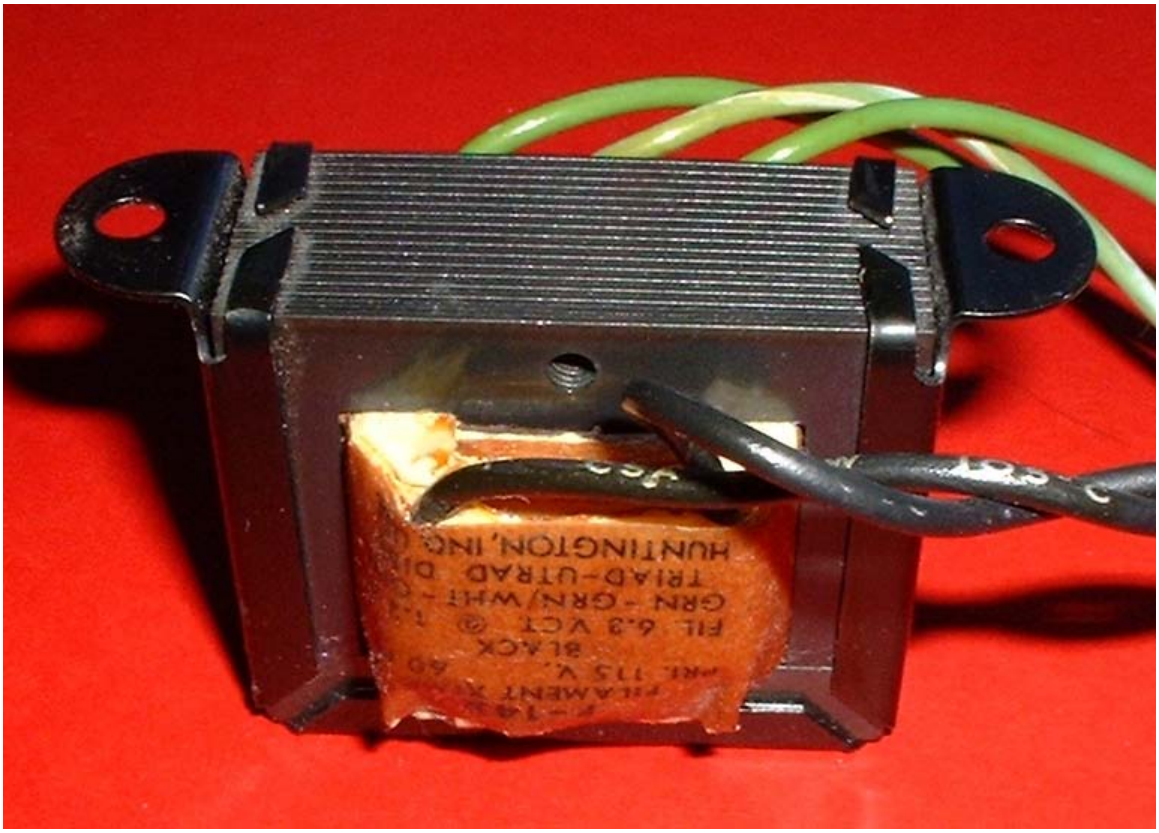
Transformers can be classified in many different ways; an incomplete list is:

- *By power capacity*: from a fraction of a volt-ampere (VA) to over a thousand MVA;
- *By frequency range*: power-, audio-, or radio frequency;
- *By voltage class*: from a few volts to hundreds of kilovolts;
- *By cooling type*: air-cooled, oil-filled, fan-cooled, or water-cooled;

- *By application*: such as power supply, impedance matching, output voltage and current stabilizer, or circuit isolation;
- *By purpose*: distribution, rectifier, arc furnace, amplifier output, etc.;
- *By winding turns ratio*: step-up, step-down, isolating with equal or near-equal ratio, variable, multiple windings.

Construction

Cores



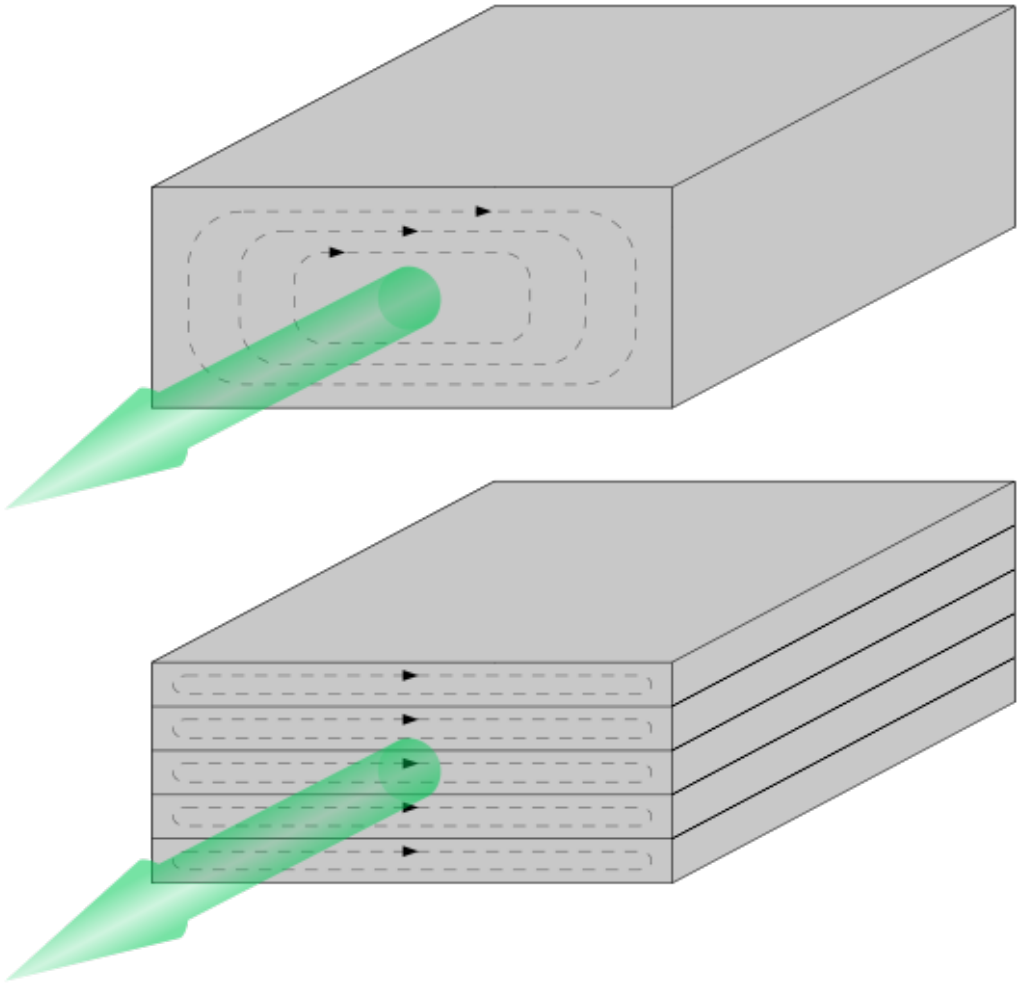
Laminated core transformer showing edge of laminations at top of photo

Laminated steel cores

Transformers for use at power or audio frequencies typically have cores made of high permeability silicon steel. The steel has a permeability many times that of free space, and the core thus serves to greatly reduce the magnetizing current, and confine the flux to a path which closely couples the windings. Early transformer developers soon realized that cores constructed from solid ion resulted in prohibitive eddy-current losses, and their designs mitigated this effect with cores consisting of bundles of insulated iron wires. Later designs constructed the core by stacking layers of thin steel laminations, a principle that has remained in use. Each lamination is insulated from its neighbors by a thin non-

conducting layer of insulation. The universal transformer equation indicates a minimum cross-sectional area for the core to avoid saturation.

The effect of laminations is to confine eddy currents to highly elliptical paths that enclose little flux, and so reduce their magnitude. Thinner laminations reduce losses, but are more laborious and expensive to construct. Thin laminations are generally used on high frequency transformers, with some types of very thin steel laminations able to operate up to 10 kHz.



Laminating the core greatly reduces eddy-current losses

One common design of laminated core is made from interleaved stacks of E-shaped steel sheets capped with I-shaped pieces, leading to its name of "E-I transformer". Such a design tends to exhibit more losses, but is very economical to manufacture. The cut-core or C-core type is made by winding a steel strip around a rectangular form and then bonding the layers together. It is then cut in two, forming two C shapes, and the core assembled by binding the two C halves together with a steel strap. They have the

advantage that the flux is always oriented parallel to the metal grains, reducing reluctance.

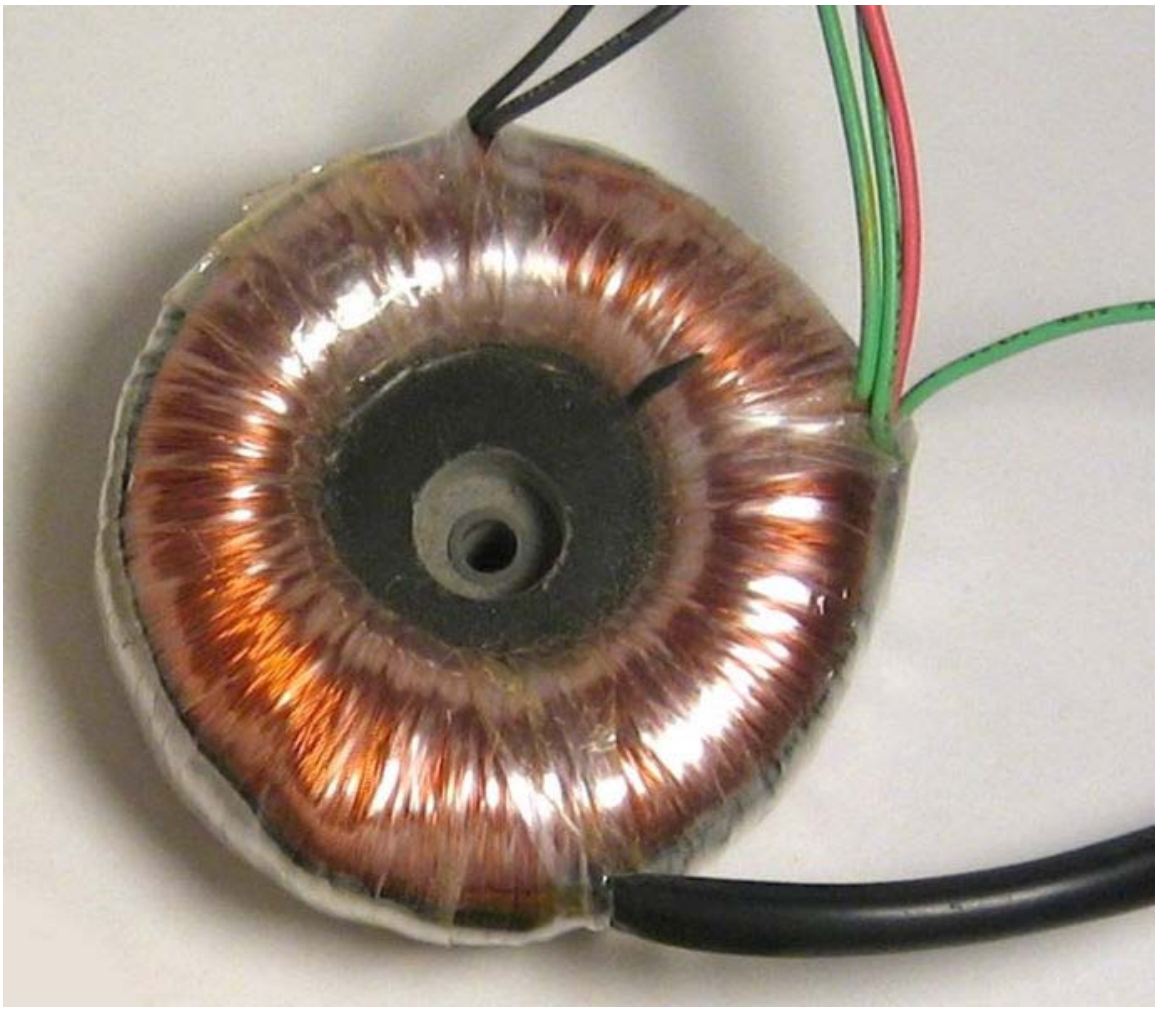
A steel core's remanence means that it retains a static magnetic field when power is removed. When power is then reapplied, the residual field will cause a high inrush current until the effect of the remaining magnetism is reduced, usually after a few cycles of the applied alternating current. Overcurrent protection devices such as fuses must be selected to allow this harmless inrush to pass. On transformers connected to long, overhead power transmission lines, induced currents due to geomagnetic disturbances during solar storms can cause saturation of the core and operation of transformer protection devices.

Distribution transformers can achieve low no-load losses by using cores made with low-loss high-permeability silicon steel or amorphous (non-crystalline) metal alloy. The higher initial cost of the core material is offset over the life of the transformer by its lower losses at light load.

Solid cores

Powdered iron cores are used in circuits (such as switch-mode power supplies) that operate above main frequencies and up to a few tens of kilohertz. These materials combine high magnetic permeability with high bulk electrical resistivity. For frequencies extending beyond the VHF band, cores made from non-conductive magnetic ceramic materials called ferrites are common. Some radio-frequency transformers also have movable cores (sometimes called 'slugs') which allow adjustment of the coupling coefficient (and bandwidth) of tuned radio-frequency circuits.

Toroidal cores



Small toroidal core transformer

Toroidal transformers are built around a ring-shaped core, which, depending on operating frequency, is made from a long strip of silicon steel or permalloy wound into a coil, powdered iron, or ferrite. A strip construction ensures that the grain boundaries are optimally aligned, improving the transformer's efficiency by reducing the core's reluctance. The closed ring shape eliminates air gaps inherent in the construction of an E-I core. The cross-section of the ring is usually square or rectangular, but more expensive cores with circular cross-sections are also available. The primary and secondary coils are often wound concentrically to cover the entire surface of the core. This minimizes the length of wire needed, and also provides screening to minimize the core's magnetic field from generating electromagnetic interference.

Toroidal transformers are more efficient than the cheaper laminated E-I types for a similar power level. Other advantages compared to E-I types, include smaller size (about half), lower weight (about half), less mechanical hum (making them superior in audio

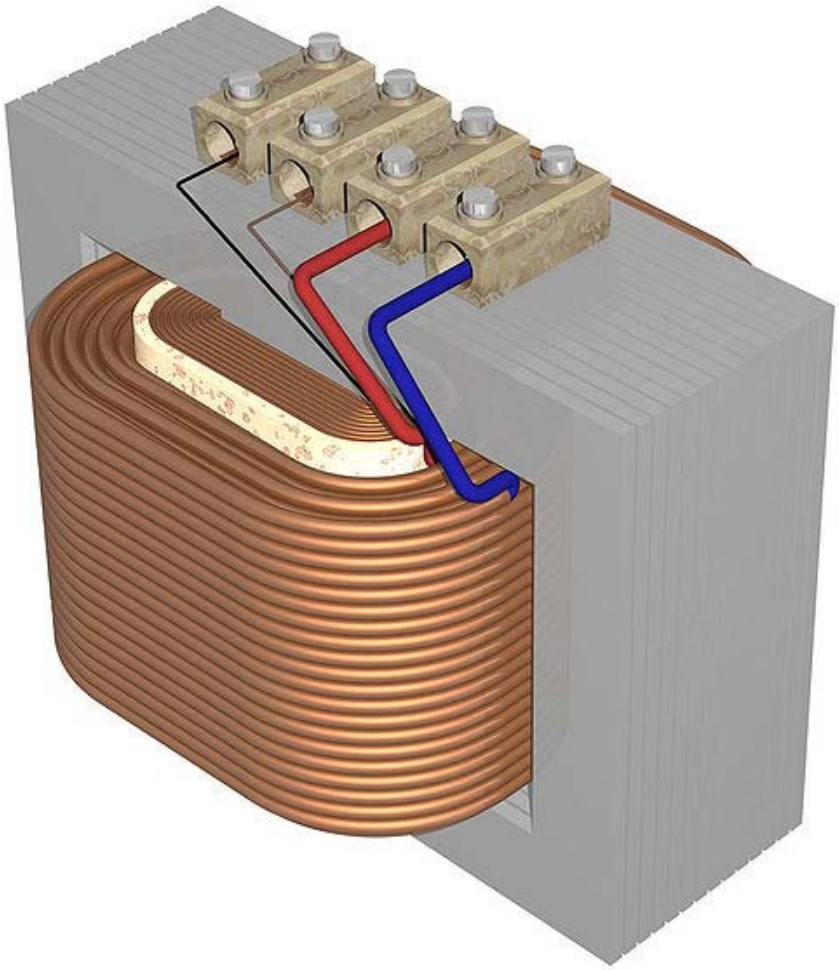
amplifiers), lower exterior magnetic field (about one tenth), low off-load losses (making them more efficient in standby circuits), single-bolt mounting, and greater choice of shapes. The main disadvantages are higher cost and limited power capacity. Because of the lack of a residual gap in the magnetic path, toroidal transformers also tend to exhibit higher inrush current, compared to laminated E-I types.

Ferrite toroidal cores are used at higher frequencies, typically between a few tens of kilohertz to hundreds of megahertz, to reduce losses, physical size, and weight of switch-mode power supplies. A drawback of toroidal transformer construction is the higher labor cost of winding. This is because it is necessary to pass the entire length of a coil winding through the core aperture each time a single turn is added to the coil. As a consequence, toroidal transformers are uncommon above ratings of a few kVA. Small distribution transformers may achieve some of the benefits of a toroidal core by splitting it and forcing it open, then inserting a bobbin containing primary and secondary windings.

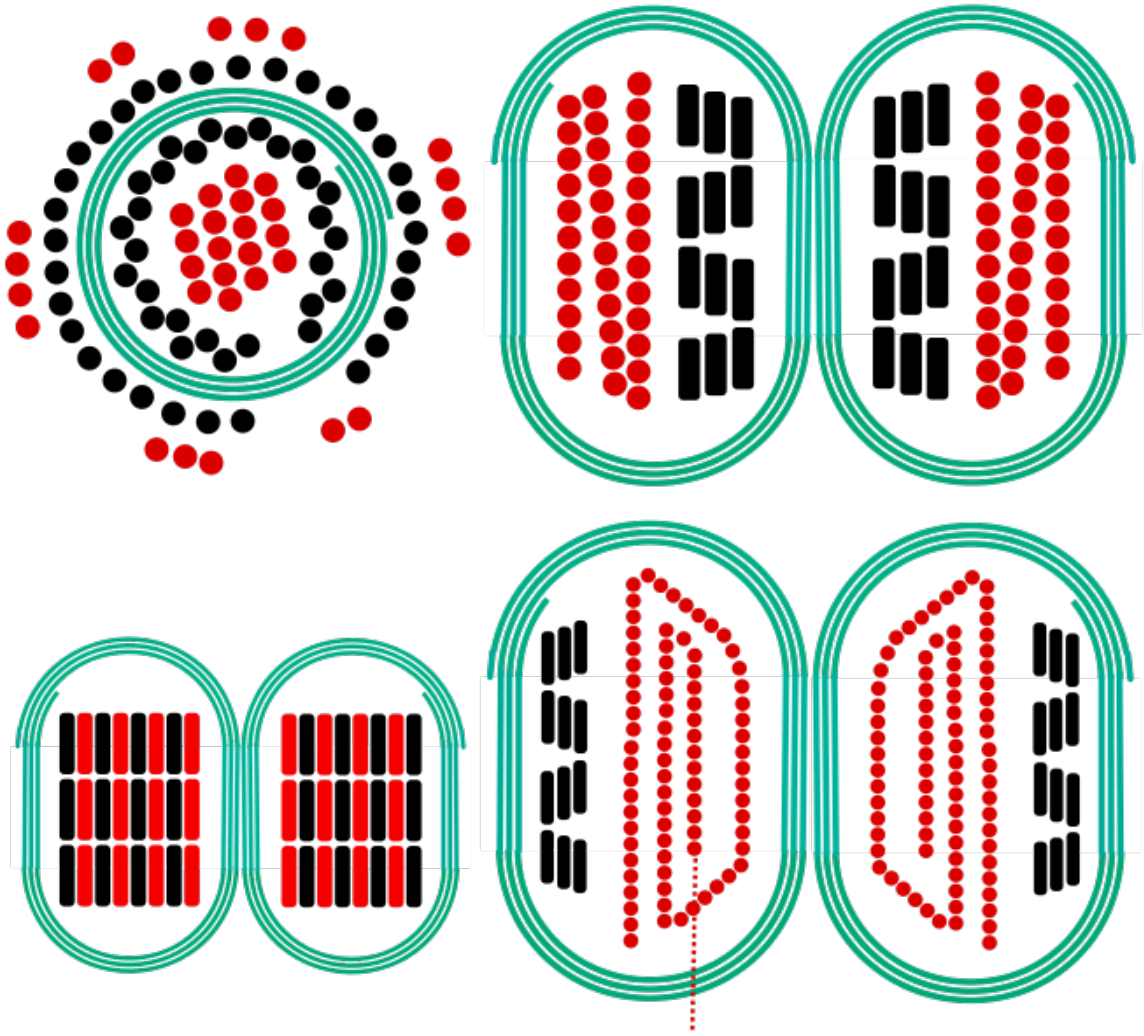
Air cores

A physical core is not an absolute requisite and a functioning transformer can be produced simply by placing the windings near each other, an arrangement termed an "air-core" transformer. The air which comprises the magnetic circuit is essentially lossless, and so an air-core transformer eliminates loss due to hysteresis in the core material. The leakage inductance is inevitably high, resulting in very poor regulation, and so such designs are unsuitable for use in power distribution. They have however very high bandwidth, and are frequently employed in radio-frequency applications, for which a satisfactory coupling coefficient is maintained by carefully overlapping the primary and secondary windings. They're also used for resonant transformers such as Tesla coils where they can achieve reasonably low loss in spite of the high leakage inductance.

Windings



Windings are usually arranged concentrically to minimize flux leakage.



Cut view through transformer windings. White: insulator. Green spiral: Grain oriented silicon steel. Black: Primary winding made of oxygen-free copper. Red: Secondary winding. Top left: Toroidal transformer. Right: C-core, but E-core would be similar. The black windings are made of film. Top: Equally low capacitance between all ends of both windings. Since most cores are at least moderately conductive they also need insulation. Bottom: Lowest capacitance for one end of the secondary winding needed for low-power high-voltage transformers. Bottom left: Reduction of leakage inductance would lead to increase of capacitance.

The conducting material used for the windings depends upon the application, but in all cases the individual turns must be electrically insulated from each other to ensure that the current travels throughout every turn. For small power and signal transformers, in which currents are low and the potential difference between adjacent turns is small, the coils are often wound from enamelled magnet wire, such as Formvar wire. Larger power transformers operating at high voltages may be wound with copper rectangular strip conductors insulated by oil-impregnated paper and blocks of pressboard.

High-frequency transformers operating in the tens to hundreds of kilohertz often have windings made of braided Litz wire to minimize the skin-effect and proximity effect losses. Large power transformers use multiple-stranded conductors as well, since even at low power frequencies non-uniform distribution of current would otherwise exist in high-current windings. Each strand is individually insulated, and the strands are arranged so that at certain points in the winding, or throughout the whole winding, each portion occupies different relative positions in the complete conductor. The transposition equalizes the current flowing in each strand of the conductor, and reduces eddy current losses in the winding itself. The stranded conductor is also more flexible than a solid conductor of similar size, aiding manufacture.

For signal transformers, the windings may be arranged in a way to minimize leakage inductance and stray capacitance to improve high-frequency response. This can be done by splitting up each coil into sections, and those sections placed in layers between the sections of the other winding. This is known as a stacked type or interleaved winding.

Both the primary and secondary windings on power transformers may have external connections, called taps, to intermediate points on the winding to allow selection of the voltage ratio. In distribution transformers the taps may be connected to an automatic on-load tap changer for voltage regulation of distribution circuits. Audio-frequency transformers, used for the distribution of audio to public address loudspeakers, have taps to allow adjustment of impedance to each speaker. A center-tapped transformer is often used in the output stage of an audio power amplifier in a push-pull circuit. Modulation transformers in AM transmitters are very similar.

Certain transformers have the windings protected by epoxy resin. By impregnating the transformer with epoxy under a vacuum, one can replace air spaces within the windings with epoxy, thus sealing the windings and helping to prevent the possible formation of corona and absorption of dirt or water. This produces transformers more suited to damp or dirty environments, but at increased manufacturing cost.

Coolant

Cut-away view of three-phase oil-cooled transformer. The oil reservoir is visible at the top. Radiative fins aid the dissipation of heat.

High temperatures will damage the winding insulation. Small transformers do not generate significant heat and are cooled by air circulation and radiation of heat. Power transformers rated up to several hundred kVA can be adequately cooled by natural convective air-cooling, sometimes assisted by fans. In larger transformers, part of the design problem is removal of heat. Some power transformers are immersed in transformer oil that both cools and insulates the windings. The oil is a highly refined mineral oil that remains stable at transformer operating temperature. Indoor liquid-filled transformers are required by building regulations in many jurisdictions to use a non-

flammable liquid, or to be located in fire-resistant rooms. Air-cooled dry transformers are preferred for indoor applications even at capacity ratings where oil-cooled construction would be more economical, because their cost is offset by the reduced building construction cost.

The oil-filled tank often has radiators through which the oil circulates by natural convection; some large transformers employ forced circulation of the oil by electric pumps, aided by external fans or water-cooled heat exchangers. Oil-filled transformers undergo prolonged drying processes to ensure that the transformer is completely free of water vapor before the cooling oil is introduced. This helps prevent electrical breakdown under load. Oil-filled transformers may be equipped with Buchholz relays, which detect gas evolved during internal arcing and rapidly de-energize the transformer to avert catastrophic failure. Oil-filled transformers may fail, rupture, and burn, causing power outages and losses. Installations of oil-filled transformers usually includes fire protection measures such as walls, oil containment, and fire-suppression sprinkler systems.

Polychlorinated biphenyls have properties that once favored their use as a coolant, though concerns over their environmental persistence led to a widespread ban on their use. Today, non-toxic, stable silicone-based oils, or fluorinated hydrocarbons may be used where the expense of a fire-resistant liquid offsets additional building cost for a transformer vault. Before 1977, even transformers that were nominally filled only with mineral oils may also have been contaminated with polychlorinated biphenyls at 10-20 ppm. Since mineral oil and PCB fluid mix, maintenance equipment used for both PCB and oil-filled transformers could carry over small amounts of PCB, contaminating oil-filled transformers.

Some "dry" transformers (containing no liquid) are enclosed in sealed, pressurized tanks and cooled by nitrogen or sulfur hexafluoride gas.

Experimental power transformers in the 2 MVA range have been built with superconducting windings which eliminates the copper losses, but not the core steel loss. These are cooled by liquid nitrogen or helium.

Insulation drying

Construction of oil-filled transformers requires that the insulation covering the windings be thoroughly dried before the oil is introduced. There are several different methods of drying. Common for all is that they are carried out in vacuum environment. The vacuum makes it difficult to transfer energy (heat) to the insulation. For this there are several different methods. The traditional drying is done by circulating hot air over the active part and cycle this with periods of vacuum (hot-air vacuum drying, HAV). More common for larger transformers is to use evaporated solvent which condenses on the colder active part. The benefit is that the entire process can be carried out at lower pressure and without influence of added oxygen. This process is commonly called vapour-phase drying (VPD).

For distribution transformers, which are smaller and have a smaller insulation weight, resistance heating can be used. This is a method where current is injected in the windings to heat the insulation. The benefit is that the heating can be controlled very well and it is energy efficient. The method is called low-frequency heating (LFH) since the current is injected at a much lower frequency than the nominal of the grid, which is normally 50 or 60 Hz. A lower frequency reduces the effect of the inductance in the transformer, so the voltage can be reduced.

Terminals

Very small transformers will have wire leads connected directly to the ends of the coils, and brought out to the base of the unit for circuit connections. Larger transformers may have heavy bolted terminals, bus bars or high-voltage insulated bushings made of polymers or porcelain. A large bushing can be a complex structure since it must provide careful control of the electric field gradient without letting the transformer leak oil.

Applications

Image of an electrical substation in Melbourne, Australia showing 3 of 5 220kV/66kV transformers, each with a capacity of 185MVA

A major application of transformers is to increase voltage before transmitting electrical energy over long distances through wires. Wires have resistance and so dissipate electrical energy at a rate proportional to the square of the current through the wire. By transforming electrical power to a high-voltage (and therefore low-current) form for transmission and back again afterward, transformers enable economical transmission of power over long distances. Consequently, transformers have shaped the electricity supply industry, permitting generation to be located remotely from points of demand. All but a tiny fraction of the world's electrical power has passed through a series of transformers by the time it reaches the consumer.

Transformers are also used extensively in electronic products to step down the supply voltage to a level suitable for the low voltage circuits they contain. The transformer also electrically isolates the end user from contact with the supply voltage.

Signal and audio transformers are used to couple stages of amplifiers and to match devices such as microphones and record players to the input of amplifiers. Audio transformers allowed telephone circuits to carry on a two-way conversation over a single pair of wires. A balun transformer converts a signal that is referenced to ground to a signal that has balanced voltages to ground, such as between external cables and internal circuits.

Chapter- 10

Control Relay and Eddy Current Brake

Control relay

A **control relay** is an electromechanical device which activates one or more switches according to the current through a coil not connected to the switches. The thyristor is a semiconductor device which carries out similar functions.

A relay is essentially an electromagnet with two possible states arranged so that when there is sufficient current the core of the relay's coil attracts a ferromagnetic armature which mechanically operates switches; a spring holds the armature away from the core when not actuated. The spring is designed to snap the contacts between two stable mechanical states; there should not be a range of coil current which allows the contacts to be in an intermediate state.

A relay allows circuits to be switched by electrical equipment: for example, a timer circuit with a relay could switch power at a preset time. For many years relays were the standard method of controlling industrial electronic systems. A number of relays could be used together to carry out complex functions (relay logic). The principle of relay logic is based on relays which energize and de-energize associated contacts. Relay logic is the predecessor of ladder logic, which is commonly used in Programmable logic controllers.

Operation

When electric current passes through a coil, magnetic north and south poles are produced across the gap separating the coil and armature. The relay is actuated, or latched, when sufficient current through the coil makes the attractive force between the core and armature overcome the spring tension. The relay remains latched as long as at least a specific value of *holding current*, which can be lower than the actuating current, flows through the coil. When the current through the coil is reduced below this value, the magnetic attraction of the armature becomes too weak to hold it in the actuated position and the spring snaps it to the non-actuated position.

There are many considerations involved in the correct selection of a control relay for a particular application. These considerations include factors such as speed of operation, sensitivity, and hysteresis. Although typical control relays operate in the 5 ms to 20 ms

range, relays with switching speeds as fast as 100 us are available. Reed relays which are actuated by low currents and switch fast are suitable for controlling small currents.

As for any switch, the current through the relay contacts (unrelated to the current through the coil) must not exceed a certain value to avoid damage. In the particular case of high-inductance circuits such as motors other issues must be addressed. When a power source is connected to an inductance, an input surge current which may be several times larger than the steady current exists. When the circuit is broken, the current cannot change instantaneously, which creates a potentially damaging spark across the separating contacts.

Consequently for relays which may be used to control inductive loads we must specify the maximum current that may flow through the relay contacts when it actuates, the *make rating*; the continuous rating; and the *break rating*. The make rating may be several times larger than the continuous rating, which is itself larger than the break rating.

Derating factors

Type of load	**% of rated value**
Resistive	75
Inductive	35
Motor	20
Filament	10
Capacitive	75

Control relays should not be operated above rated temperature because of resulting increased degradation and fatigue. Common practice is to derate 20 degrees Celsius from the maximum rated temperature limit. Relays operating at rated load are also affected by their environment. Oil vapors may greatly decrease the contact tip life, and dust or dirt may cause the tips to burn before their normal life expectancy. Control relay life cycle varies from 50,000 to over one million cycles depending on the electrical loads of the contacts, duty cycle, application, and the extent to which the relay is derated. When a control relay is operating at its derated value, it is controlling a lower value of current than its maximum make and break ratings. This is often done to extend the operating life of the control relay. Table 1 lists the relay and switch derating factors for typical industrial control applications.

Eddy current brake

An eddy current brake of a German ICE 3 in action.

An **eddy current brake**, like a conventional friction brake, is responsible for slowing an object, such as a train or a roller coaster. Unlike friction brakes, which apply pressure on two separate objects, eddy current brakes slow an object by creating eddy currents through electromagnetic induction which create resistance, and in turn either heat or electricity.

Construction and operation

Circular eddy current brake

Circular eddy current brake on 700 Series Shinkansen

Electromagnetic brakes are similar to electrical motors; non-ferromagnetic metal discs (rotors) are connected to a rotating coil, and a magnetic field between the rotor and the coil creates a resistance used to generate electricity or heat. When electromagnets are used, control of the braking action is made possible by varying the strength of the magnetic field. A braking force is possible when electric current is passed through the electromagnets. The movement of the metal through the magnetic field of the

electromagnets creates eddy currents in the discs. These eddy currents generate an opposing magnetic field, which then resists the rotation of the discs, providing braking force. The net result is to convert the motion of the rotors into heat in the rotors.

Japanese Shinkansen trains had employed circular eddy current brake system on trailer cars since 100 Series Shinkansen. However, N700 Series Shinkansen abandoned eddy current brakes in favour of regenerative brakes since 14 of the 16 cars in the trainset used electric motors.

Linear eddy current brake

The principle of the linear eddy current brake has been described by the French physicist Foucault, hence in French the eddy current brake is called the "frein à courants de Foucault".

The linear eddy current brake consists of a magnetic yoke with electrical coils positioned along the rail, which are being magnetized alternating as south and north magnetic poles. This magnet does not touch the rail, as with the magnetic brake, but is held at a constant small distance from the rail (approximately seven millimeters). It does not move along the rail, exerting only a vertical pull on the rail.

When the magnet is moved along the rail, it generates a non-stationary magnetic field in the head of the rail, which then generates electrical tension (Faraday's induction law), and causes eddy currents. These disturb the magnetic field in such a way that the magnetic force is diverted to the opposite of the direction of the movement, thus creating a horizontal force component, which works against the movement of the magnet.

The braking energy of the vehicle is converted in eddy current losses which lead to a warming of the rail. (The regular magnetic brake, in wide use in railways, exerts its braking force by friction with the rail, which also creates heat.)

The eddy current brake does not have any mechanical contact with the rail, and thus no wear, and creates no noise or odor. The eddy current brake is unusable at low speeds, but can be used at high speeds both for emergency braking and for regular braking.

The TSI (Technical Specifications for Interoperability) of the EU for trans-European high speed rail recommends that all newly built high speed lines should make the eddy current brake possible.

Eddy current brakes at the Intamin roller coaster *Goliath* in Walibi World (Netherlands)

The first train in commercial circulation to use such a braking is the ICE 3.

Modern roller coasters use this type of braking, but utilize permanent magnets instead of electromagnets, and require no electricity. However, their braking strength cannot be adjusted as easily as with an electromagnet.

Chapter- 11

Electrical Steel

Electrical steel, also called **lamination steel**, **silicon electrical steel**, **silicon steel** or **transformer steel**, is specialty steel tailored to produce certain magnetic properties, such as a small hysteresis area (small energy dissipation per cycle, or low core loss) and high permeability.

The material is usually manufactured in the form of cold-rolled strips less than 2 mm thick. These strips are called laminations when stacked together to form a core. Once assembled, they form the laminated cores of transformers or the stator and rotor parts of electric motors. Laminations may be cut to their finished shape by a punch and die, or in smaller quantities may be cut by a laser, or by wire erosion.

Metallurgy

Electrical steel is an iron alloy which may have from zero to 6.5% silicon (Si:5Fe). Silicon significantly increases the electrical resistivity of the steel, which decreases the induced eddy currents and thus reduces the core loss. Manganese and aluminum can be added up to 0.5%.

Increasing the amount of silicon inhibits eddy currents and narrows the hysteresis loop of the material, thus lowering the core losses. However, the grain structure hardens and embrittles the metal, which adversely affects the workability of the material, especially when rolling it. When alloying, the concentration levels of carbon, sulfur, oxygen and nitrogen must be kept low, as these elements indicate the presence of carbides, sulfides, oxides and nitrides. These compounds, even in particles as small as one micrometer in diameter, increase hysteresis losses while also decreasing magnetic permeability. The presence of carbon has a more detrimental effect than sulfur or oxygen. Carbon also causes magnetic aging when it slowly leaves the solid solution and precipitates as carbides, thus resulting in an increase in power loss over time. For these reasons, the carbon level is kept to 0.005% or lower. The carbon level can be reduced by annealing the steel in a decarburizing atmosphere, such as hydrogen.

Physical properties examples

Melting point: ~1,500 °C (example for ~3.1% silicon content)

Density: 7,650 kg/m^3 (example for 3% silicon content)

Resistivity: 47.2×10^{-8} (Ω·m) (example for 3% silicon content)

Grain orientation

There are two main types of electrical steel: grain-oriented and non-oriented.

Grain-oriented electrical steel usually has a silicon level of 3% (Si:11Fe). It is processed in such a way that the optimum properties are developed in the rolling direction, due to a tight control (proposed by Norman P. Goss) of the crystal orientation relative to the sheet. Due to the special orientation, the magnetic flux density is increased by 30% in the coil rolling direction, although its magnetic saturation is decreased by 5%. It is used for the cores of high-efficiency transformers, electric motor and generators. Cold Rolled Grain-oriented steel is often abbreviated to CRGO.

Non-oriented electrical steel usually has a silicon level of 2 to 3.5% and has similar magnetic properties in all directions, which makes it isotropic. It is less expensive and is used in applications where the direction of magnetic flux is changing, such as electric motors and generators. It is also used when efficiency is less important or when there is insufficient space to correctly orient components to take advantage of the anisotropic properties of grain-oriented electrical steel. Cold Rolled Non Grain-oriented steel is often abbreviated to CRNGO.

Amorphous steel

Transformers with amorphous steel cores can have core losses of one-third that of conventional steels. This material is a metallic glass prepared by pouring molten alloy steel onto a rotating cooled wheel, which cools the metal at a rate of about one megakelvin per second, so fast that crystals do not form. Amorphous steel has poorer mechanical properties and as of 2010 costs about twice as much as conventional steel, making it cost-effective only for some large distribution-type transformers.

Lamination coatings

Electrical steel is usually coated to increase electrical resistance between laminations, to provide resistance to corrosion or rust, and to act as a lubricant during die cutting. There are various coatings, organic and inorganic, and the coating used depends on the application of the steel. The type of coating selected depends on the heat treatment of the laminations, whether the finished lamination will be immersed in oil, and the working temperature of the finished apparatus. Very early practice was to insulate each lamination with a layer of paper or a varnish coating, but this reduced the stacking factor of the core and limited the maximum temperature of the core.

Magnetic properties

The magnetic properties of electrical steel are dependent on heat treatment, as increasing the average crystal size decreases the hysteresis loss. Hysteresis loss is determined by a standard test and for common grades of electrical steel may range from about 2 to 10 watts per kilogram (1 to 5 watts per pound) at 60 Hz and 1.5 tesla magnetic field strength. Semi-processed electrical steels are delivered in a state that, after punching the final shape, a final heat treatment develops the desired 150-micrometer grain size. The fully processed steels are usually delivered with insulating coating, full heat treatment, and defined magnetic properties, for applications where the punching operation does not significantly degrade the material properties. Excessive bending, incorrect heat treatment, or even rough handling of core steel can adversely affect its magnetic properties and may also increase noise due to magnetostriction

Magnetic properties of electrical steels are tested using the internationally standardised Epstein frame method.

Practical concerns

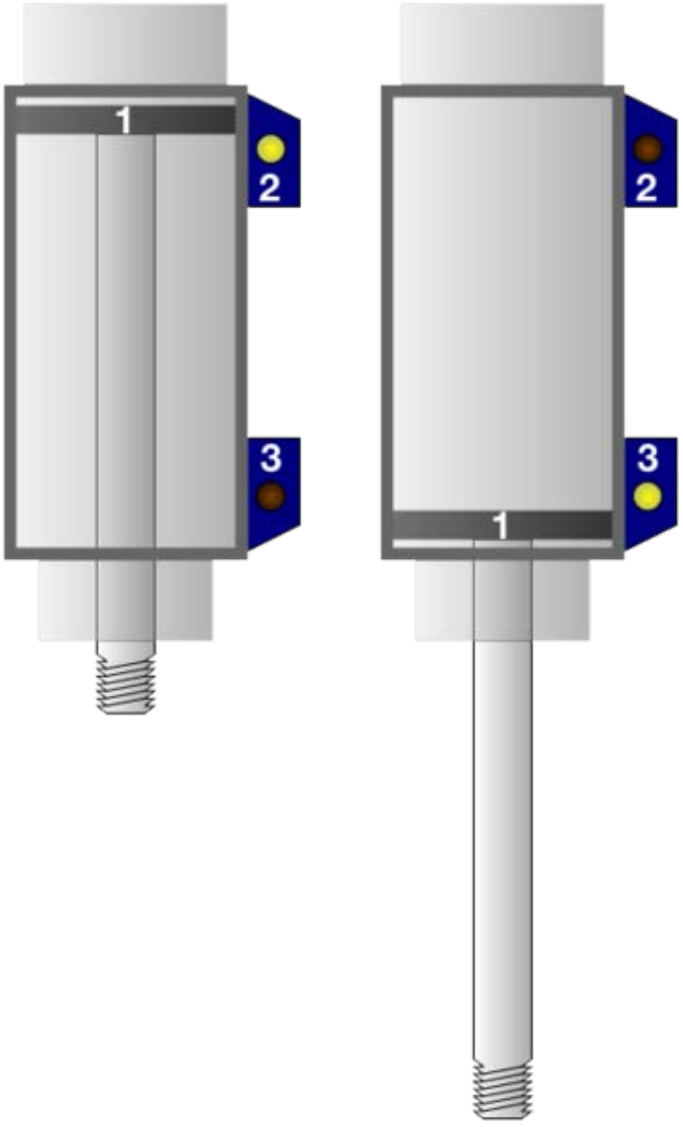
Core steel is much more costly than mild steel—in 1981 it was more than twice the cost per unit weight.

The size of magnetic domains in the sheet can be reduced by scribing the surface of the sheet with a laser, or mechanically. This greatly reduces the hysteresis losses in the assembled core.

Chapter- 12

Hall Effect Sensor and Induction Loop

Hall effect sensor



The magnetic piston (1) in this pneumatic cylinder will cause the Hall effect sensors (2 and 3) mounted on its outer wall to activate when it is fully retracted or extended.

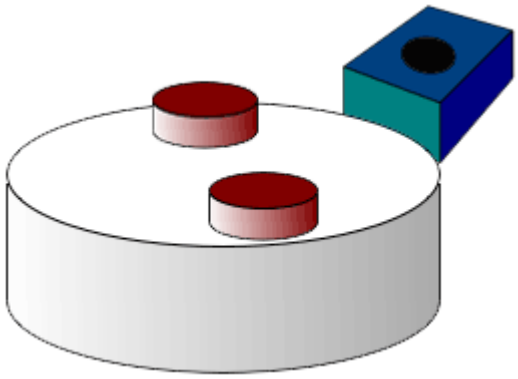
Clutch with Hall Effect sensor.

A **Hall effect sensor** is a transducer that varies its output voltage in response to changes in magnetic field. Hall sensors are used for proximity switching, positioning, speed detection, and current sensing applications.

In its simplest form, the sensor operates as an analogue transducer, directly returning a voltage. With a known magnetic field, its distance from the Hall plate can be determined. Using groups of sensors, the relative position of the magnet can be deduced.

Electricity carried through a conductor will produce a magnetic field that varies with current, and a Hall sensor can be used to measure the current without interrupting the circuit. Typically, the sensor is integrated with a wound core or permanent magnet that surrounds the conductor to be measured.

Frequently, a Hall sensor is combined with circuitry that allows the device to act in a digital (on/off) mode, and may be called a switch in this configuration. Commonly seen in industrial applications such as the pictured pneumatic cylinder, they are also used in consumer equipment; for example some computer printers use them to detect missing paper and open covers. When high reliability is required, they are used in keyboards.



Hall sensors are commonly used to time the speed of wheels and shafts, such as for internal combustion engine ignition timing, tachometers and anti-lock braking systems. They are used in brushless DC electric motors to detect the position of the permanent magnet. In the pictured wheel with two equally spaced magnets, the voltage from the sensor will peak twice for each revolution. This arrangement is commonly used to regulate the speed of disc drives.

Hall probe

A hall probe contains an indium compound crystal such as indium antimonide, mounted on an aluminum backing plate, and encapsulated in the probe head. The plane of the crystal is perpendicular to the probe handle. Connecting leads from the crystal are brought down through the handle to the circuit box.

When the Hall Probe is held so that the magnetic field lines are passing at right angles through the sensor of the probe, the meter gives a reading of the value of magnetic flux density (B). A current is passed through the crystal which, when placed in a magnetic field has a “Hall Effect” voltage developed across it. The Hall Effect is seen when a conductor is passed through a uniform magnetic field. The natural electron drift of the charge carriers causes the magnetic field to apply a Lorentz force (the force exerted on a charged particle in an electromagnetic field) to these charge carriers. The result is what is seen as a charge separation, with a build up of either positive or negative charges on the bottom or on the top of the plate. The crystal measures 5 mm square. The probe handle, being made of a non-ferrous material, has no disturbing effect on the field.

A Hall Probe is enough to measure the Earth's magnetic field. It must be held so that the Earth's field lines are passing directly through it. It is then rotated quickly so the field lines pass through the sensor in the opposite direction. The change in the flux density reading is double the Earth's magnetic flux density. A hall probe must first be calibrated against a known value of magnetic field strength. For a solenoid the hall probe is placed in the center.

Hall Effect Sensor Interface

Hall effect sensor may require analog circuitry to be interfaced to microprocessors. These interfaces may include inpu diagnostics, fault protection for transient conditions and short/open circuit detection. It may also provide and monitor the current to the hall effect sensor itself. There are precision IC products available to handle these features.

Induction loop

Induction loop is a term used to describe an electromagnetic communication- and detection system, relying on the fact that a moving magnet will induce an electrical current in a nearby conducting wire. Induction loops are used for transmission and reception of communication signals, or for detection of metal objects in metal detectors or vehicle presence indicators. A common modern use for induction loops is to provide hearing assistance to hearing-aid users.

Implementation

An example of the Inductance loop installed in the road for cars and bikes

The "aerial" system of an induction loop installation can consist of one or more loops of a conductive element.

In industrial applications this might be a large single- or multi-turn, loop, or a complex multi-lobed, phase coincident sub-loop design, most effectively mounted above the required reception area in industrial applications.

An audio induction loop might have one or more loops sometimes with a phase shift between them, and either near to or around the area in which a hearing aid user would be present. Many different configurations can be used depending on the application.

Such an induction loop receiver is classically a very small iron-cored inductor (telecoil), although Rediffusion demonstrated a prototype Hall-Effect system in its PLL FM system.

The system commonly uses an analogue power amplifier matched to the low impedance of the transmission loop. The transmission is normally direct rather than superimposed or modulated upon a carrier, though multi-channel systems have been implemented using modulation.

Vehicle detection (inductive) loops are used to count vehicles passing or arriving at a certain point, for instance approaching a traffic light, and in motorway traffic management. An insulated, electrically conducting loop is installed under the road. An electrical voltage is generated when a ferrous (containing iron or steel) body passes close to the wire/loop

Other definitions

A different sort of "induction loop" is applied to metal detectors, where a large coil, which forms part of a resonant circuit, is effectively "detuned" by the coil's proximity to a conductive object. The detected object may be metallic, (metal and cable detection) or conductive/capacitive (stud/cavity detection) Other configurations of this equipment use two or more receiving coils, and the detected object modifies the inductive coupling or alters the phase angle of the voltage induced in the receiving coils relative to the oscillator coil.

An increasingly common application is for providing hearing aid-compatible "assistive listening" telecoil. In this application a loop or series of loops is used to provide an audio frequency oscillating magnetic field in an area where a hearing aid user may be present. Many hearing aids contain a telecoil which allows the user to receive and hear the magnetic field and remove the normal audio signal provided from the hearing aid microphone site. These loops are often referred to as a hearing loop or Audio induction loop.

Modern day applications

- vehicle detection (e.g. at traffic lights)
- car park Parking Guidance and Information systems
- metal detectors
- audio induction loops

Historical applications

- An Anti-submarine indicator loop was a device used to detect submarines and surface vessels using specially designed submerged cables connected to a galvanometer.

Chapter- 13

Magnetic Amplifier

The **magnetic amplifier** (colloquially known as a "mag amp") is an electromagnetic device for amplifying electrical signals. The magnetic amplifier was invented early in the 20th century, and was used as an alternative to vacuum tube amplifiers where robustness and high current capacity were required. World War II Germany perfected this type of amplifier, and it was used for instance in the V-2 rocket. The magnetic amplifier was most prominent in power control and low-frequency signal applications from 1947 to about 1957, when the transistor began to supplant it. The magnetic amplifier has now been largely superseded by the transistor-based amplifier, except in a few safety critical, high reliability or extremely demanding applications. Combinations of transistor and mag-amp techniques are still used.

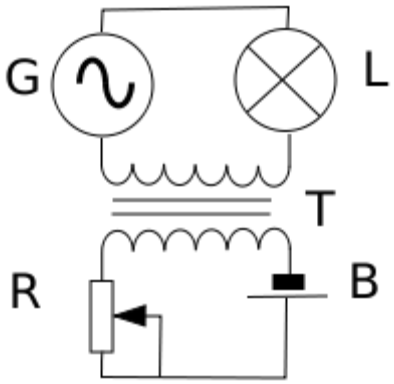
Strengths

The magnetic amplifier is a static device with no moving parts. It has no wear-out mechanism and has a good tolerance to mechanical shock and vibration. It requires no warm-up time. Multiple isolated signals may be summed by additional control windings on the magnetic cores. The windings of a magnetic amplifier have a higher tolerance to momentary overloads than comparable solid-state devices. The magnetic amplifier is also used as a transducer in applications such as current measurement and the flux gate compass

Limitations

The gain available from a single stage is limited and low compared to electronic amplifiers. Frequency response of a high gain amplifier is limited to about one-tenth the excitation frequency, although this is often mitigated by exciting magnetic amplifiers with currents at higher than utility frequency. Solid-state amplifiers can be more compact and efficient than magnetic amplifiers. The bias and feedback windings are not unilateral, and may couple energy back from the controlled circuit into the control circuit. This complicates the design of multistage amplifiers when compared with electronic devices.

Principle of operation



A saturable reactor, illustrating the principle of a magnetic amplifier

Visually a mag amp device may resemble a transformer but the operating principle is quite different from a transformer - essentially the mag amp is a saturable reactor. It makes use of magnetic saturation of the core, a non-linear property of a certain class of transformer cores. For controlled saturation characteristics the magnetic amplifier employs core materials that have been designed to have a specific B-H curve shape that is highly rectangular, in contrast to the slowly-tapering B-H curve of softly saturating core materials that are often used in normal transformers.

The typical magnetic amplifier consists of two physically separate but similar transformer magnetic cores, each of which has two windings - a control winding and an AC winding. A small DC current from a low impedance source is fed into the series-connected control windings. The AC windings may be connected either in series or in parallel, the configurations resulting in different types of mag amps. The amount of control current fed into the control winding sets the point in the AC winding waveform at which either core will saturate. In saturation, the AC winding on the saturated core will go from a high impedance state ("off") into a very low impedance state ("on") - that is, the control current controls at which voltage the mag amp switches "on".

A relatively small DC current on the control winding is able to control or switch large AC currents on the AC windings. This results in current amplification.

Applications

Magnetic amplifiers were important as modulation and control amplifiers in the early development of voice transmission by radio. A magnetic amplifier was used as voice modulator for a 2 kilowatt Alexanderson alternator, and magnetic amplifiers were used in the keying circuits of large high-frequency alternators used for radio communications. Magnetic amplifiers were also used to regulate the speed of Alexanderson alternators to maintain the accuracy of the transmitted radio frequency.

The ability to control large currents with small control power made magnetic amplifiers useful for control of lighting circuits, for stage lighting and for advertising signs. Saturable reactor amplifiers were used for control of power to industrial furnaces. Small magnetic amplifiers were used for radio tuning indicators, control of small motor and cooling fan speed, control of battery chargers.

Magnetic amplifiers were used extensively as the switching element in early switched-mode (SMPS) power supplies, as well as in lighting control. Semiconductor based solid-state switches have largely superseded them, though recently there has been some regained interest in using mag amps in compact and reliable switching power supplies. PC ATX power supplies often use mag amps for secondary side voltage regulation.

Magnetic amplifiers are still used in some arc welders.

Magnetic amplifier transformer cores designed specifically for switch mode power supplies are currently manufactured by several large electromagnetics companies, including Metglas and Mag-Inc.

Magnetic amplifiers can be used for measuring high DC-voltages without direct connection to the high voltage and are therefore still used in the HVDC-technique.

Magnetic amplifiers were used by locomotives to detect wheel slip, until replaced by Hall Effect current transducers. The cables from two traction motors passed through the core of the device. During normal operation the resultant flux was zero as both currents were the same and in opposite directions. However, the currents would differ during wheel slip, producing a resultant flux that acted as the Control winding, developing a voltage across a resistor in series with the AC winding which was sent to the wheel slip correction circuits.

History

Early development

A voltage source and a series connected variable resistor may be regarded as a direct current signal source for a low resistance load such as the control coil of a saturable reactor which amplifies the signal. Thus, in principle, a saturable reactor is already an amplifier, although before 20th century they were used for simple tasks, such as controlling lighting and electrical machinery as early as 1885.

In the early 20th Century, the General Electric Company, under the direction of engineer E. F. W. Alexanderson, developed a system of transoceanic radio communications, using continuous wave transmission over great distances. Alexanderson drew upon the work of Nikola Tesla and Reginald Fessenden as the inspiration for his system.

The result of this work was the 2 kW Alexanderson alternator, which produced radio frequencies from 50 to 100 kHz and which critics had previously denounced as

impractical. Later, Guglielmo Marconi took a vested interest in the project and, in 1915, witnessed a demonstration of a new, 50 kW, 50 kHz alternator.

The experimental telegraphy and telephony demonstrations made during 1917 attracted the attention of the US Government, especially in light of partial failures in the transoceanic cable that snaked across the bottom of the Atlantic Ocean. The 50 kW alternator was commandeered by the US Navy and put into service in January 1918 and was used until 1920, when a 200 kW generator-alternator set was built and installed.

Usage in radio

Magnetic amplifiers were used early on to control large, high-power alternators by turning them on and off for telegraphy or to vary the signal for voice modulation. However, the alternator's frequency limits were rather low to where a frequency multiplier had to be utilized to generate higher radio frequencies than the alternator was capable of producing. Even so, early magnetic amplifiers incorporating powdered-iron cores were incapable of producing radio frequencies above approximately 200 kHz. Other core materials, such as ferrite cores and oil-filled transformers, would have to be developed to allow the amplifier to produce higher frequencies.

Usage in aircraft

Magnetic amplifiers were used in aircraft systems (avionics) before the advent of high reliability semiconductors. They were important in implementing early autoland systems and Concorde made use of the technology for the control of its engine air intakes before subsequent development of a replacement system using digital electronics.

Usage in computing

Magnetic amplifiers were widely studied during the 1950s as a potential switching element for mainframe computers. Mag amps could be used to sum several inputs in a single core, which was very useful in the arithmetic logic unit (ALU). Custom tubes could do the same, but transistors could not, so the mag amp was able to combine the advantages of tubes and transistors in an era when the latter were expensive and unreliable.

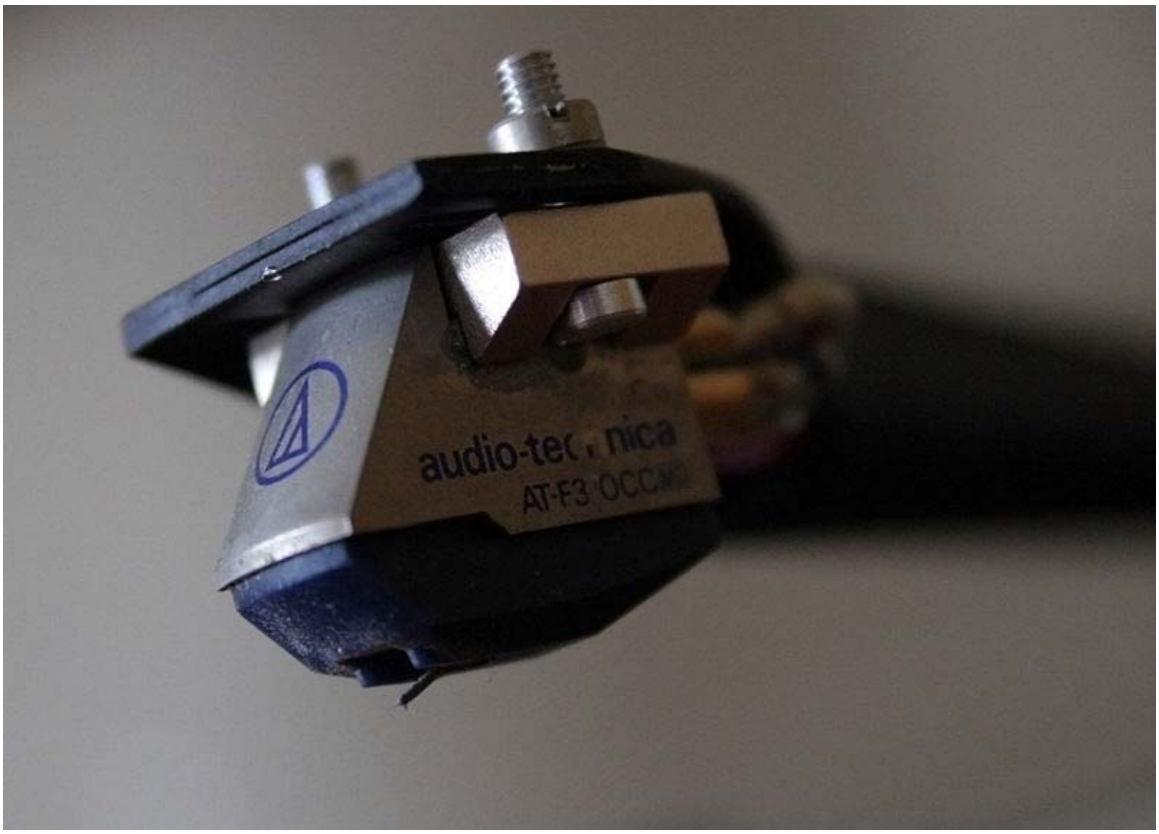
However, that era was very short, lasting from the mid 1950s to about 1960, at which point new fabrication techniques were producing great improvements in transistors and dramatically lowering their price points. Only one large-scale mag amp machine was put into production, the UNIVAC Solid State, but a number of contemporary late-1950's/early-1960s computers made some use of the technology, like the Ferranti Orion.

Misnomer uses

In the 1970s, Robert Carver designed and produced several high quality high-powered audio amplifiers, calling them magnetic amplifiers. In fact, they were in most respects conventional audio amplifier designs with an unusual power supply circuit.

Chapter- 14

Magnetic Cartridge



An Audio Technica AT-F3 MC cartridge

A **magnetic cartridge** is a transducer used for the playback of gramophone records on a turntable or phonograph. It converts mechanical vibrational energy from a stylus riding in a spiral record groove into an electrical signal that is subsequently amplified and then converted back to sound by a loudspeaker system.

History

The first electric pick-ups were developed in about 1925. They used a piezo-electric crystal of quartz, stimulated by a stylus made of sapphire or diamond. The magnetic

cartridge is presently the most common form of sound pickup used and came into use in the 1950s, following the introduction of magnetic cutter heads around 1945 for mastering records.

Types

In high-fidelity systems, crystal and ceramic pickups have been replaced by the magnetic cartridge, using either a **moving magnet** or a **moving coil**.

Compared to the crystal and ceramic pickups, the magnetic cartridge gives improved playback fidelity and reduces record wear by tracking the groove with lighter pressure. Magnetic cartridges use much lower tracking forces and thus damage the record grooves less. They also have a lower output voltage than a crystal or ceramic pickup, in the range of only a few millivolts, thus requiring greater amplification.

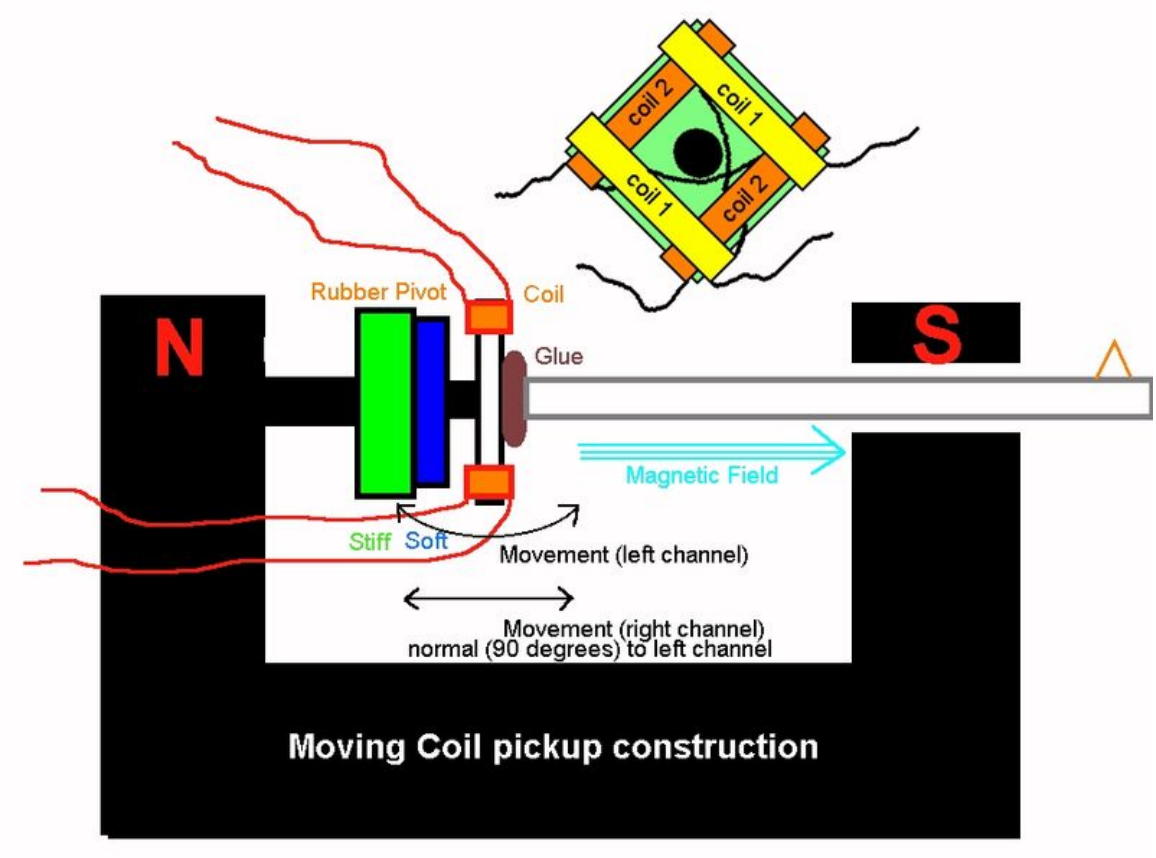
Moving Magnet (MM) cartridges

In a moving magnet cartridge, the stylus cantilever carries a tiny permanent magnet, which is positioned between two sets of fixed coils (in a stereophonic cartridge), forming a tiny electromagnetic generator. As the magnet vibrates in response to the stylus following the record groove, it induces a tiny current in the coils.

Because the magnet is small and has little mass, and is not coupled mechanically to the generator (as in a ceramic cartridge), a properly adjusted stylus follows the groove more faithfully while requiring less tracking force (the downward pressure on the stylus).

There is a sub-category. Moving iron and induced magnet types (ADC being a well known example) which have the magnet fixed and move a piece of iron or other ferrous alloy in the field of the magnet to produce the signal within the fixed coils.

Moving Coil (MC) cartridges



The MC design is again a tiny electromagnetic generator, but (unlike an MM design) with the magnet and coils reversed: the coils are attached to the stylus, and move within the field of a permanent magnet. The coils are tiny and made from very fine wire, so are even lighter than the small magnet used in an MM cartridge, thus improving the tracking ability of the cartridge. This can give extended frequency response as well as greater fidelity.

A disadvantage however is that moving-coil cartridges generate an even lower voltage signal than a moving-magnet type cartridge. This is because the moving coil cannot be large enough (it would be too heavy) to generate equivalent voltage levels. The resulting signal is only a few hundred microvolts, and thus more easily swamped by noise, induced hum, etc. Thus it is more challenging to design a preamplifier with the extremely low noise inputs needed for moving-coil cartridge, therefore a "step up transformer" is sometimes used instead.

Moving coil cartridges are extremely small precision instruments and are therefore generally expensive, but are frequently preferred by audiophiles due to their better performance.

Moving Micro Cross (MMC) cartridges

The MMC design was invented and patented by Bang & Olufsen. Since it often uses a special mounting, it can be mostly found in Bang & Olufsen turntables (which also cannot use another type of cartridge). Apart from being produced for the Bang & Olufsen mounting system, the SP12 and SP14 were also available in standard 1/2" mount.

The MMC cartridge is a *moving iron* design. Magnets and coils are stationary while the *micro cross* moves with the stylus, thereby varying the distances between the arms of the cross and the magnets. The design obviously offers more freedom concerning magnet and coil mass (compared to MC and MM cartridges). For example, the MMC20 (presented in 1978) uses four coils wound around the magnetic cores with 1200 turns each. Minimizing the moving mass also reduces the unavoidable wear on the records.

It is also claimed that the MMC design allows for superior channel separation, since each channel's movements appear on a separate axis.

Moving Magnet vs. Moving Coil debate

Moving magnet cartridges are more commonly found at the 'lower-end' of the market, while the 'higher-end' tends to be dominated by moving coil designs. The debate as to whether MM or MC designs can ultimately produce the better sound is often heated and subjective. The distinction between the two is often blurred by cost and design considerations - i.e. can an MC cartridge requiring another step-up amplification outperform well made MM cartridges that need simpler front-end stages? Every now and then a design comes along to re-open this debate. A good example being the Linn K9 (now discontinued) - regarded by some as one of the better MM designs and competitive with MC alternatives costing more. Amongst others, Grace, ADC and Grado also manufactured notably good designs.

"Decca" Cartridges (aka "Moving Iron")

The Decca phono cartridges were a unique design, with fixed magnets and coils. The stylus shaft was composed of the diamond tip, a short piece of soft iron, and an L-shaped cantilever made of non-magnetic steel. Since the iron was placed very close to the tip (within 1 mm), the motions of the tip could be tracked very accurately. Decca engineers called this "positive scanning". Vertical and lateral compliance was controlled by the shape and thickness of the cantilever. Decca cartridges had a reputation for being very musical; however early versions required more tracking force than competitive designs - making record wear a concern.

Stereo reproduction

One reason that magnetic cartridges superseded the crystal pick-up was the relative ease with which it could be made to reproduce stereo recordings, which were introduced in 1958. In a stereo recording, the two channels are arranged to drive the record cutter head

at an angle of 45° to the vertical, effectively encoding each channel in the left and right V-shaped walls of the record groove. This system worked well, since it provided full compatibility with a monaural pick-up, so stereo records could be played on older mono equipment. To reproduce the stereo signal, the cartridge simply arranges pairs of coils at 45° to complement the cutting process. With careful design, the coils can be shielded from each other electrically and mechanically such that stereo separation is maximised.

Comparison with crystal technology

Piezoelectric crystal or ceramic pickups had a few clear advantages. They were much easier and cheaper to make and were more robust than the delicate magnetic pickup. In addition, the output voltage from a crystal pickup is relatively large, requiring less amplification, which helped improve the signal-to-noise ratio. However, the signal from a crystal is not an accurate reproduction of the recording, as there is a lot of distortion introduced. The stylus is coupled to the crystal in a fairly rigid manner, which is also not as good at following rapid changes in the record grooves, so frequency response also suffers. This requires a greater tracking force, which in turn wears the records out faster. (This has earned cheap portable record players, the nickname "portable grinding wheel", in some circles.)

By contrast, since the magnetic cartridge is not mechanically coupled, the stylus and lever arm weight can be made exceedingly small. This gives extended frequency response and low distortion. The distortion is further minimised by the fact that there is more inherent linearity in the induction principle than there is in the piezo-electric one. Since the lighter stylus requires very low tracking forces, it requires a more sophisticated counterweighted arm, but reduces record wear.

The output from the magnetic cartridge is only a few millivolts compared to several tenths of a volt from a crystal or ceramic pickup. This requires an additional preamplifier stage. Careful design and shielding in the signal cables and amplifier is needed to prevent unwanted noise (shot noise or EMI). The magnetic induction principle also naturally leads to a linearly rising response with increasing frequency, and this needs to be compensated for to give correct (flat) frequency reproduction. Conversely, the very low bass frequencies are not efficiently picked up, so a strong bass boost is needed. This can amplify unwanted low frequency noise such as that from the turntable motor and drive mechanism itself (rumble). Crystal pickups do not suffer from these drawbacks and give a much better bass performance, so they may be preferred in some applications (such as DJ-ing), where robustness and good bass are favoured over highest fidelity reproduction. The moving coil pickups tend to have an even lower level of output, and so usually require a step-up transformer or very special preamplifiers to bring these signals up to the level of input that a standard amplifier requires. These special preamplifiers must have very low noise indeed (some have even been cryogenically cooled), and hence tend to be expensive. Many audiophiles claim that the benefits are worthwhile. There are higher output moving coil pickup cartridges that can be used directly into a normal moving magnet input, but these are in the minority.

The bass boost and high frequency rolloff can be conveniently incorporated into the preamplifier which implements the RIAA equalization curve, a noise reduction technique used on all modern records.

The magnetic nature of the modern pickup means it must be shielded from external fields, especially from the loudspeakers of the same system, or unpleasant and possibly damaging feedback can occur. For this reason, the cartridge itself has a shield, usually of mu-metal, to help screen out unwanted fields.

The stylus

The stylus, or "needle", is a crucial part of the record player (or 'phonograph' or 'gramophone', both terms now archaic), as it is the one part of the system that actually contacts the recorded disc and transfers its vibrations to the rest of the system. It is the part which also suffers the greatest wear. There are two desired qualities in a stylus: first, that it faithfully follows the contours of the recorded groove and transfers the vibration to the system, and second, that it does not damage the recorded disc.

History of materials

Early phonograph styli in mechanical players were steel, or even fibre, needles, usually with a shank about 1/8" (3 mm) in diameter, ground to a sharp point. These were easily replaceable by the user, as they had a very limited life and wore out fairly rapidly with use. Extensive play tended to wear records out as well as needles.

When the electronic phonograph was introduced, styli were included as part of the pickup cartridge.

In early times, wear of cylinders and glass styli was problematic. By 1908, sapphire styli were being attached to the rice paper diaphragms, and in 1912 Edison began to use the diamond stylus. Typical low cost crystal cartridges of the 1950s tracked at 5-8 grams of force and used replaceable osmium tipped steel styli called *needles*, which styli might last only five to ten plays before needing replacement. To help make the stylus last longer, sapphire styli for 78 rpm records or diamond styli for 33 1/3 rpm LPs were re-introduced. A 78 rpm stylus typically has a 3 mil (0.08 mm) diameter while a stylus for 33 1/3 rpm *microgroove* LP discs has a 1 mil (0.03 mm) diameter tip to fit the narrower groove. Sapphire might be good for 40 or 50 hours while a diamond would last at least ten times that long. Typically, these early cartridges were of the "flip-over" type; the cartridge had a stylus on either side, one for 78 rpm discs, the other for 33 and 45 rpm ("microgroove") records. The entire cartridge could be rotated 180° by means of a knob or lever at the end of the tonearm to use the desired stylus.

Later, starting in the 1960s, most manufacturers settled on diamond-tipped styli for all cartridges. Magnetic cartridges lowered tracking forces to 1-2 grams, and with the obsolescence of 78's, diamond became the standard stylus material. Moving magnet

cartridges often have replaceable styli; most moving coil cartridges do not offer user-replaceable styli, although some manufacturers offer a trade-in or a re-tipping service.

Magnetic cartridge manufacturers usually provide a specialized range of styli for DJ use. More rugged conical styli are required due to the frequent reversals of direction involved in scratching and back-cueing.

Stylus shape

The physical shape of the stylus has a bearing on its performance. The most obvious shape is the spherical stylus (also known as "conical"), where the tip of the stylus is ground to a hemisphere for playing monophonic recordings or for rugged use. However, this shape is unable to faithfully track all possible variations in a record groove. Better quality LP styli use an elliptical or "line contact" shape, arranged with its 0.02 mm (0.0007 inch) long axis across the record groove. The short axis may be from 0.005 to 0.01 mm (.0002 to .0004 inch) depending on the particular design. This shape followed the undulations of the groove better than spherical styli because they more closely resembled the triangular cutter used to create the groove.

In later years, bi-radial styli appeared where, in the perpendicular axis, the tip of the stylus was ground to a smaller radius than the main body. The result was a stylus that rode slightly lower in the groove and even more closely matched the shape of the triangular cutter than the elliptical design. There were several variations on the basic bi-radial design depending on the manufacturer who produced them.

Cantilevers

The cantilever is the arm that connects the stylus to the magnet or pickup coils. Most cartridges have cantilevers made from aluminium or boron; some very expensive models have ruby, diamond, beryllium or carbon fiber cantilevers chosen for their exceptional stiffness.