Electrical Components used in Electrical Engineering

Meri Marlow

First Edition, 2012

ISBN 978-81-323-3542-9

Published by:
University Publications
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.

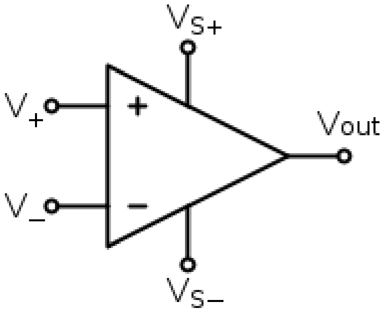
An **operational amplifier** ("op-amp") is a DC-coupled high-gain electronic voltage amplifier with a differential input and, usually, a single-ended output. An op-amp produces an output voltage that is typically hundreds of thousands times larger than the voltage *difference* between its input terminals.

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
- $V_{\mathrm{S+}}$: positive power supply
- $V_{\mathrm{S-}}$: negative power supply

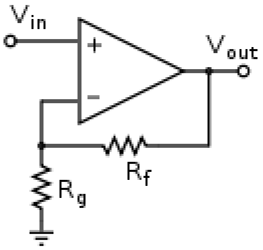
The power supply pins ($V_{\mathrm{S+}}$ and $V_{\mathrm{S-}}$) can be labeled in different ways. Despite different labeling, the function remains the same — to provide additional power for amplification of the signal. Often these pins are left out of the diagram for clarity, and the power configuration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

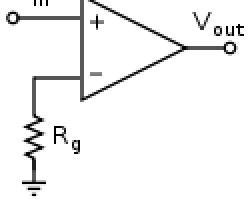
$$V_{\text{out}} = (V_+ - V_-)\,A_{OL}$$

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.



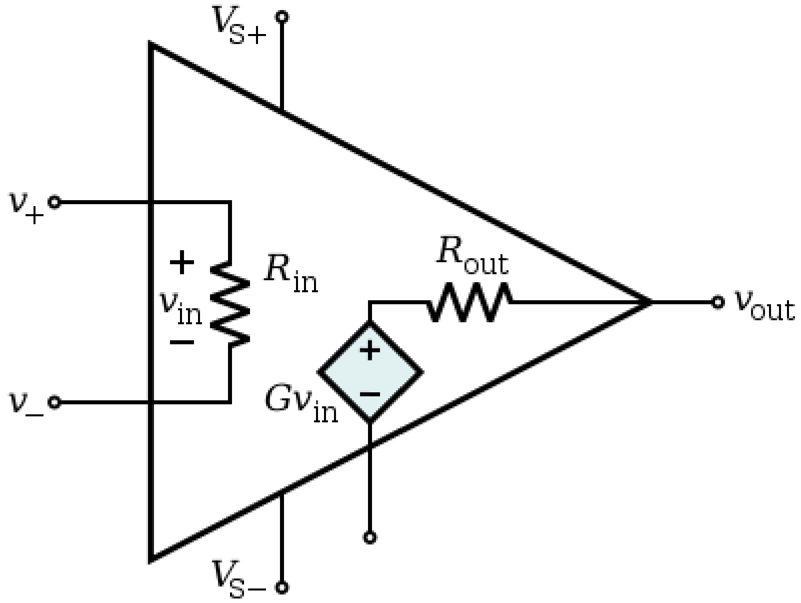
Adding negative feedback via the voltage divider R_f,R_g reduces the gain. Equilibrium will be established when V_{out} is just sufficient to reach around and "pull" the inverting input to the same voltage as V_{in}. As a simple example, if V_{in} = 1 V and $R_f = R_g$, V_{out} will be 2 V, the amount 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:

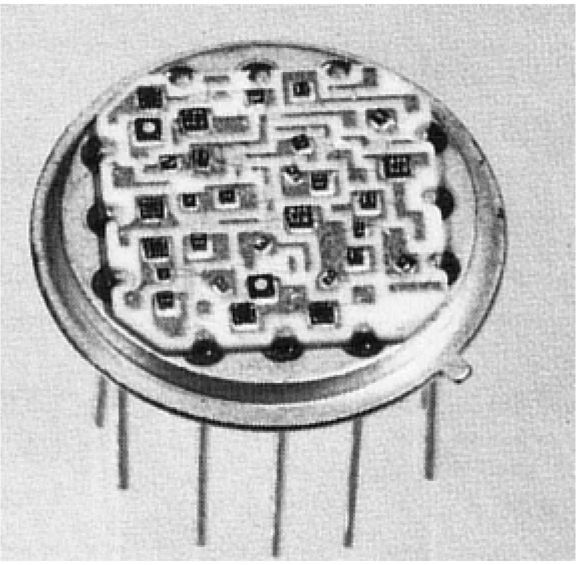
- Infinite open-loop gain (when doing theoretical analysis, a limit may be taken as open loop gain A_{OL} goes to infinity).
- Infinite voltage range available at the output (v_{out}) (in practice the voltages available from the output are limited by the supply voltages $V_{S+}$and V_{S-}). The power 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, $R_{\mathrm{in}} = \infty$, 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 $v_{\mathrm{out}} = 0$).
- 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., $R_{\mathrm{out}} = 0$, 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

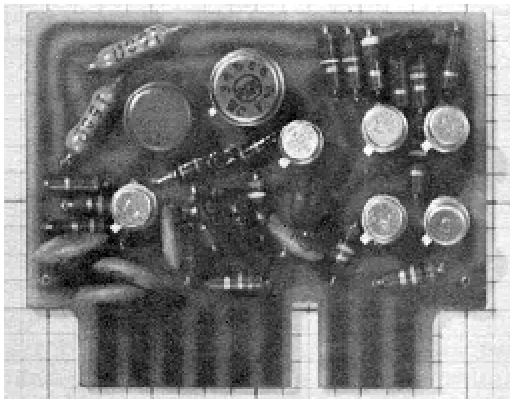
In 1947, the operational amplifier was first formally defined and named in a paper by Professor John R. Ragazzini of Columbia University. In this same paper a footnote mentioned an op-amp design by a student that would turn out to be quite significant. This op-amp, designed by Loebe Julie, was superior 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.

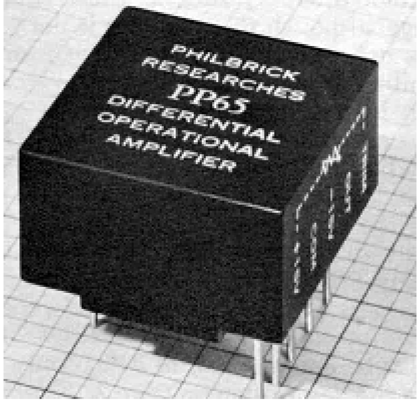
1961: First discrete IC op-amps



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:

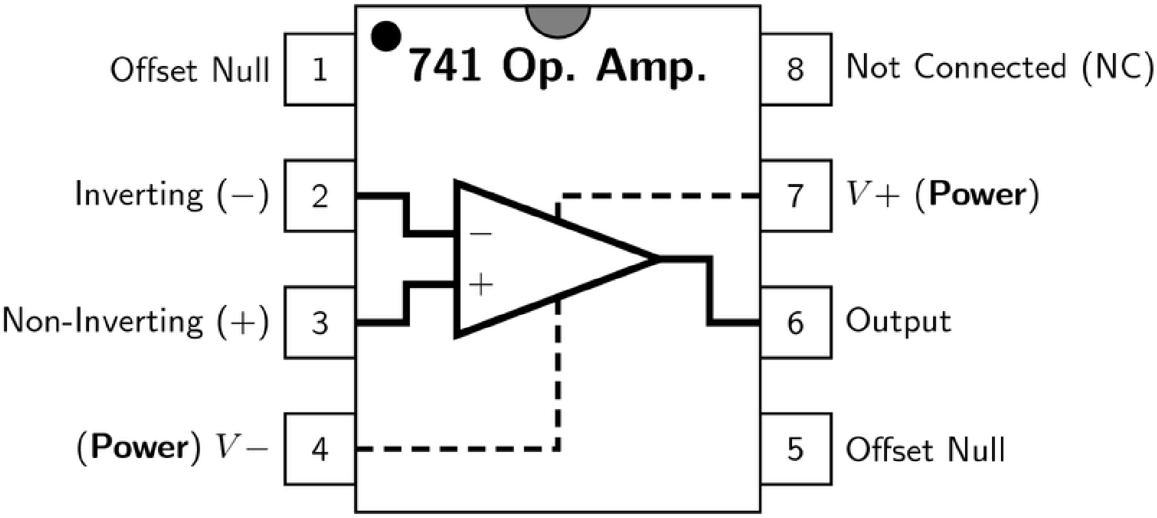
- discrete (built from individual transistors or tubes/valves)
- IC (fabricated in an Integrated circuit) - most common
- hybrid

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.

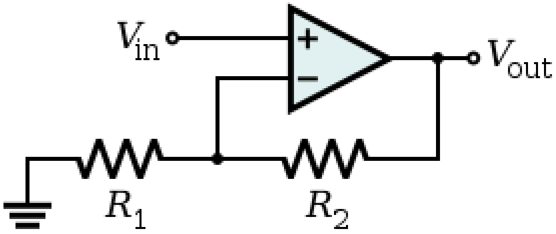
Circuit design follows the same lines for all electronic circuits. A specification is drawn up governing what the circuit is required to do, with allowable limits. For example, the gain may be required to be 100 times, with a tolerance of 5% but drift of less than 1% in a specified temperature range; 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:

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

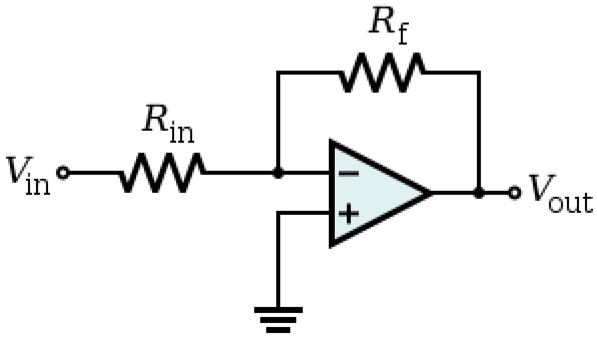
Solving for V_{out}:

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

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:

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

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:

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

Substituting this into the gain equation and solving for V_out:

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

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

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

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.

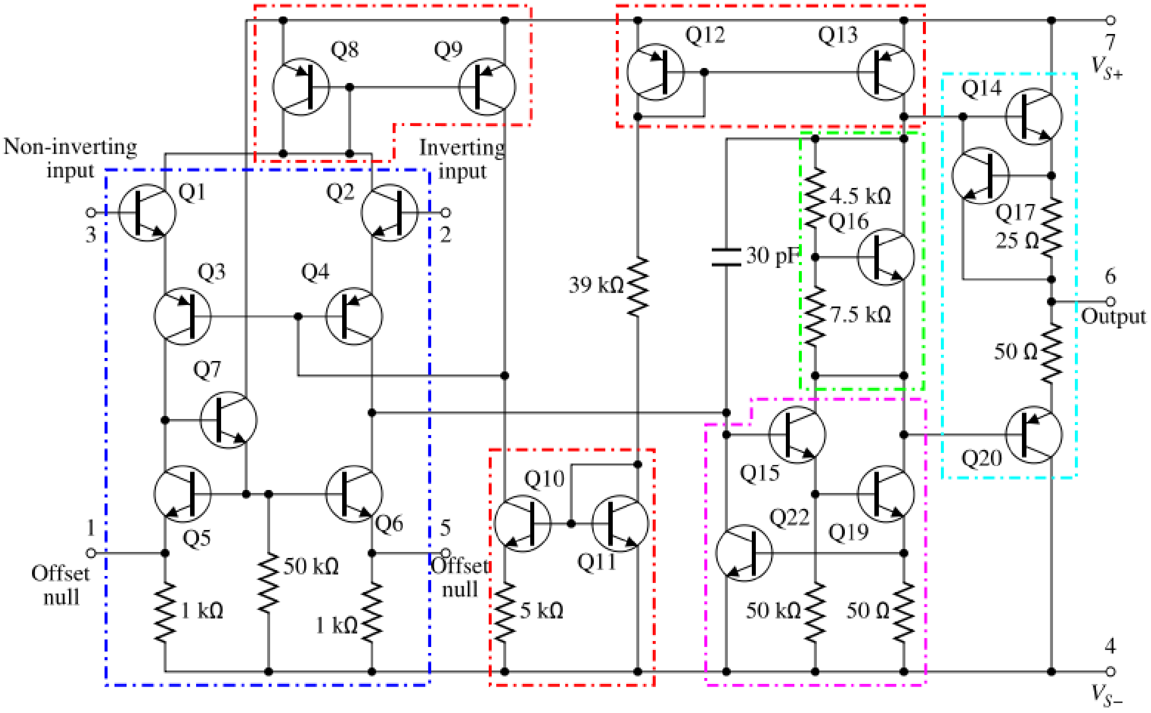
Limited dissipated power

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.

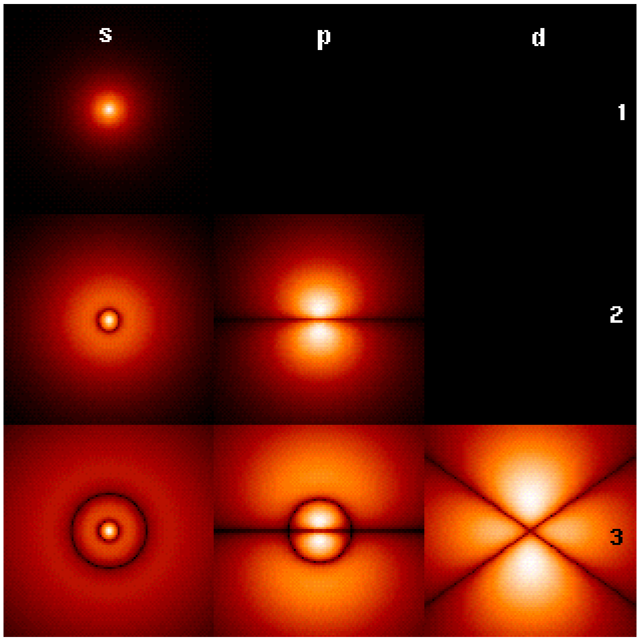
Silicon is used to create most semiconductors commercially. Dozens of other materials are used, including germanium, gallium arsenide, and silicon carbide. A pure semiconductor is often called an “intrinsic” semiconductor. The electronic properties and the conductivity of a semiconductor can be changed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

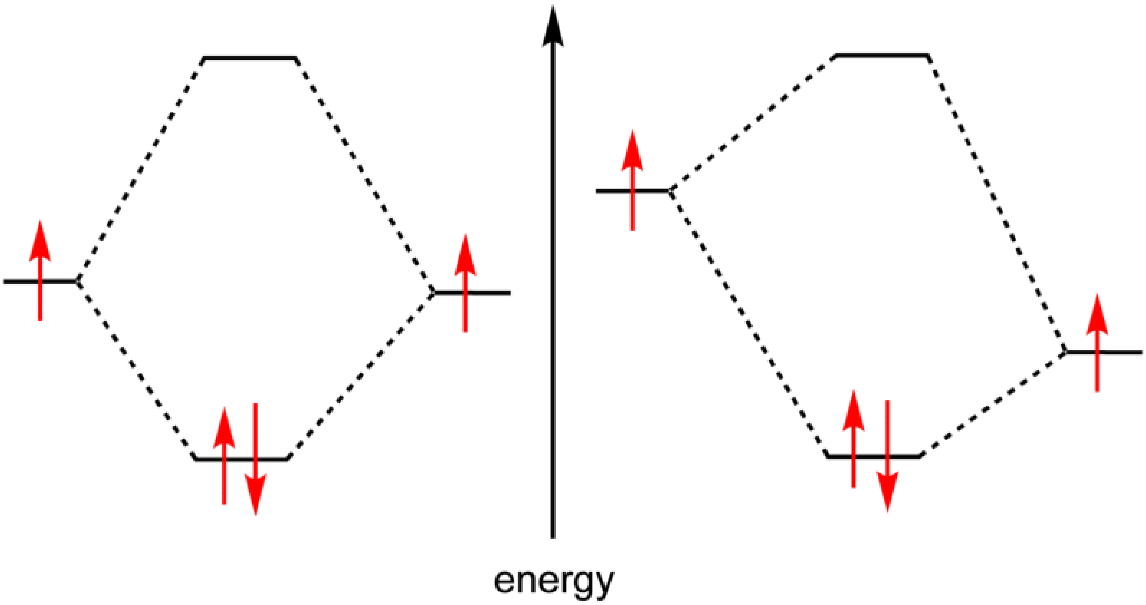
There are three popular ways to classify the electronic structure of a crystal.

- Band structure

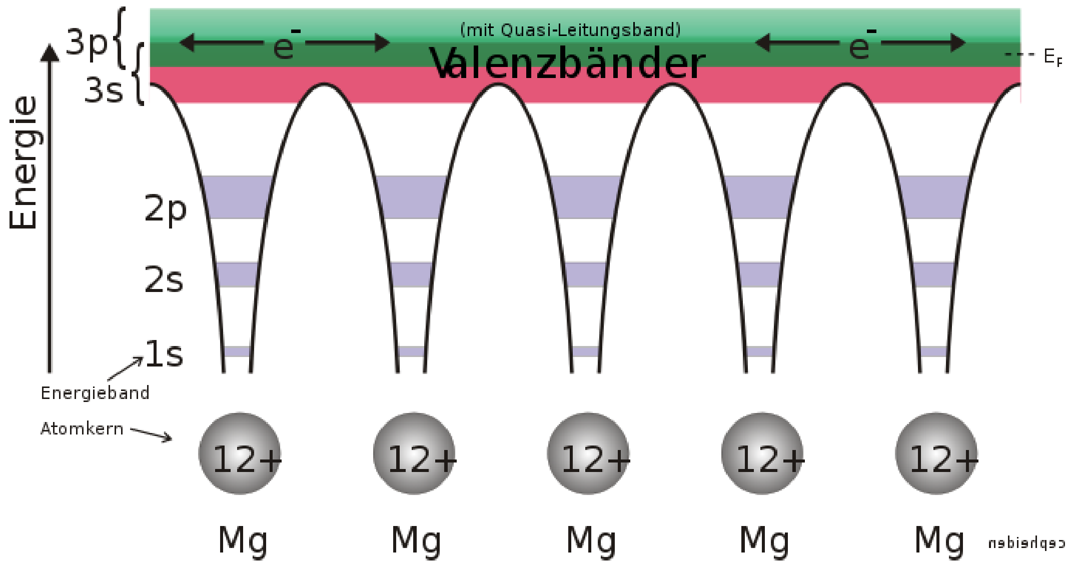
atoms – crystal – vacuum



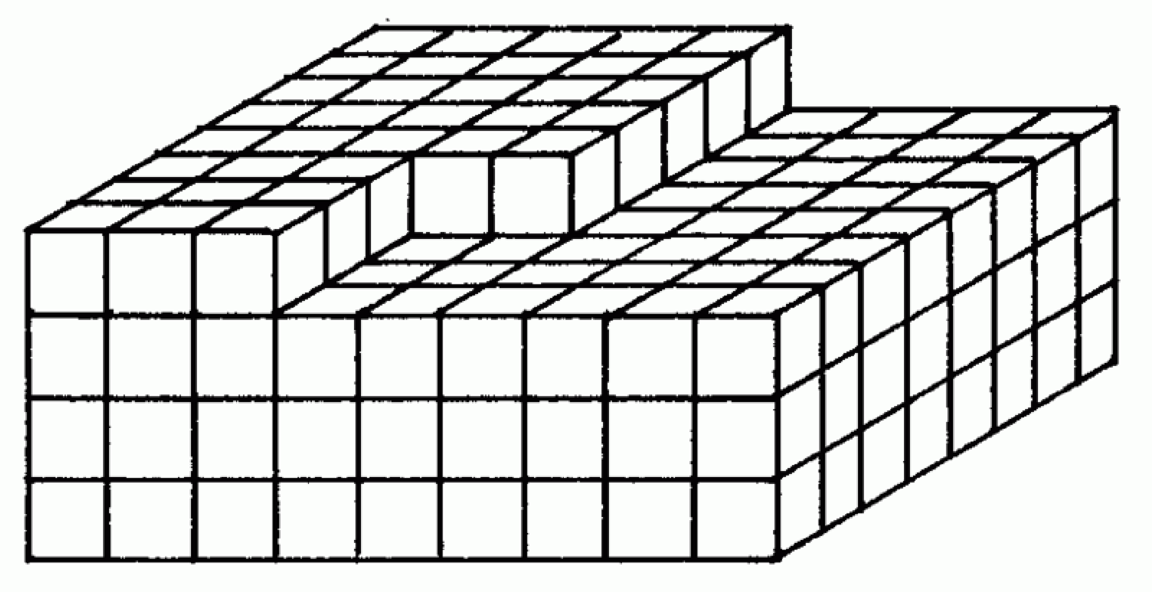
In a single H-atom an electron resides in well known orbitals. Note that the orbitals are called s,p,d in order of increasing circular current.



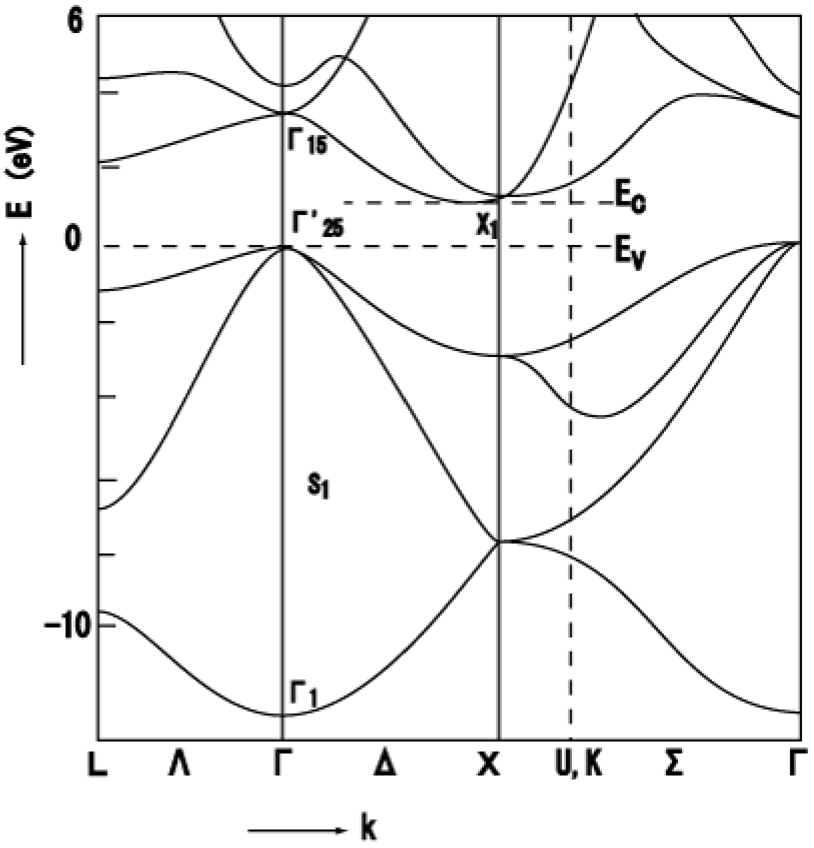
Putting two atoms together leads to delocalized orbitals across two atoms, yielding a covalent bond. Due to the Pauli exclusion principle, every state can contain only one electron.



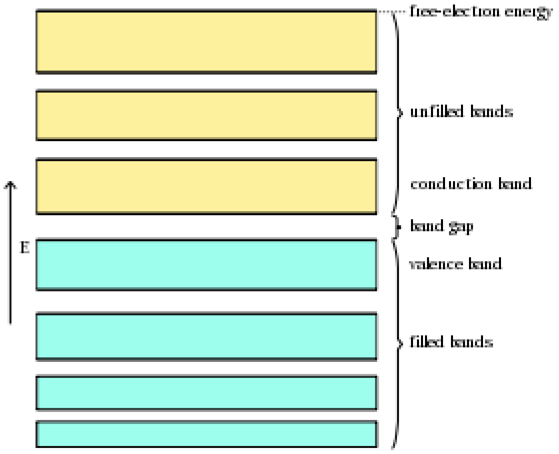
This can be continued with more atoms. Note: This picture shows a metal, not an actual semiconductor.



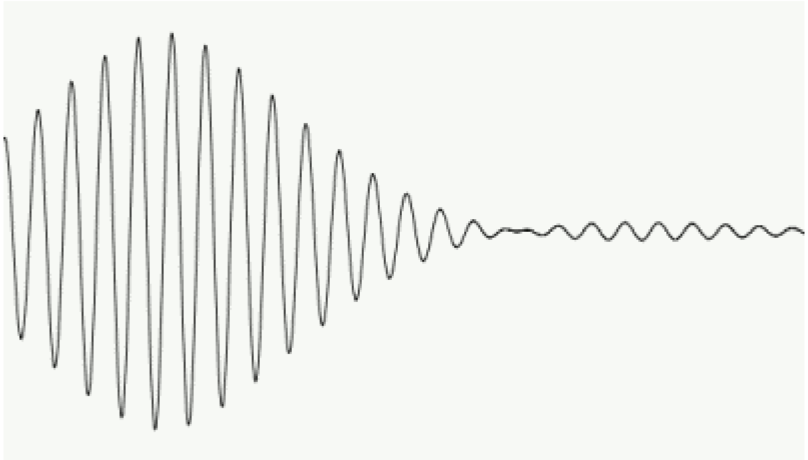
Continuing to add creates a crystal, which may then be cut into a tape and fused together at the ends to allow circular currents.



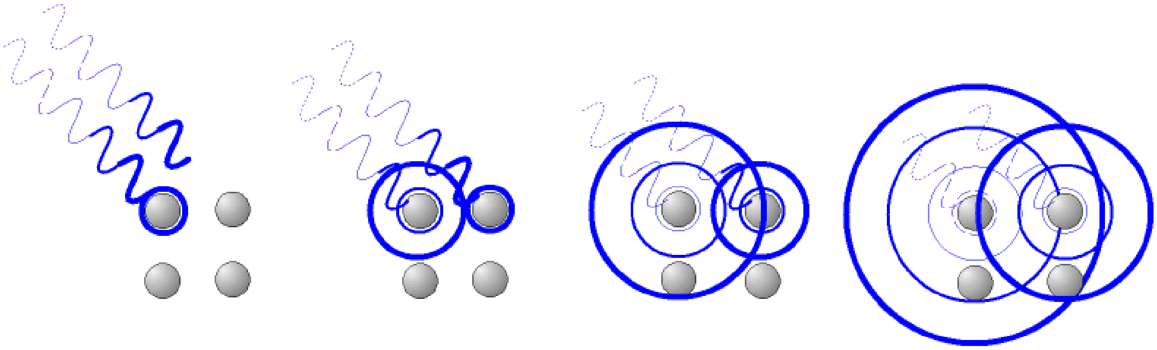
For this regular solid the band structure can be calculated or measured.



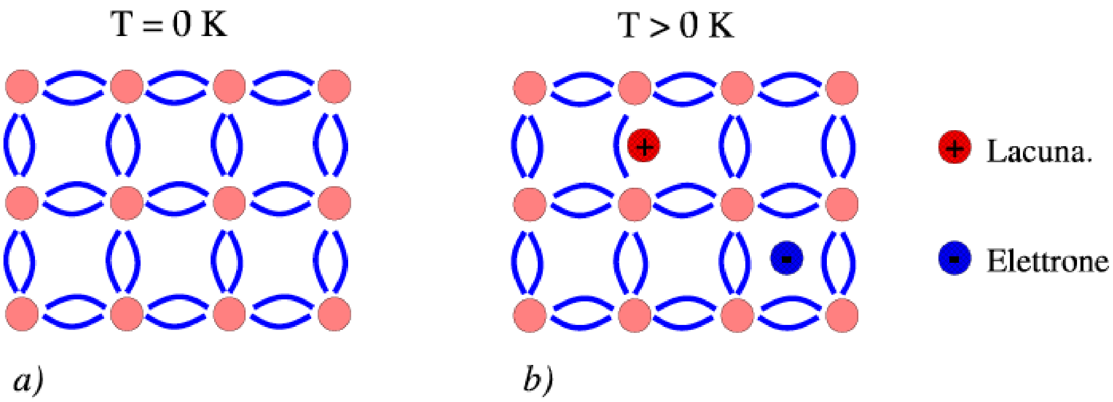
Integrating over the k axis gives the bands of a semiconductor showing a full valence band and an empty conduction band. Generally stopping at the vacuum level is undesirable, because some people want to calculate: photoemission, inverse photoemission



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.

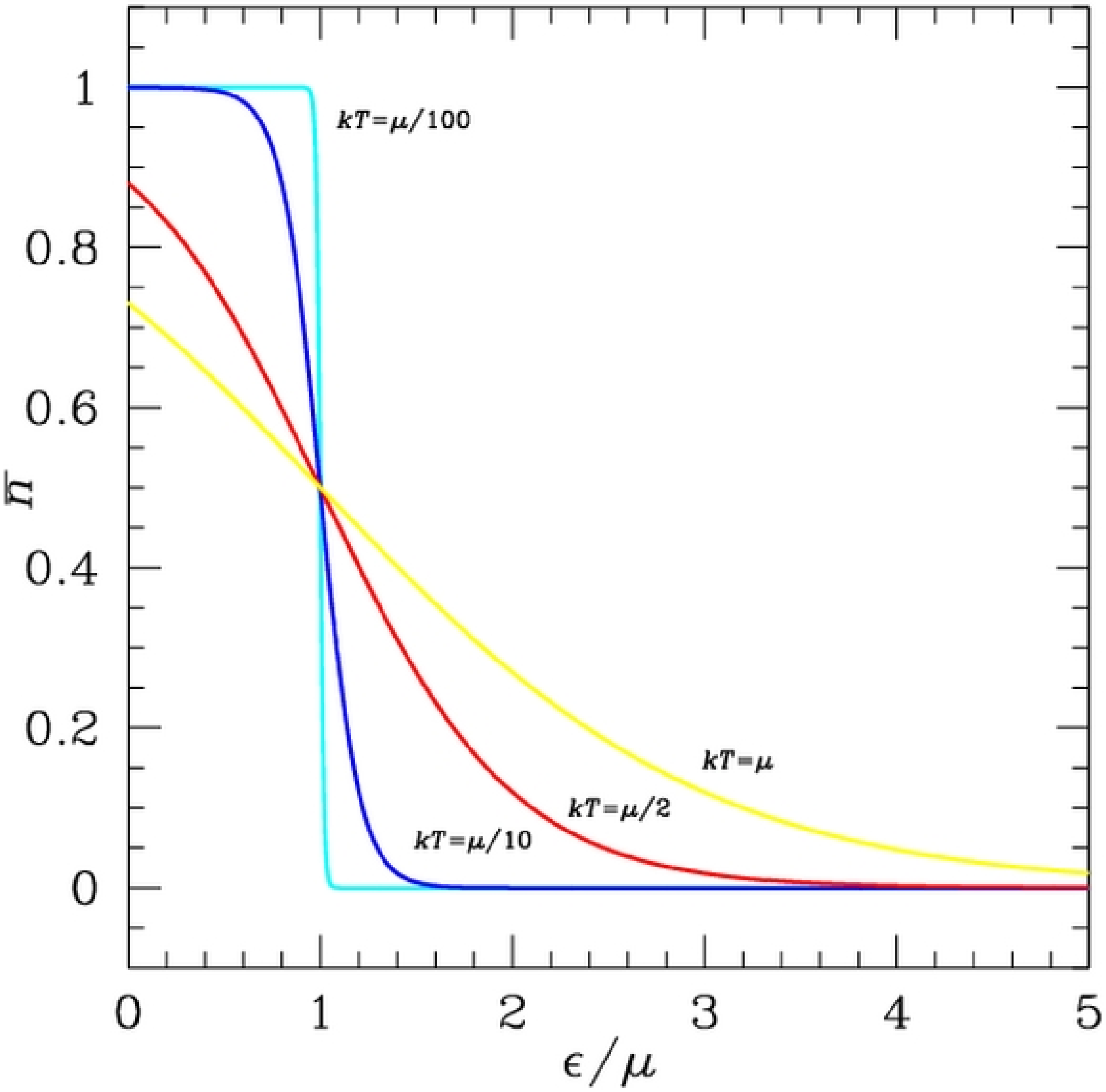
With covalent bonds, an electron moves by hopping to a neighboring bond. The Pauli exclusion principle requires the electron to be lifted into the higher anti-bonding state of that bond. For delocalized states, for example in one dimension – that is in a nanowire, for every energy there is a state with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

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

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

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

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

Carrier generation and recombination

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

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

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

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

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

Semi-insulators

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

Doping

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

Dopants

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

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

Carrier concentration

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

$$n = p = n_i.$$

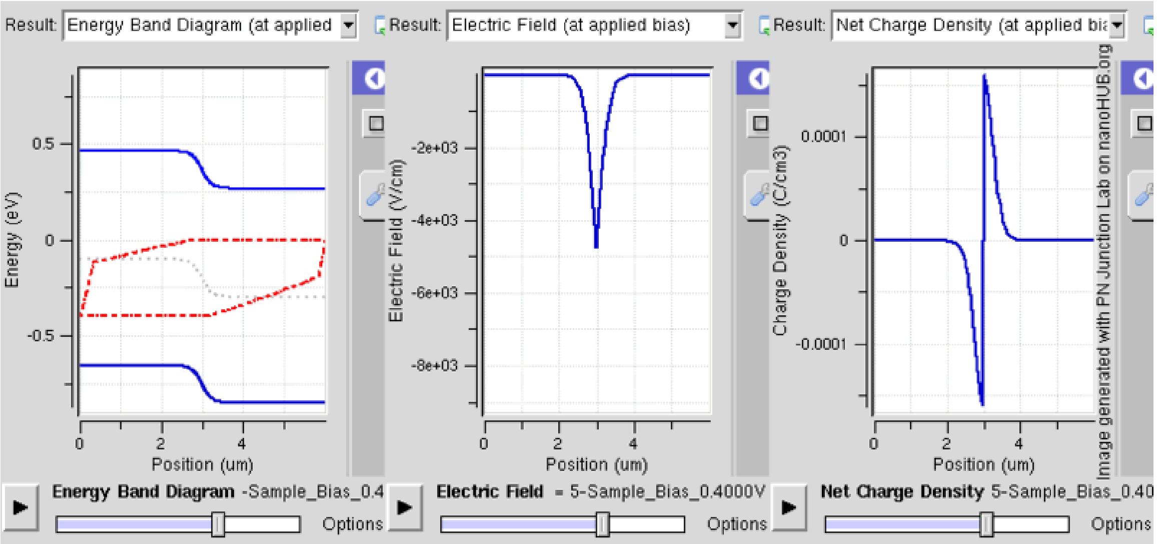
If we have a non-intrinsic semiconductor in thermal equilibrium the relation becomes:

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

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

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

Effect on band structure


Result: Energy Band Diagram (at applied
Result: Electric Field (at applied bias)
Result: Net Charge Density (at applied bia
Energy Band Diagram -Sample_Bias_0.4
Electric Field = 5-Sample_Bias_0.4000V
Net Charge Density 5-Sample_Bias_0.40

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

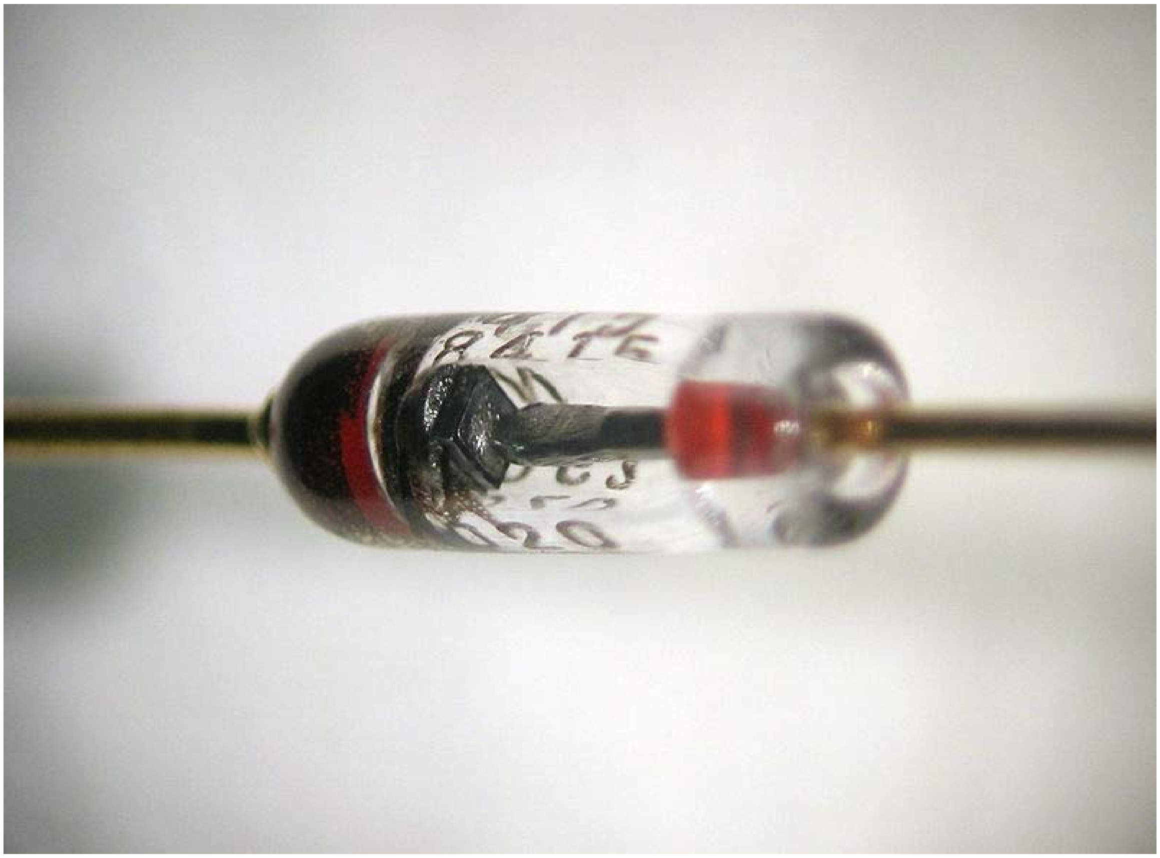


Figure 1: Closeup of a diode, showing the square shaped semiconductor crystal *(black object on left)*.

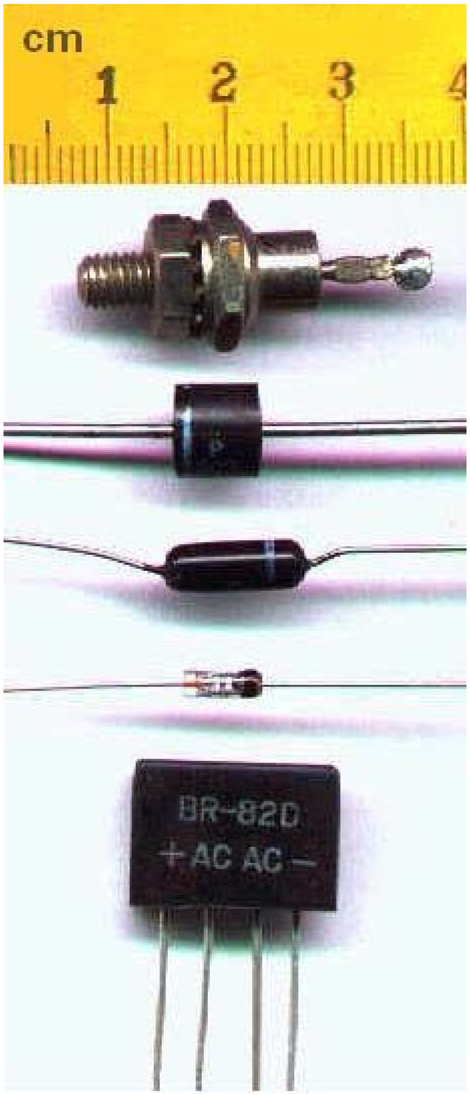


Figure 2: Various semiconductor diodes. Bottom: A bridge rectifier. In most diodes, a white or black painted band identifies the cathode terminal, that is, the terminal which conventional current flows out of when the diode is conducting.

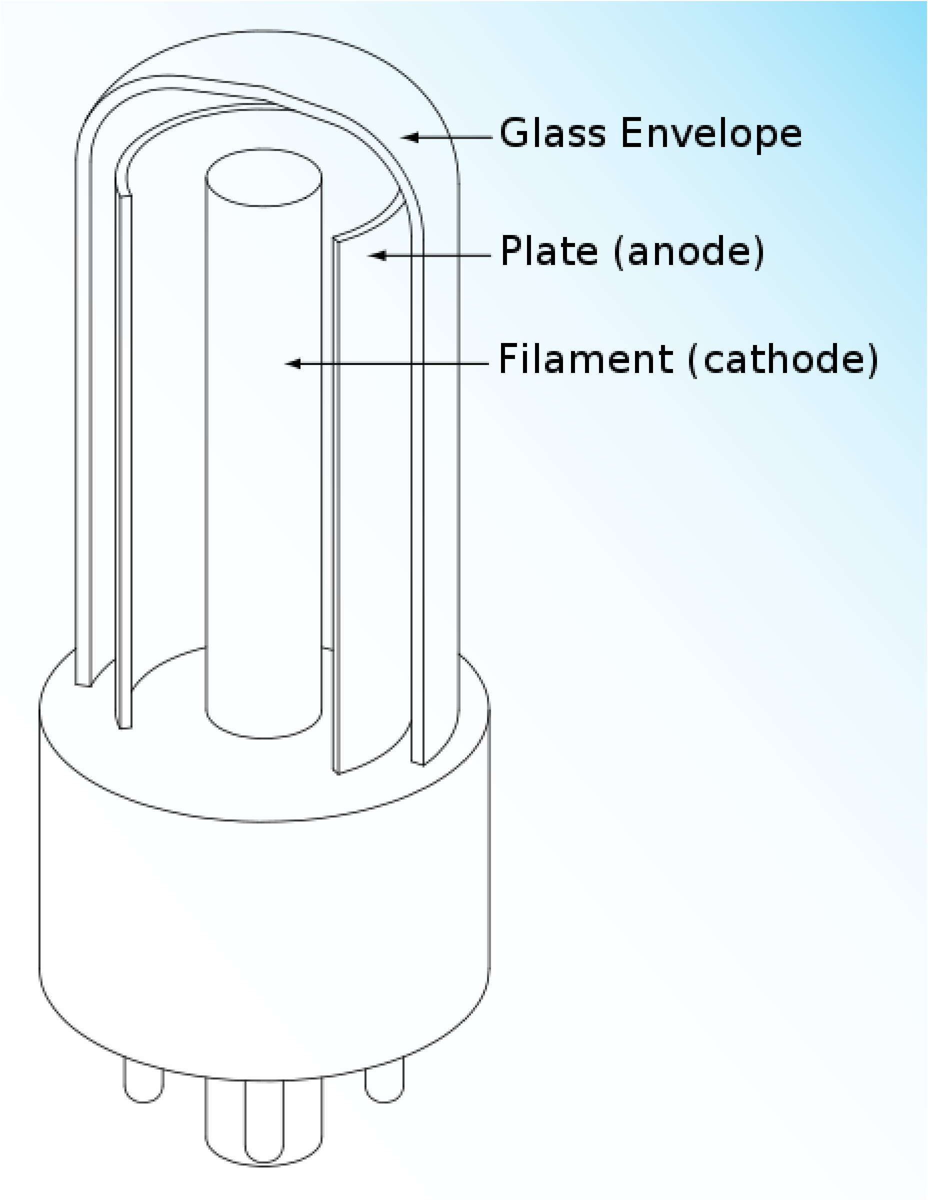


Figure 3: Structure of a vacuum tube diode. The filament may be bare, or more commonly (as shown here), embedded within and insulated from an enclosing cathode

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 *ὅδος*), meaning “path”.

Thermionic and gaseous state diodes

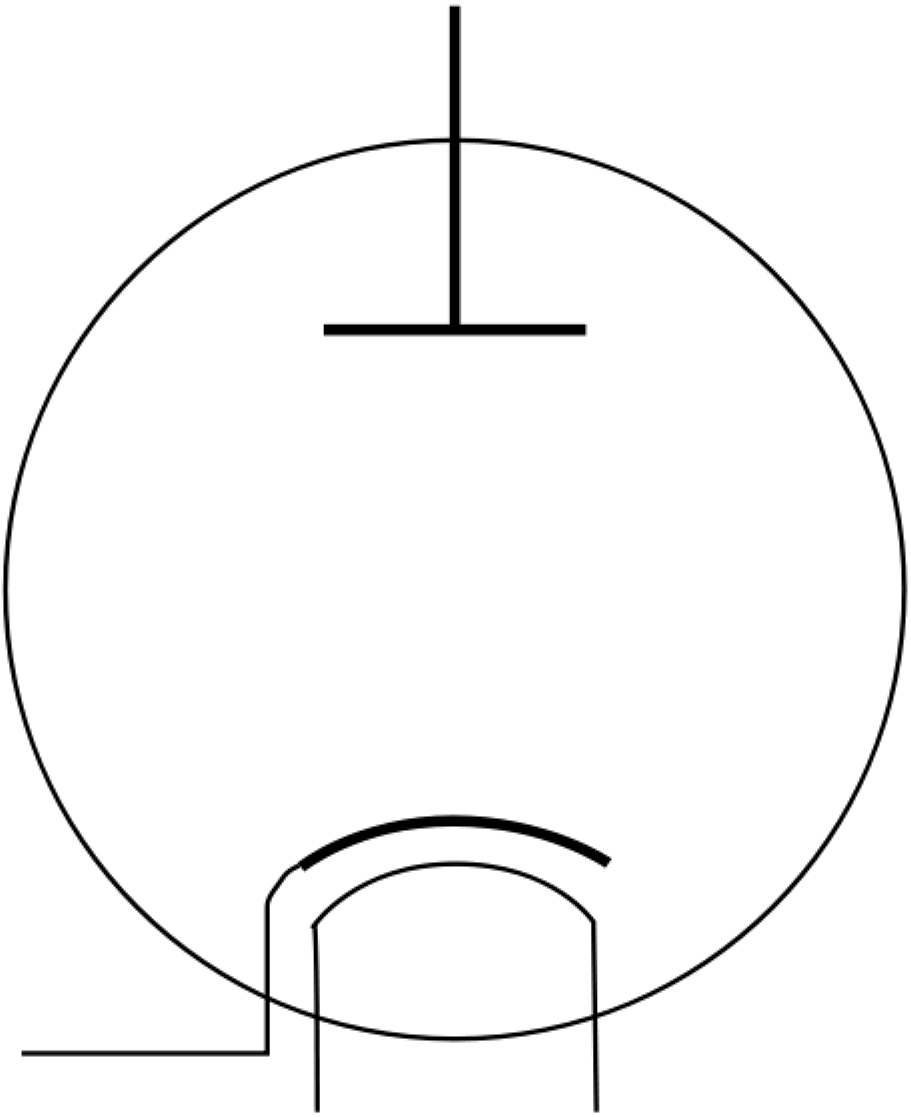


Figure 4: The symbol for an indirect heated vacuum tube diode. From top to bottom, the components are the anode, the cathode, and the heater filament.

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

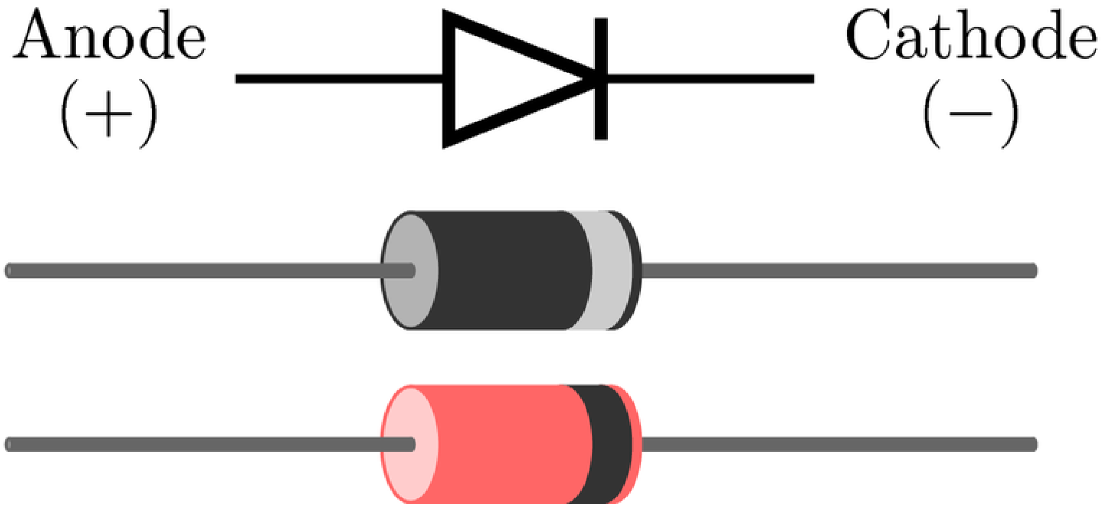


Figure 7: Typical diode packages in same alignment as diode symbol. Thin bar depicts the cathode.

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.

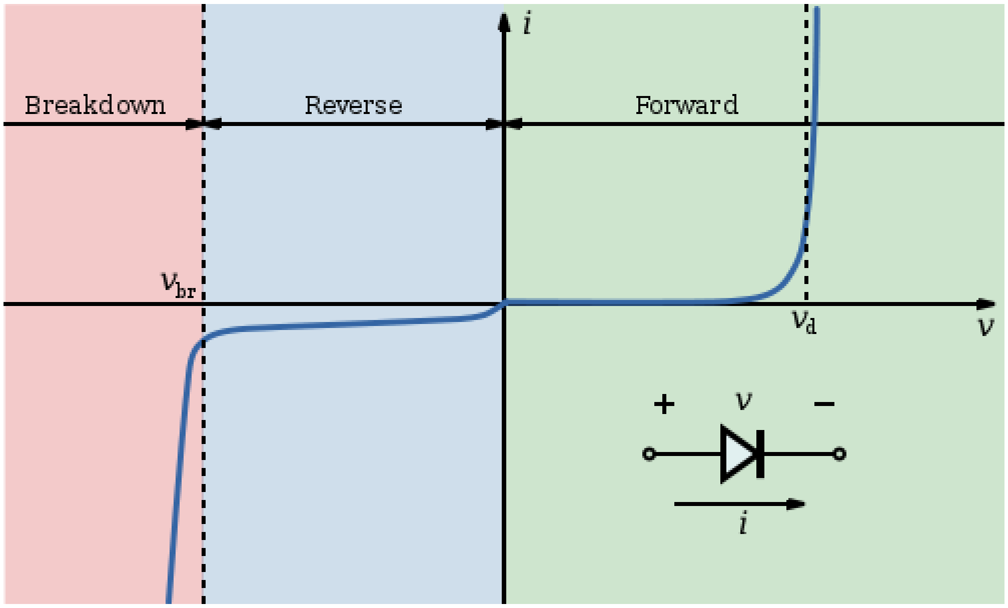


Figure 5: I–V characteristics of a P-N junction diode (not to scale).

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

$$I = I_{\mathrm{S}} e^{V_{\mathrm{D}}/(nV_{\mathrm{T}})}$$

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

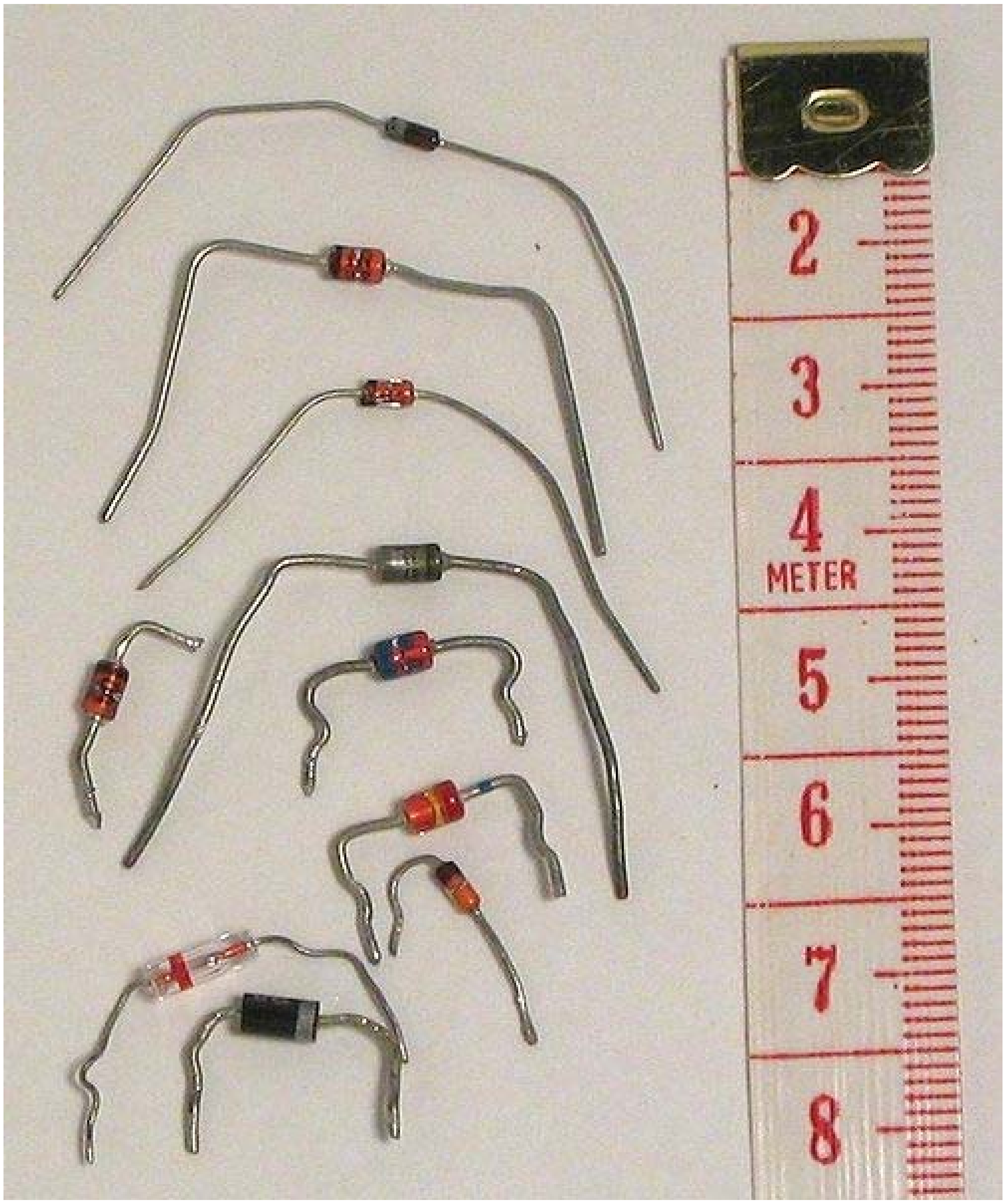


Figure 8: Several types of diodes. The scale is centimeters.

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

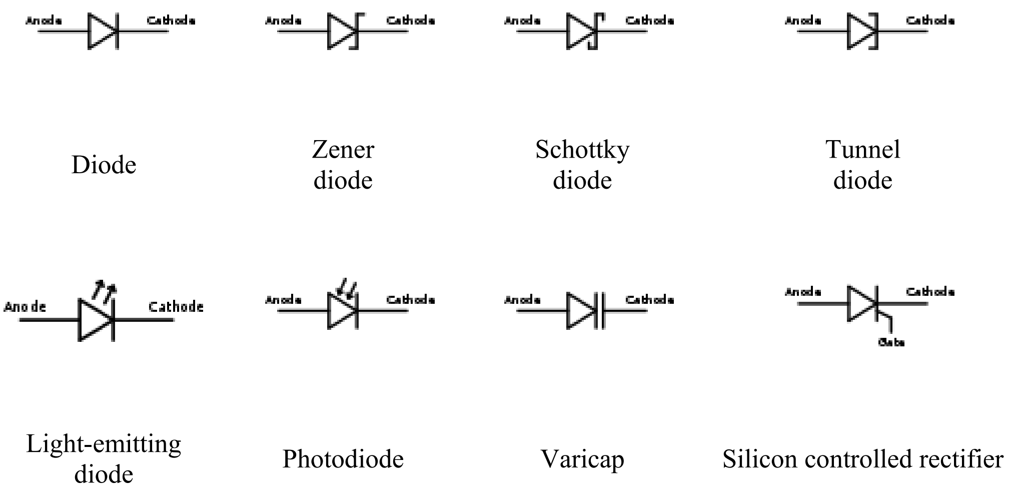
These are used as voltage-controlled capacitors. These are important in PLL (phase-locked loop) and FLL (frequency-locked loop) circuits, allowing tuning circuits, such as those in television receivers, to lock quickly, replacing older designs that took a long time to warm up and lock. A PLL is faster 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.



Diode

Zener diode

Schottky diode

Tunnel diode

Light-emitting diode

Photodiode

Varicap

Silicon controlled rectifier

Figure 6: Some diode symbols.

Numbering and coding schemes

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

EIA/JEDEC

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

Pro Electron

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

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

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

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

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

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

Related devices

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

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

Applications

Radio demodulation

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

Power conversion

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

Over-voltage protection

Diodes are frequently used to conduct damaging high voltages away from sensitive electronic devices. They are usually reverse-biased (non-conducting) under normal circumstances. When the voltage rises above the normal range, the diodes become forward-biased (conducting). For example, diodes are used in (stepper motor and H-bridge) motor controller and relay circuits to de-energize coils rapidly without the damaging voltage spikes that would otherwise occur. (Any diode used in such an application is called a flyback diode). Many integrated circuits also incorporate diodes on the connection pins to prevent external voltages from damaging their sensitive transistors. Specialized diodes are used to protect from over-voltages at higher power.

Logic gates

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

Ionizing radiation detectors

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

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

Temperature measurements

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

Current steering

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

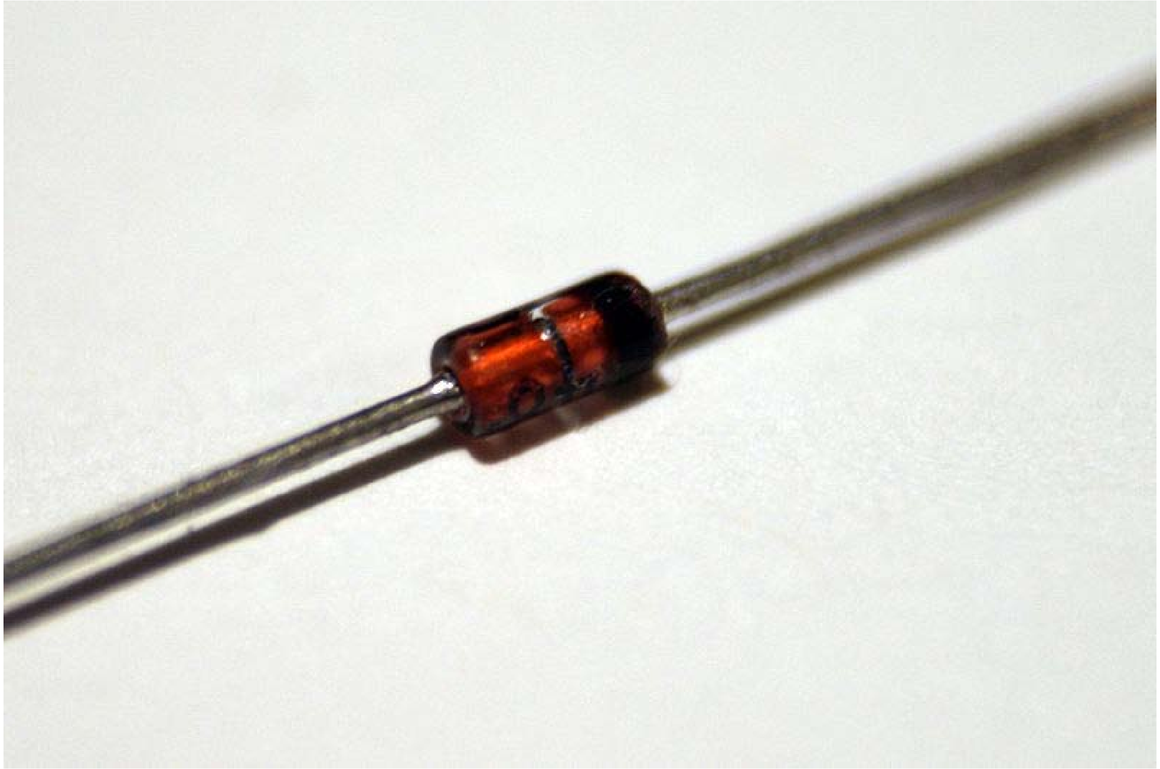
Diodes are also used in electronic musical keyboards. To reduce the amount of wiring needed in electronic musical keyboards, these instruments often use keyboard matrix circuits. The keyboard controller scans the rows and columns to determine which note the player has pressed. The problem with matrix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

Abbreviations

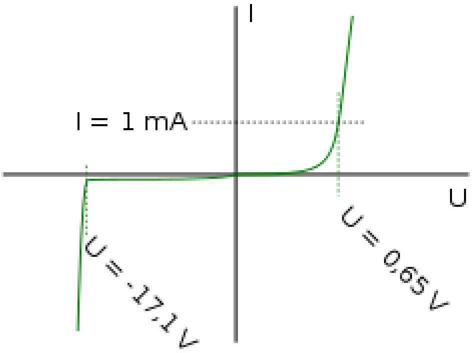
Diodes are usually referred to as *D* for diode on PCBs. Sometimes the abbreviation *CR* for **crystal rectifier** is used.

Chapter- 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.

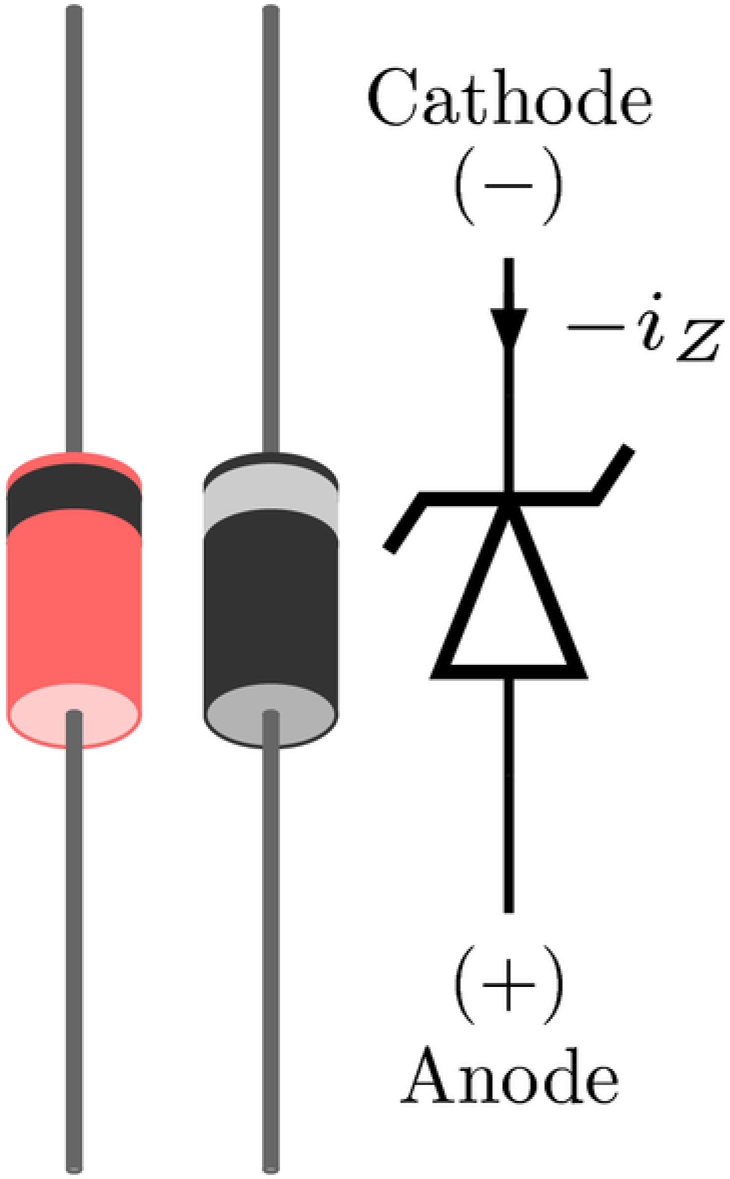
A Zener diode exhibits almost the same properties, except the device is specially designed so as to have a greatly reduced breakdown voltage, the so-called Zener voltage. By contrast with the conventional device, a reverse-biased Zener diode will exhibit a controlled breakdown and allow the current to 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.

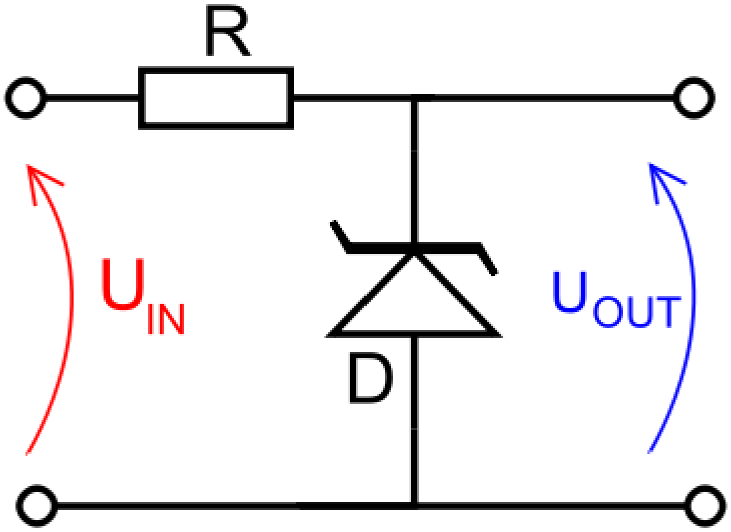
All such diodes, regardless of breakdown voltage, are usually marketed under the umbrella term of "Zener diode".

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.

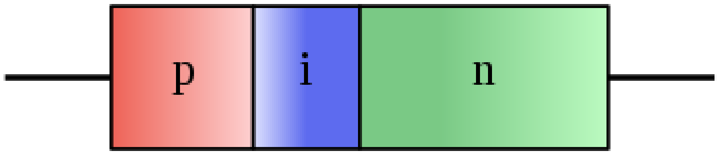
These devices are also encountered, typically in series with a base-emitter junction, in transistor stages where selective choice of a device centered around the avalanche/Zener point can be used to introduce compensating temperature co-efficient balancing of the transistor PN junction. An example of 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

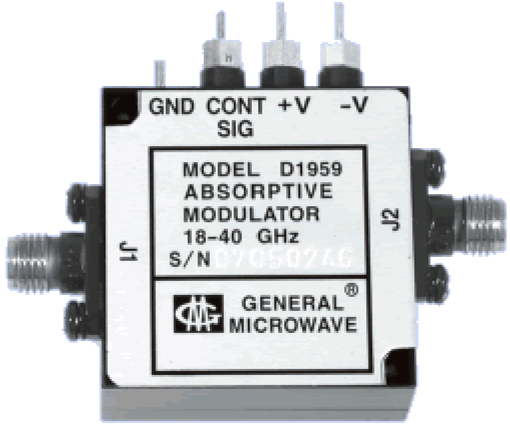
Under zero or reverse bias, a PIN diode has a low capacitance. The low capacitance will not pass much of an RF signal. Under a forward bias of 1 mA, a typical PIN diode will have an RF resistance of about 1 ohm, making it a good RF conductor. Consequently, the PIN diode makes a good RF switch.

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.

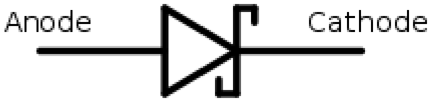
Typically, amorphous silicon thin-film cells use PIN structures. On the other hand, CdTe cells use NIP structure, a variation of the PIN structure. In a NIP structure, an intrinsic CdTe layer is sandwiched by n-doped CdS and p-doped ZnTe. The photons are incident on the n-doped layer unlike a PIN diode.

Example Diodes

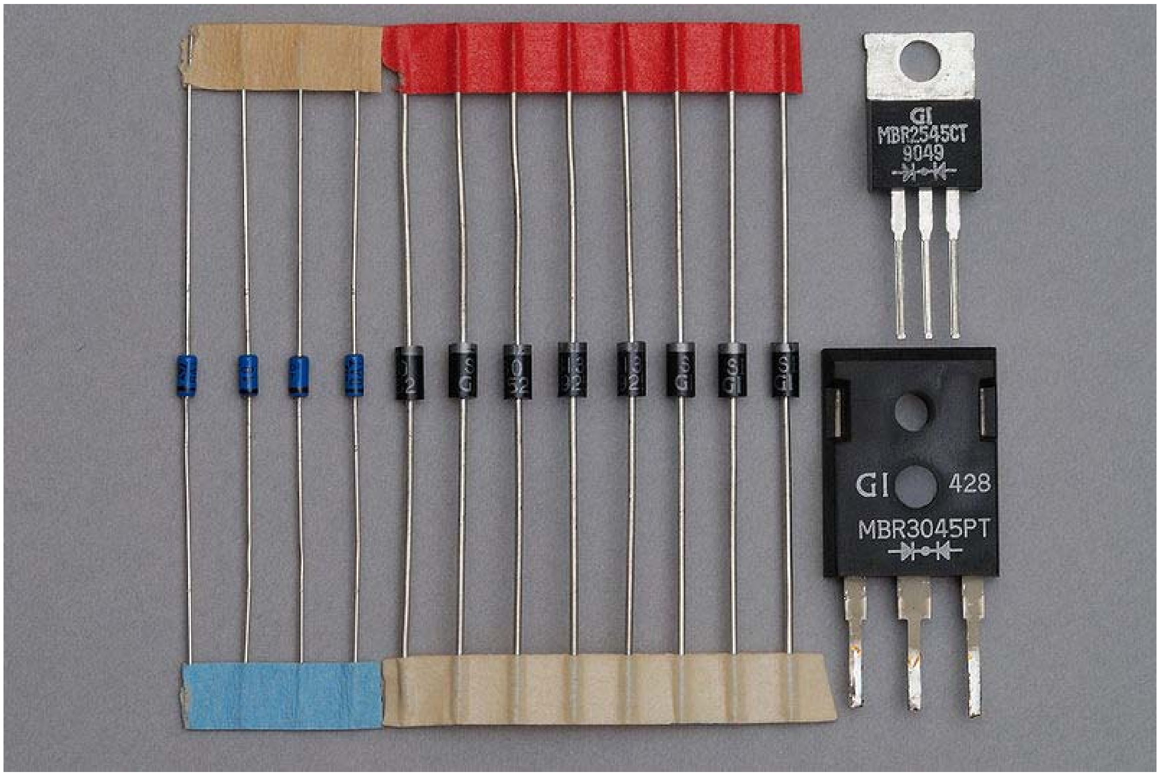
SFH203 or BPW43 are cheap general purpose PIN diodes in 5 mm clear plastic case with bandwidth over 100 MHz. They are used in RONJA telecommunication systems and other circuitry applications.

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.

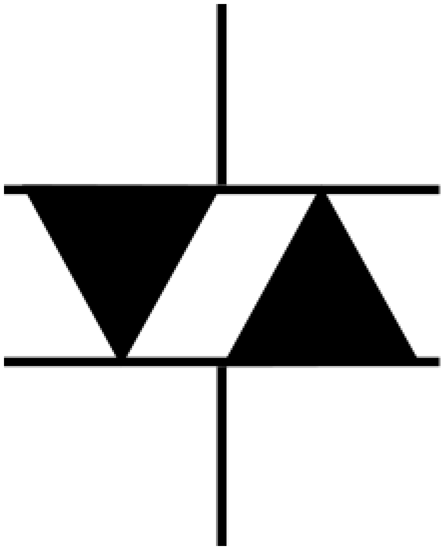
RF noise generation

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.

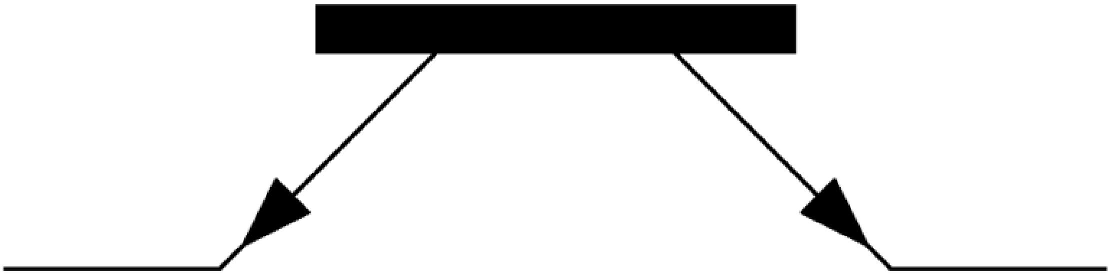
Microwave frequency generation

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.

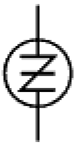
When this occurs, diode enters the region of negative dynamic resistance, leading to a decrease in the voltage drop across the diode and, usually, a sharp increase in current through the diode. The diode remains "in conduction" until the current through it drops below a value characteristic for the device, 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.

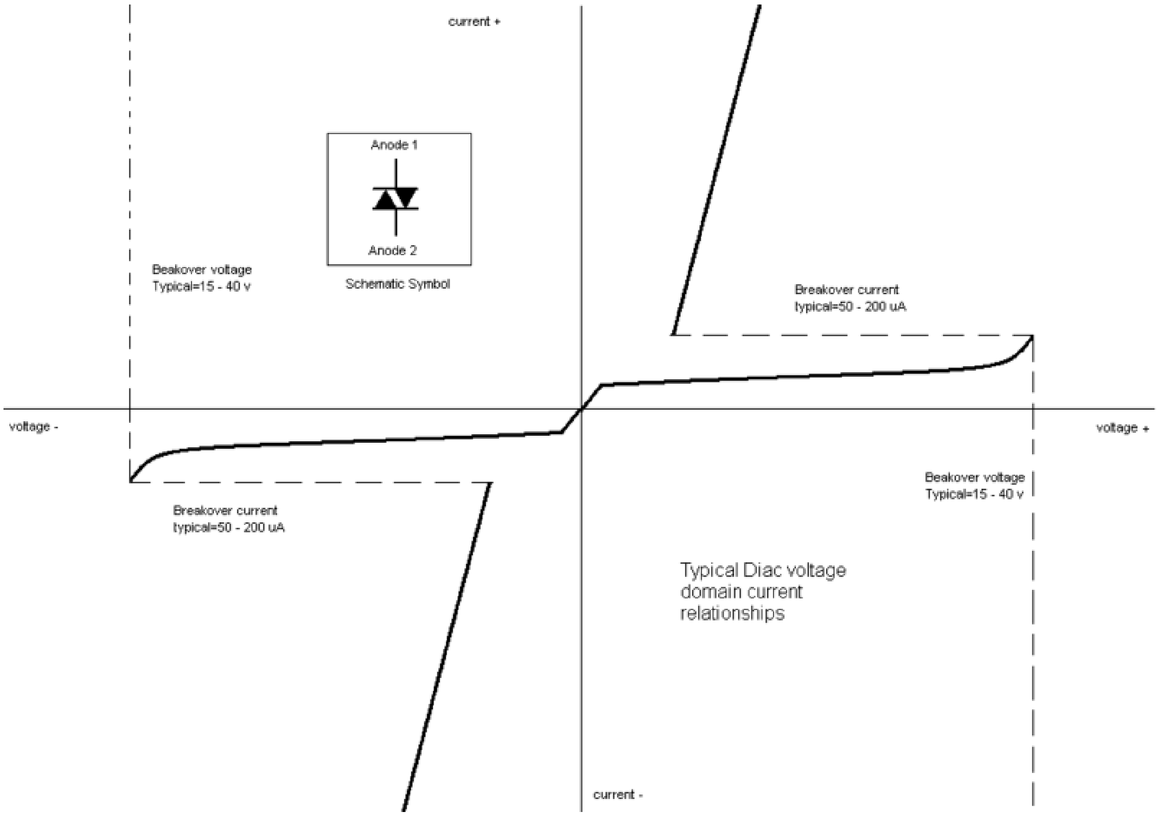
DIACs have no gate electrode, unlike some other thyristors that they are commonly used to trigger, such as TRIACs. Some TRIACs contain a built-in DIAC in series with the TRIAC's "gate" terminal for this purpose.

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.

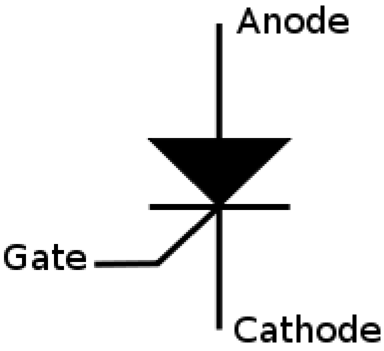
The operation of the SIDAC is functionally similar to that of a spark gap. The SIDAC remains nonconducting until the applied voltage meets or exceeds its rated breakover voltage. Once entering this conductive state going through the negative dynamic resistance region, the SIDAC continues to conduct, 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

Thyristor



Circuit symbol for a thyristor

An SCR rated about 100 amperes, 1200 volts mounted on a heat sink - the two small wires are the gate trigger leads

A **thyristor** is a solid-state semiconductor device with four layers of alternating N and P-type material. They act as bistable switches, conducting when their gate receives a current pulse, and continue to conduct while they are forward biased (that is, while the voltage across the device is not reversed).

Some sources define silicon controlled rectifiers and thyristors as synonymous.

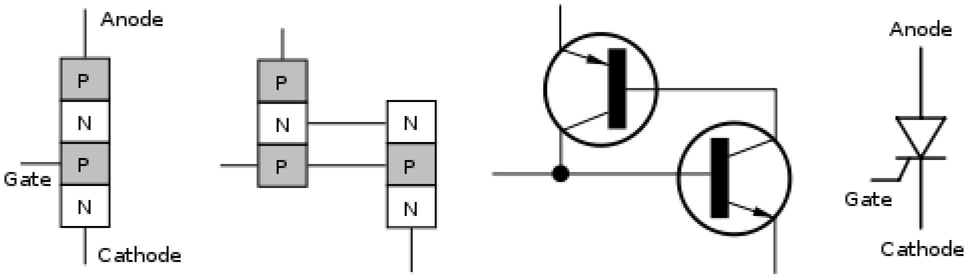
Other sources define thyristors as a larger set of devices with at least four layers of alternating N and P-type material, including:

- Silicon controlled rectifier (SCR)

- Gate turn-off thyristor (GTO)
- Triode AC switch (TRIAC)
- Static Induction Transistor/Thyristor (SIT/SITh)
- MOS Controlled Thyristor (MCT)
- Distributed Buffer - Gate Turn-off Thyristor (DB-GTO)
- Integrated gate commutated thyristor (IGCT)
- MOS composite static induction thyristor/CSMT
- Reverse conducting thyristor

Function

The thyristor is a four-layer, three terminal semiconducting device, with each layer consisting of alternately N-type or P-type material, for example P-N-P-N. The main terminals, labelled anode and cathode, are across the full four layers, and the control terminal, called the gate, is attached to p-type material near to the cathode. (A variant called an SCS—Silicon Controlled Switch—brings all four layers out to terminals.) The operation of a thyristor can be understood in terms of a pair of tightly coupled bipolar junction transistors, arranged to cause the self-latching action:

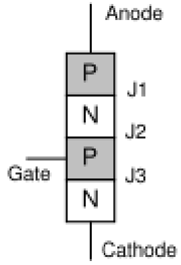


Thyristors have three states:

1. Reverse blocking mode — Voltage is applied in the direction that would be blocked by a diode
2. Forward blocking mode — Voltage is applied in the direction that would cause a diode to conduct, but the thyristor has not yet been triggered into conduction
3. Forward conducting mode — The thyristor has been triggered into conduction and will remain conducting until the forward current drops below a threshold value known as the "holding current"

Function of the gate terminal

The thyristor has three p-n junctions (serially named J_1, J_2, J_3 from the anode).



Layer diagram of thyristor.

When the anode is at a positive potential V_{AK} with respect to the cathode with no voltage applied at the gate, junctions J_1 and J_3 are forward biased, while junction J_2 is reverse biased. As J_2 is reverse biased, no conduction takes place (Off state). Now if V_{AK} is increased beyond the breakdown voltage V_{BO} of the thyristor, avalanche breakdown of J_2 takes place and the thyristor starts conducting (On state).

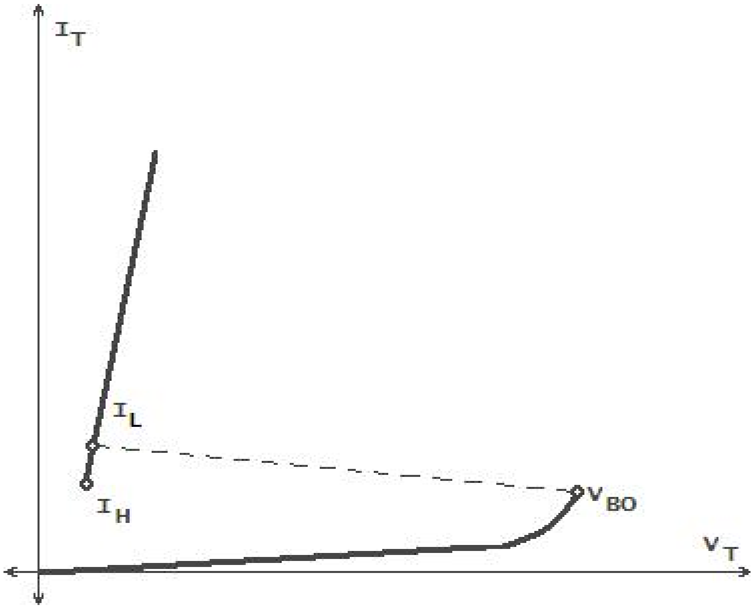
If a positive potential V_G is applied at the gate terminal with respect to the cathode, the breakdown of the junction J_2 occurs at a lower value of V_{AK}. By selecting an appropriate value of V_G, the thyristor can be switched into the on state suddenly.

Once avalanche breakdown has occurred, the thyristor continues to conduct, irrespective of the gate voltage, until: (a) the potential V_G is removed or (b) the current through the device (anode−cathode) is less than the holding current specified by the manufacturer. Hence V_G can be a voltage pulse, such as the voltage output from a UJT relaxation oscillator.

These gate pulses are characterized in terms of gate trigger voltage (V_{GT}) and gate trigger current (I_{GT}). Gate trigger current varies inversely with gate pulse width in such a way that it is evident that there is a minimum gate charge required to trigger the thyristor.

Switching characteristics

In a conventional thyristor, once it has been switched on by the gate terminal, the device remains latched in the on-state (*i.e.* does not need a continuous supply of gate current to conduct), providing the anode current has exceeded the latching current (I_L). As long as the anode remains positively biased, it cannot be switched off until the anode current falls below the holding current (I_H).



V - *I* characteristics.

A thyristor can be switched off if the external circuit causes the anode to become negatively biased. In some applications this is done by switching a second thyristor to discharge a capacitor into the cathode of the first thyristor. This method is called forced commutation.

After a thyristor has been switched off by forced commutation, a finite time delay must have elapsed before the anode can again be positively biased *and* retain the thyristor in the off-state. This minimum delay is called the circuit commutated turn off time (t_Q). Attempting to positively bias the anode within this time causes the thyristor to be self-triggered by the remaining charge carriers (holes and electrons) that have not yet recombined.

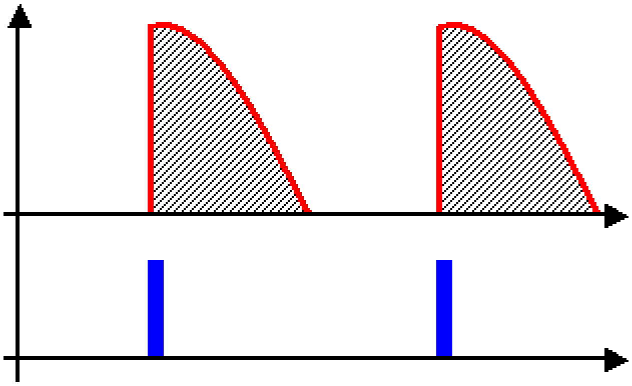
For applications with frequencies higher than the domestic AC mains supply (e.g. 50 Hz or 60 Hz), thyristors with lower values of t_Q are required. Such fast thyristors are made by diffusing into the silicon heavy metals ions such as gold or platinum which act as charge combination centres. Alternatively, fast thyristors may be made by neutron irradiation of the silicon.

History

The Silicon Controlled Rectifier (SCR) or Thyristor proposed by William Shockley in 1950 and championed by Moll and others at Bell Labs was developed in 1956 by power engineers at General Electric (G.E.) led by Gordon Hall and commercialized by G.E.'s Frank W. "Bill" Gutzwiller.

A bank of six 2000 A Thyristors (white pucks).

Applications



Load voltage regulated by thyristor phase control.
Red trace: load voltage
Blue trace: trigger signal.

Thyristors are mainly used where high currents and voltages are involved, and are often used to control alternating currents, where the change of polarity of the current causes the device to switch off automatically; referred to as Zero Cross operation. The device can be said to operate *synchronously* as, once the device is open, it conducts current in phase with the voltage applied over its cathode to anode junction with no further gate modulation being required to replicate; the device is biased *fully on*. This is not to be confused with symmetrical operation, as the output is unidirectional, flowing only from cathode to anode, and so is asymmetrical in nature.

Thyristors can be used as the control elements for phase angle triggered controllers, also known as phase fired controllers.

They can also be found in power supplies for digital circuits, where they are used as a sort of "circuit breaker" or "crowbar" to prevent a failure in the power supply from damaging downstream components. A thyristor is used in conjunction with a zener diode attached to its gate, and when the output voltage of the supply rises above the zener voltage, the thyristor will conduct, then short-circuit the power supply output to ground (and in general blowing an upstream fuse).

The first large scale application of thyristors, with associated triggering diac, in consumer products related to stabilized power supplies within color television receivers in the early 1970s. The stabilized high voltage DC supply for the receiver was obtained by moving the switching point of the thyristor device up and down the falling slope of the positive going half of the AC supply input (if the rising slope was used the output voltage would always rise towards the peak input voltage when the device was triggered and thus defeat the aim of regulation). The precise switching point was determined by the load on the output DC supply as well fluctuations on the input AC supply.

Thyristors have been used for decades as lighting dimmers in television, motion pictures, and theater, where they replaced inferior technologies such as autotransformers and rheostats. They have also been used in photography as a critical part of flashes (strobes).

Snubber circuits

Thyristors can be triggered by a high rate of rise of off-state voltage. This is prevented by connecting a resistor-capacitor (RC) snubber circuit between the anode and cathode terminals in order to limit the dV/dt (i.e., rate of change of voltage versus time).

HVDC electricity transmission

Two of three thyristor valve stacks used for long distance transmission of power from Manitoba Hydro dams

Since modern thyristors can switch power on the scale of megawatts, thyristor valves have become the heart of high-voltage direct current (HVDC) conversion either to or from alternating current. In the realm of this and other very high power applications, both electronically switched (ETT) and light switched (LTT) thyristors are still the primary choice. The valves are arranged in stacks usually suspended from the ceiling of a transmission building called a valve hall. Thyristors are arranged into a Graetz bridge circuit and to avoid harmonics are connected in series to form a 12 pulse converter. Each thyristor is cooled with deionized water, and the entire arrangement becomes one of multiple identical modules forming a layer in a multilayer valve stack called a *quadruple* *valve*. Three such stacks are typically hung from the ceiling of the valve building of a long distance transmission facility.

Comparisons to other devices

The functional drawback of a thyristor is that, like a diode, it only conducts in one direction. A similar self-latching 5-layer device, called a TRIAC, is able to work in both directions. This added capability, though, also can become a shortfall. Because the TRIAC can conduct in both directions, reactive loads can cause it to fail to turn off during the zero-voltage instants of the ac power cycle. Because of this, use of TRIACs with (for example) heavily-inductive motor loads usually requires the use of a "snubber" circuit around the TRIAC to assure that it will turn off with each half-cycle of mains power. Inverse parallel SCRs can also be used in place of the triac; because each SCR in the pair has an entire half-cycle of reverse polarity applied to it, the SCRs, unlike TRIACs, are sure to turn off. The "price" to be paid for this arrangement, however, is the added complexity of two separate but essentially identical gating circuits.

Etymology

An earlier gas filled tube device called a Thyratron provided a similar electronic switching capability, where a small control voltage could switch a large current. It is from a combination of "thyratron" and "transistor" that the term "thyristor" is derived.

Although thyristors are heavily used in megawatt scale rectification of AC to DC, in low and medium power (from few tens of watts to few tens of kilowatts) they have almost been replaced by other devices with superior switching characteristics like MOSFETs or IGBTs. One major problem associated with SCRs is that they are not fully controllable switches. The GTO (Gate Turn-off Thyristor) and IGCT are two related devices which address this problem. In high-frequency applications, thyristors are poor candidates due to large switching times arising from bipolar conduction. MOSFETs, on the other hand, have much faster switching capability because of their unipolar conduction (only majority carriers carry the current).

Failure modes

As well as the usual failure modes due to exceeding voltage, current or power ratings, thyristors have their own particular modes of failure, including:

- **Turn on di/dt** — in which the rate of rise of on-state current after triggering is higher than can be supported by the spreading speed of the active conduction area (SCRs & triacs).
- **Forced commutation** — in which the transient peak reverse recovery current causes such a high voltage drop in the sub-cathode region that it exceeds the reverse breakdown voltage of the gate cathode diode junction (SCRs only).

- **Switch on dv/dt** — the thyristor can be spuriously fired without trigger from the gate if the rate of rise of voltage anode to cathode is too great.

Silicon carbide thyristors

In recent years, some manufacturers have developed thyristors using Silicon carbide (SiC) as the semiconductor material. These have applications in high temperature environments, being capable of operating at temperatures up to 350 °C.

Types of thyristor

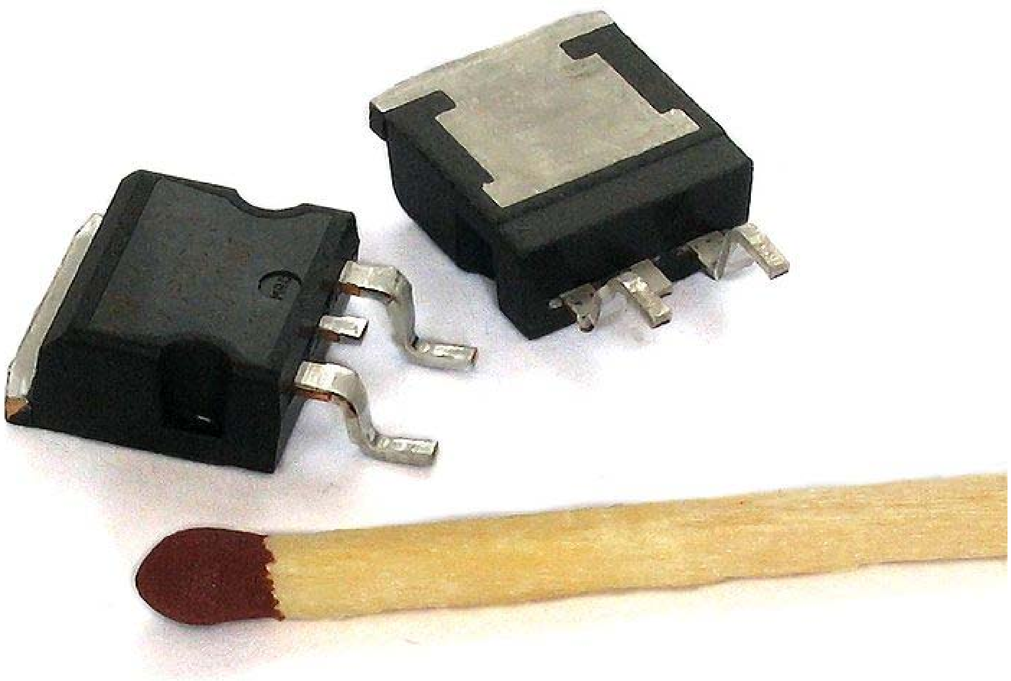
- SCR — Silicon Controlled Rectifier
- ASCR — Asymmetrical SCR
- RCT — Reverse Conducting Thyristor
- LASCR — Light Activated SCR, or LTT — Light triggered thyristor
- BOD — Breakover Diode — A gateless thyristor triggered by avalanche current
  - Shockley diode — Unidirectional trigger and switching device
  - Dynistor — Unidirectional switching device
  - DIAC — Bidirectional trigger device
  - SIDAC — Bidirectional switching device
  - Trisil, SIDACtor — Bidirectional protection devices
- TRIAC — Triode for Alternating Current — A bidirectional switching device containing two thyristor structures with common gate contact
- BCT — Bidirectional Control Thyristor — A bidirectional switching device containing two thyristor structures with separate gate contacts
- GTO — Gate Turn-Off thyristor
- IGCT — Integrated Gate Commutated Thyristor
  - MA-GTO — Modified Anode Gate Turn-Off thyristor
  - DB-GTO — Distributed Buffer Gate Turn-Off thyristor
- MCT — MOSFET Controlled Thyristor — It contains two additional FET structures for on/off control.
  - BRT — Base Resistance Controlled Thyristor
- SITh — Static Induction Thyristor, or FCTh — Field Controlled Thyristor containing a gate structure that can shut down anode current flow.
- LASS — Light Activated Semiconducting Switch
- AGT — Anode Gate Thyristor — A thyristor with gate on n-type layer near to the anode
- PUT or PUJT — Programmable Unijunction Transistor — A thyristor with gate on n-type layer near to the anode used as a functional replacement for unijunction transistor
- SCS — Sylicon Controlled Switch or Thyristor Tetrode — A thyristor with both cathode and anode gates

Reverse conducting thyristor

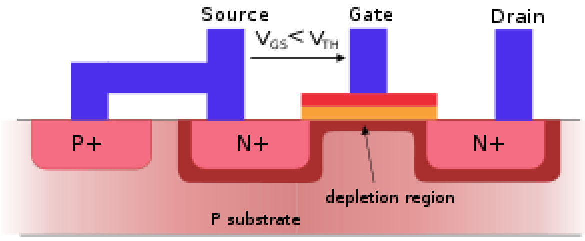
A reverse conducting thyristor (RCT) has an integrated reverse diode, so is not capable of reverse blocking. These devices are advantageous where a reverse or freewheel diode must be used. Because the SCR and diode never conduct at the same time they do not produce heat simultaneously and can easily be integrated and cooled together. Reverse conducting thyristors are often used in frequency changers and inverters.

Chapter- 9

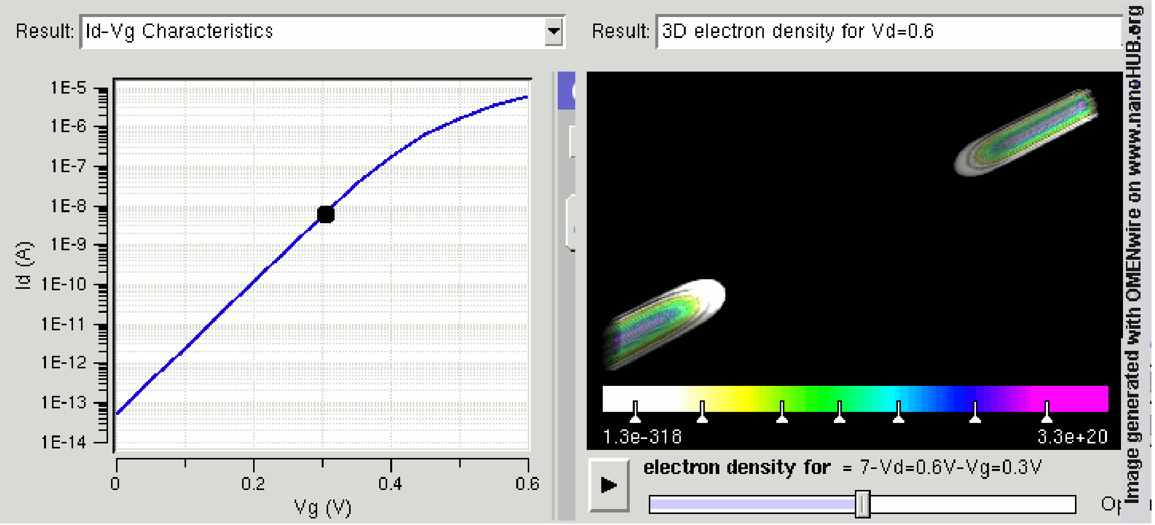
MOSFET



Two power MOSFETs in the surface-mount package D2PAK. Operating as switches, each of these components can sustain a blocking voltage of 120 volts in the *OFF* state, and can conduct a continuous current of 30 amperes in the *ON* state, dissipating up to about 100 watts and controlling a load of over 2000 watts. A matchstick is pictured for scale.



A cross section through an nMOSFET when the gate voltage V_{GS} is below the threshold for making a conductive channel; there is little or no conduction between the terminals source and drain; the switch is off. When the gate is more positive, it attracts electrons, inducing an *n*-type conductive channel in the substrate below the oxide, which allows electrons to flow between the *n*-doped terminals; the switch is on.


Result: Id-Vg Characteristics
Id (A)
1E-5
1E-6
1E-7
1E-8
1E-9
1E-10
1E-11
1E-12
1E-13
1E-14
0
0.2
0.4
0.6
Vg (V)
Result: 3D electron density for Vd=0.6
1.3e-318
3.3e+20
electron density for = 7-Vd=0.6V-Vg=0.3V
Image generated with OMENwire on www.nanoHUB.org

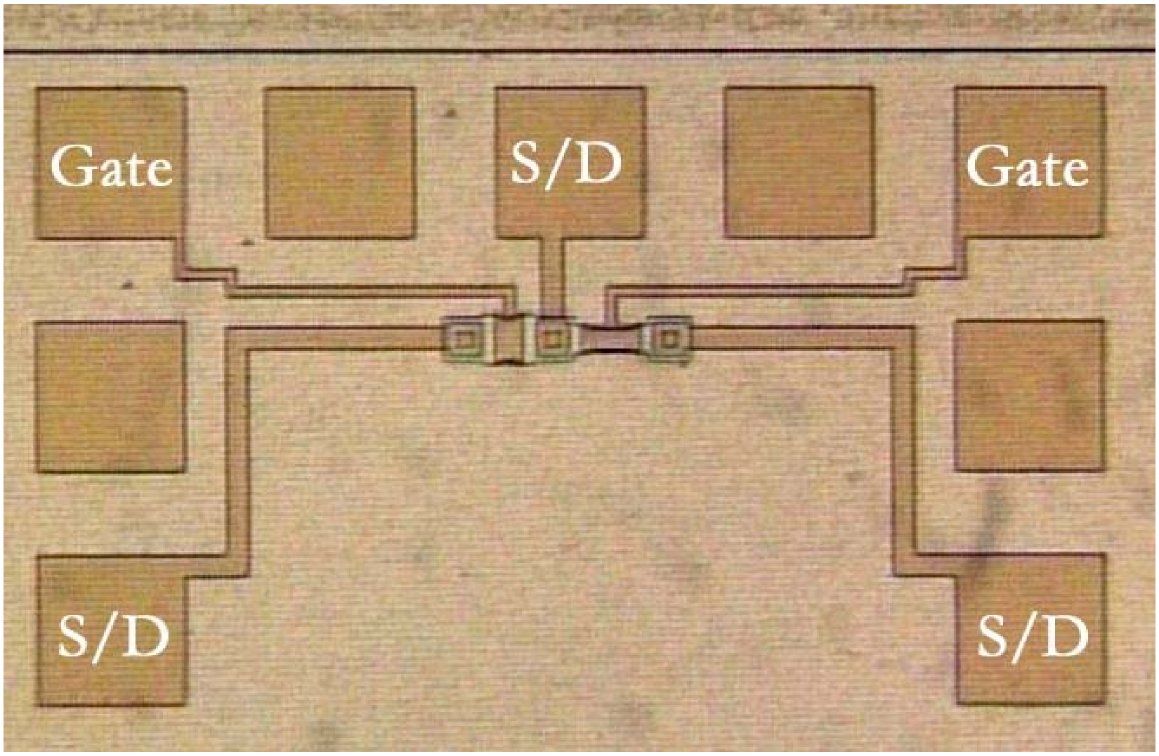
Result for formation of inversion channel (electron density) and attainment of threshold voltage (IV) in a nanowire MOSFET. Note that the threshold voltage for this device lies around 0.45V.

The **metal–oxide–semiconductor field-effect transistor** (**MOSFET**, **MOS-FET**, or **MOS FET**) is a device used for amplifying or switching electronic signals. The basic principle of this kind of transistor was first proposed by Julius Edgar Lilienfeld in 1925. In MOSFETs, a voltage on the oxide-insulated gate electrode can induce a conducting channel between the two other contacts called source and drain. The channel can be of n-type or p-type, and is accordingly called an nMOSFET or a pMOSFET (also commonly nMOS, pMOS). It is by far the most common transistor in both digital and analog circuits, though the bipolar junction transistor was at one time much more common.

The 'metal' in the name is now often a misnomer because the previously metal gate material is now often a layer of polysilicon (polycrystalline silicon). Aluminium had been the gate material until the mid 1970s, when polysilicon became dominant, due to its capability to form self-aligned gates. Metallic gates are regaining popularity, since it is difficult to increase the speed of operation of transistors without metal gates.

IGFET is a related term meaning insulated-gate field-effect transistor, and is almost synonymous with MOSFET, though it can refer to FETs with a gate insulator that is not oxide. Another synonym is MISFET for metal–insulator–semiconductor FET.

Composition



Photomicrograph of two metal-gate MOSFETs in a test pattern. Probe pads for two gates and three source/drain nodes are labeled.

Usually the semiconductor of choice is silicon, but some chip manufacturers, most notably IBM and Intel, recently started using a chemical compound of silicon and germanium (SiGe) in MOSFET channels. Unfortunately, many semiconductors with better electrical properties than silicon, such as gallium arsenide, do not form good semiconductor-to-insulator interfaces, thus are not suitable for MOSFETs. Research continues on creating insulators with acceptable electrical characteristics on other semiconductor material.

In order to overcome power consumption increase due to gate current leakage, high-κ dielectric replaces silicon dioxide for the gate insulator, while metal gates return by replacing polysilicon.

The gate is separated from the channel by a thin insulating layer, traditionally of silicon dioxide and later of silicon oxynitride. Some companies have started to introduce a high-κ dielectric + metal gate combination in the 45 nanometer node.

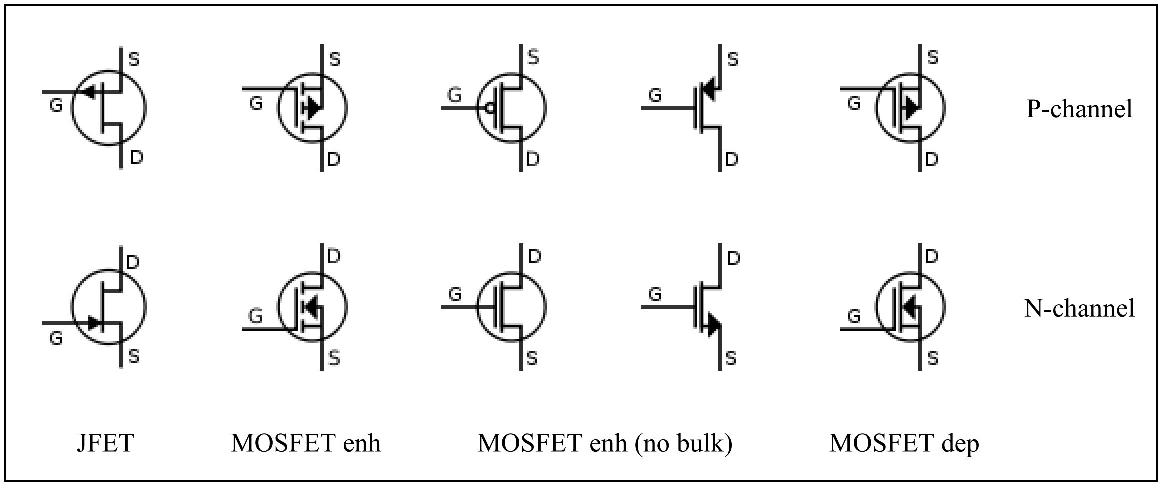
When a voltage is applied between the gate and body terminals, the electric field generated penetrates through the oxide and creates an "inversion layer" or "channel" at the semiconductor-insulator interface. The inversion channel is of the same type, P-type or N-type, as the source and drain, thus it provides a channel through which current can pass. Varying the voltage between the gate and body modulates the conductivity of this layer and allows to control the current flow between drain and source.

Circuit symbols

A variety of symbols are used for the MOSFET. The basic design is generally a line for the channel with the source and drain leaving it at right angles and then bending back at right angles into the same direction as the channel. Sometimes three line segments are used for enhancement mode and a solid line for depletion mode. Another line is drawn parallel to the channel for the gate.

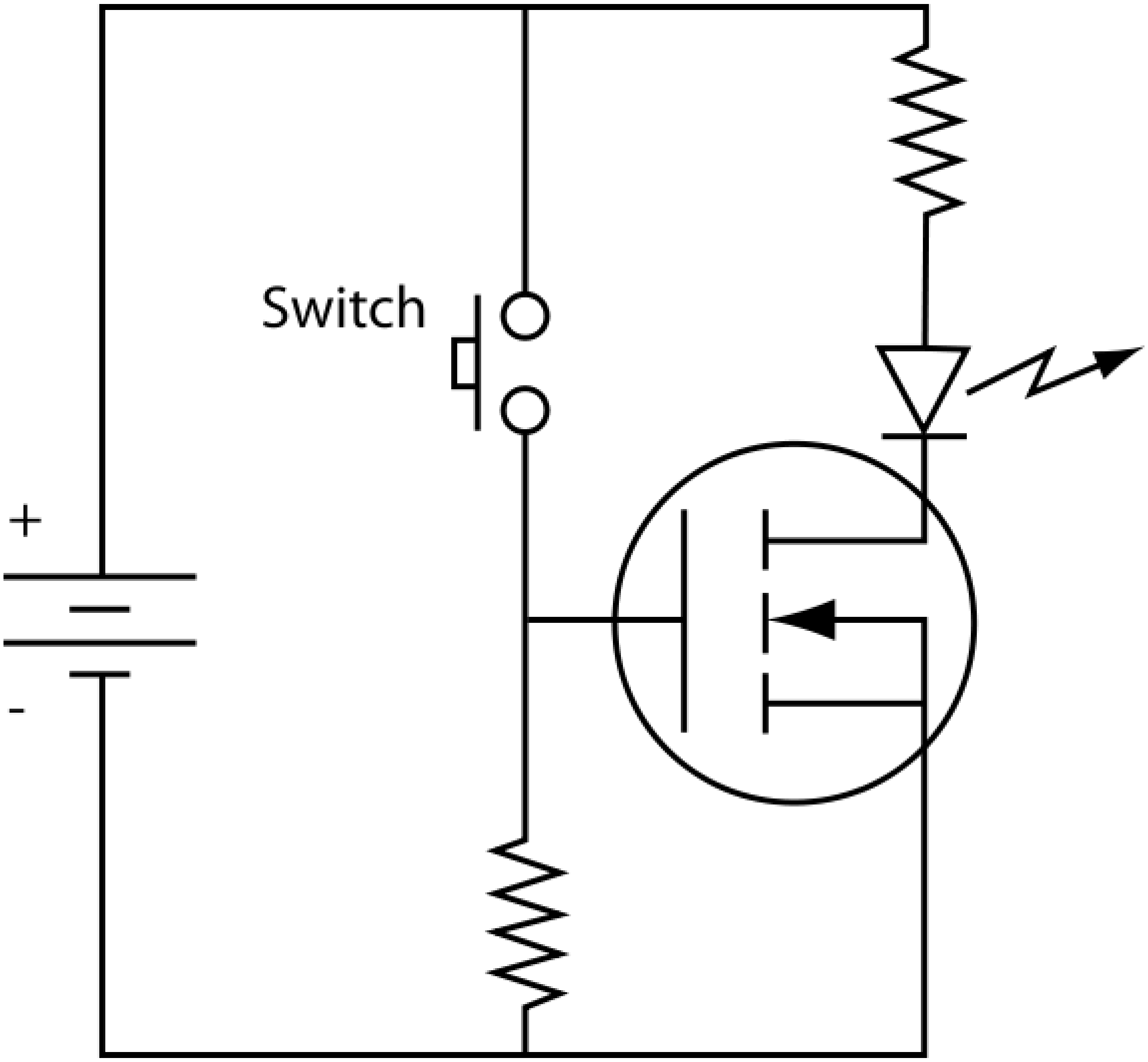
The bulk connection, if shown, is shown connected to the back of the channel with an arrow indicating PMOS or NMOS. Arrows always point from P to N, so an NMOS (N-channel in P-well or P-substrate) has the arrow pointing in (from the bulk to the channel). If the bulk is connected to the source (as is generally the case with discrete devices) it is sometimes angled to meet up with the source leaving the transistor. If the bulk is not shown (as is often the case in IC design as they are generally common bulk) an inversion symbol is sometimes used to indicate PMOS, alternatively an arrow on the source may be used in the same way as for bipolar transistors (out for nMOS, in for pMOS).

Comparison of enhancement-mode and depletion-mode MOSFET symbols, along with JFET symbols (drawn with source and drain ordered such that higher voltages appear higher on the page than lower voltages):

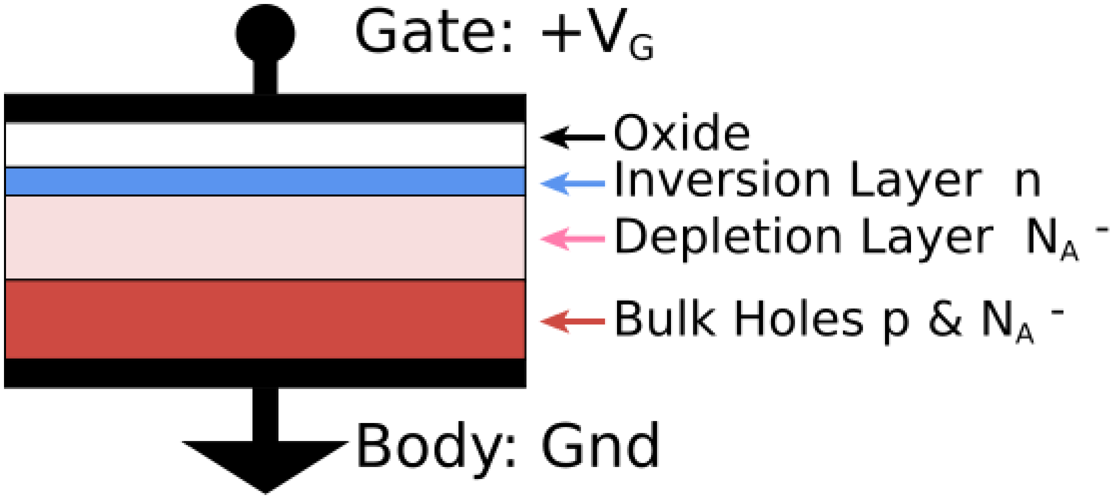


For the symbols in which the bulk, or body, terminal is shown, it is here shown internally connected to the source. This is a typical configuration, but by no means the only important configuration. In general, the MOSFET is a four-terminal device, and in integrated circuits many of the MOSFETs share a body connection, not necessarily connected to the source terminals of all the transistors.

MOSFET operation



Example application of an N-Channel MOSFET. When the switch is pushed the LED lights up.



Metal–oxide–semiconductor structure on P-type silicon

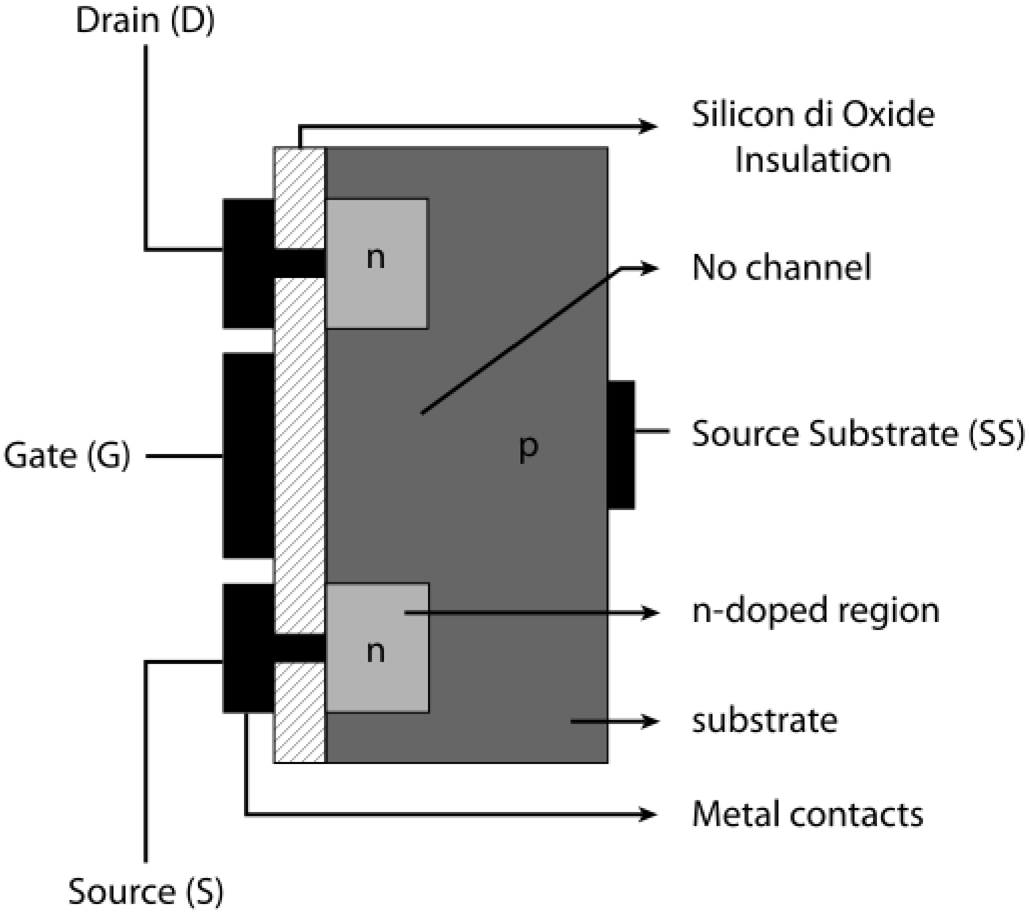
Metal–oxide–semiconductor structure

A traditional metal–oxide–semiconductor (MOS) structure is obtained by growing a layer of silicon dioxide (SiO_2) on top of a silicon substrate and depositing a layer of metal or polycrystalline silicon (the latter is commonly used). As the silicon dioxide is a dielectric material, its structure is equivalent to a planar capacitor, with one of the electrodes replaced by a semiconductor.

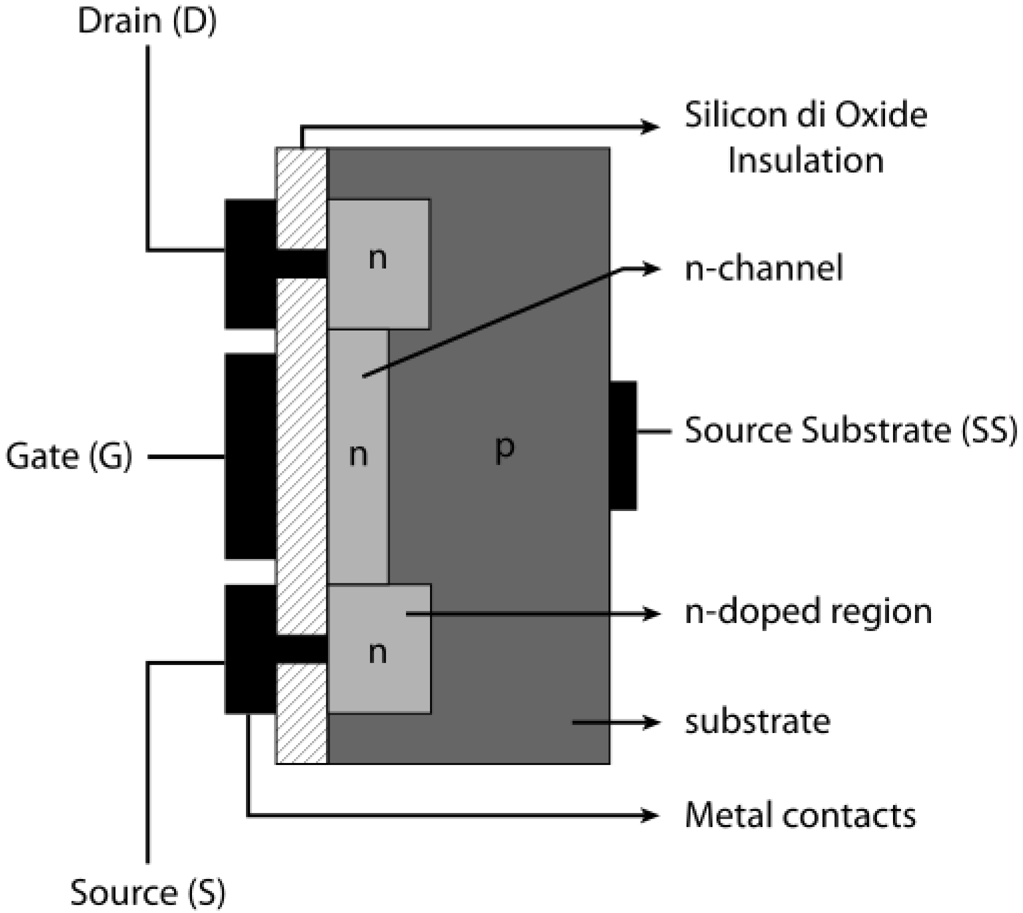
When a voltage is applied across a MOS structure, it modifies the distribution of charges in the semiconductor. If we consider a P-type semiconductor (with N_A the density of acceptors, p the density of holes; $p = N_A$ in neutral bulk), a positive voltage, V_{GB}, from gate to body creates a depletion layer by forcing the positively charged holes away from the gate-insulator/semiconductor interface, leaving exposed a carrier-free region of immobile, negatively charged acceptor ions. If V_{GB} is high enough, a high concentration of negative charge carriers forms in an **inversion layer** located in a thin layer next to the interface between the semiconductor and the insulator. Unlike the MOSFET, where the inversion layer electrons are supplied rapidly from the source/drain electrodes, in the MOS capacitor they are produced much more slowly by thermal generation through carrier generation and recombination centers in the depletion region. Conventionally, the gate voltage at which the volume density of electrons in the inversion layer is the same as the volume density of holes in the body is called the threshold voltage.

This structure with p-type body is the basis of the N-type MOSFET, which requires the addition of an N-type source and drain regions.

MOSFET structure and channel formation



Cross section of an NMOS without channel formed: OFF state



Cross section of an NMOS with channel formed: ON state

A metal–oxide–semiconductor field-effect transistor (MOSFET) is based on the modulation of charge concentration by a MOS capacitance between a **body** electrode and a **gate** electrode located above the body and insulated from all other device regions by a gate dielectric layer which in the case of a MOSFET is an oxide, such as silicon dioxide. If dielectrics other than an oxide such as silicon dioxide (often referred to as oxide) are employed the device may be referred to as a metal–insulator–semiconductor FET (MISFET). Compared to the MOS capacitor, the MOSFET includes two additional terminals (**source** and **drain**), each connected to individual highly doped regions that are separated by the body region. These regions can be either p or n type, but they must both be of the same type, and of opposite type to the body region. The source and drain (unlike the body) are highly doped as signified by a '+' sign after the type of doping.

If the MOSFET is an n-channel or nMOS FET, then the source and drain are 'n+' regions and the body is a 'p' region. As described above, with sufficient gate voltage, holes from the body are driven away from the gate, forming an inversion layer or *n-channel* at the interface between the p region and the oxide. This conducting channel extends between the source and the drain, and current is conducted through it when a voltage is applied between source and drain.

For gate voltages below the threshold value, the channel is lightly populated, and only a very small subthreshold leakage current can flow between the source and the drain.

If the MOSFET is a p-channel or pMOS FET, then the source and drain are 'p+' regions and the body is a 'n' region. When a negative gate-source voltage (positive source-gate) is applied, it creates a *p-channel* at the surface of the n region, analogous to the n-channel case, but with opposite polarities of charges and voltages. When a voltage less negative than the threshold value (a negative voltage for p-channel) is applied between gate and source, the channel disappears and only a very small subthreshold current can flow between the source and the drain.

The source is so named because it is the source of the charge carriers (electrons for n-channel, holes for p-channel) that flow through the channel; similarly, the drain is where the charge carriers leave the channel.

The device may comprise a Silicon On Insulator (SOI) device in which a Buried OXide (BOX) is formed below a thin semiconductor layer. If the channel region between the gate dielectric and a Buried Oxide (BOX) region is very thin, the very thin channel region is referred to as an Ultra Thin Channel (UTC) region with the source and drain regions formed on either side thereof in and/or above the thin semiconductor layer. Alternatively, the device may comprise a SEMiconductor On Insulator (SEMOI) device in which semiconductors other than silicon are employed. Many alternative semiconductor materials may be employed.

When the source and drain regions are formed above the channel in whole or in part, they are referred to as Raised Source/Drain (RSD) regions.

Modes of operation

The operation of a MOSFET can be separated into three different modes, depending on the voltages at the terminals. In the following discussion, a simplified algebraic model is used that is accurate only for old technology. Modern MOSFET characteristics require computer models that have rather more complex behavior.

For an **enhancement-mode, n-channel MOSFET**, the three operational modes are:

Cutoff, subthreshold, or weak-inversion mode

> **When $V_{GS} < V_{\mathrm{th}}$:**
> where V_{th} is the threshold voltage of the device.
> According to the basic threshold model, the transistor is turned off, and there is no conduction between drain and source. In reality, the Boltzmann distribution of electron energies allows some of the more energetic electrons at the source to enter the channel and flow to the drain, resulting in a subthreshold current that is an exponential function of gate–source voltage. While the current between drain and source should ideally be zero when the transistor is being used as a turned-off switch, there is a weak-inversion current, sometimes called subthreshold leakage.

In weak inversion the current varies exponentially with gate-to-source bias V_{GS} as given approximately by:

$$I_D \approx I_{D0} e^{\frac{V_{GS}-V_{th}}{nV_T}},$$

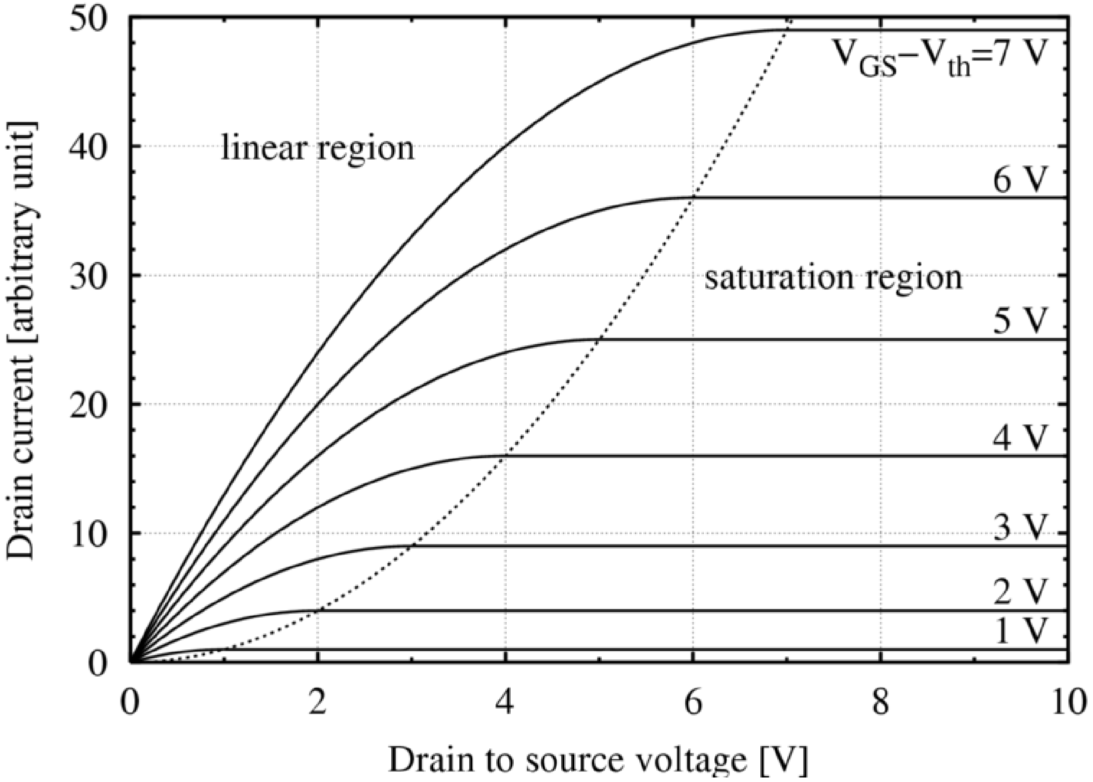
where I_{D0} = current at $V_{GS} = V_{th}$ and the slope factor n is given by

$n = 1 + C_D / C_{OX}$,

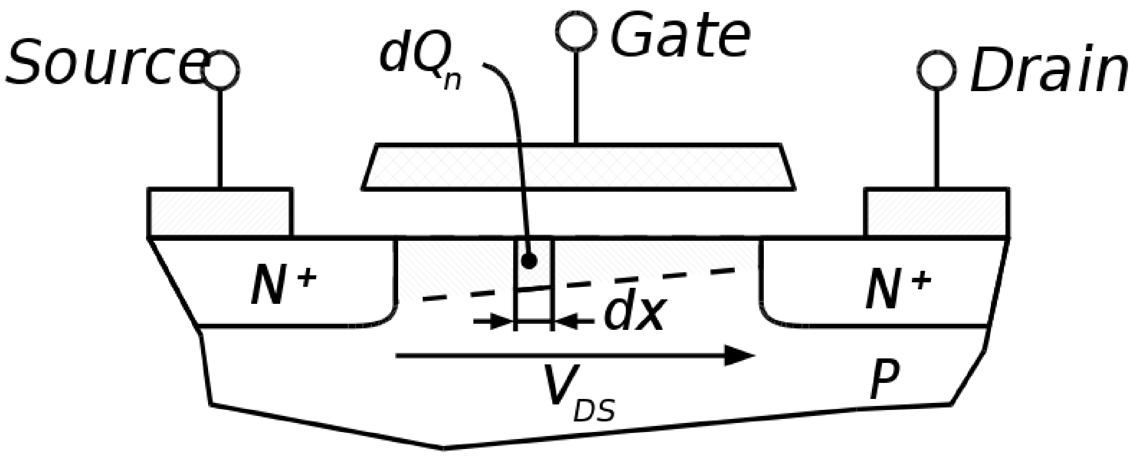
with C_D = capacitance of the depletion layer and C_{OX} = capacitance of the oxide layer. In a long-channel device, there is no drain voltage dependence of the current once $V_{DS} >> V_T$, but as channel length is reduced drain-induced barrier lowering introduces drain voltage dependence that depends in a complex way upon the device geometry (for example, the channel doping, the junction doping and so on). Frequently, threshold voltage V_{th} for this mode is defined as the gate voltage at which a selected value of current I_{D0} occurs, for example, $I_{D0} = 1\ \mu A$, which may not be the same V_{th}-value used in the equations for the following modes.

Some micropower analog circuits are designed to take advantage of subthreshold conduction. By working in the weak-inversion region, the MOSFETs in these circuits deliver the highest possible transconductance-to-current ratio, namely: $g_m / I_D = 1 / (nV_T)$, almost that of a bipolar transistor.

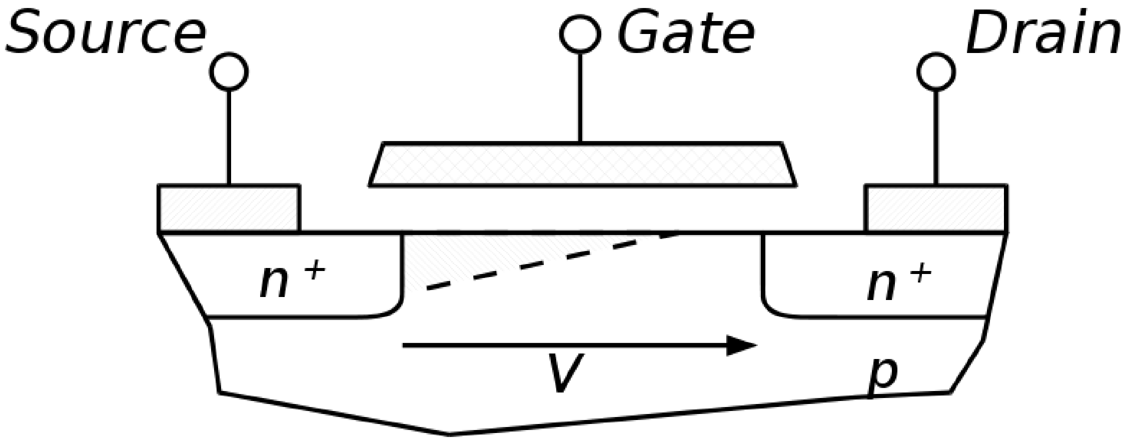
The subthreshold *I–V curve* depends exponentially upon threshold voltage, introducing a strong dependence on any manufacturing variation that affects threshold voltage; for example: variations in oxide thickness, junction depth, or body doping that change the degree of drain-induced barrier lowering. The resulting sensitivity to fabricational variations complicates optimization for leakage and performance.



MOSFET drain current vs. drain-to-source voltage for several values of $V_{GS} - V_{th}$; the boundary between **linear** (**Ohmic**) and **saturation** (**active**) modes is indicated by the upward curving parabola.



Cross section of a MOSFET operating in the linear (Ohmic) region; strong inversion region present even near drain



Cross section of a MOSFET operating in the saturation (active) region; channel exhibits **pinch-off** near drain

Triode mode or linear region (also known as the ohmic mode)

When $V_{GS} > V_{\text{th}}$ and $V_{DS} < (V_{GS} - V_{\text{th}})$

The transistor is turned on, and a channel has been created which allows current to flow between the drain and the source. The MOSFET operates like a resistor, controlled by the gate voltage relative to both the source and drain voltages. The current from drain to source is modeled as:

$$I_D = \mu_n C_{ox} \frac{W}{L} \left((V_{GS} - V_{th}) V_{DS} - \frac{V_{DS}^2}{2} \right)$$

where μ_n is the charge-carrier effective mobility, W is the gate width, L is the gate length and C_{ox} is the gate oxide capacitance per unit area. The transition from the exponential subthreshold region to the triode region is not as sharp as the equations suggest.

Saturation or active mode

When $V_{GS} > V_{\text{th}}$ and $V_{DS} > (V_{GS} – V_{\text{th}})$

The switch is turned on, and a channel has been created, which allows current to flow between the drain and source. Since the drain voltage is higher than the gate voltage, the electrons spread out, and conduction is not through a narrow channel but through a broader, two- or three-dimensional current distribution extending away from the interface and deeper in the substrate. The onset of this region is also known as **pinch-off** to indicate the lack of channel region near the drain. The drain current is now weakly dependent upon drain voltage and controlled primarily by the gate–source voltage, and modeled very approximately as:

$$I_D = \frac{\mu_n C_{ox}}{2} \frac{W}{L} (V_{GS} - V_{th})^2 (1 + \lambda V_{DS}).$$

The additional factor involving λ, the channel-length modulation parameter, models current dependence on drain voltage due to the Early effect, or channel length modulation. According to this equation, a key design parameter, the MOSFET transconductance is:

$$g_m = \frac{2I_D}{V_{GS} - V_{th}} = \frac{2I_D}{V_{ov}},$$

where the combination $V_{ov} = V_{GS} – V_{th}$ is called the **overdrive voltage**. Another key design parameter is the MOSFET output resistance r_{out} given by:

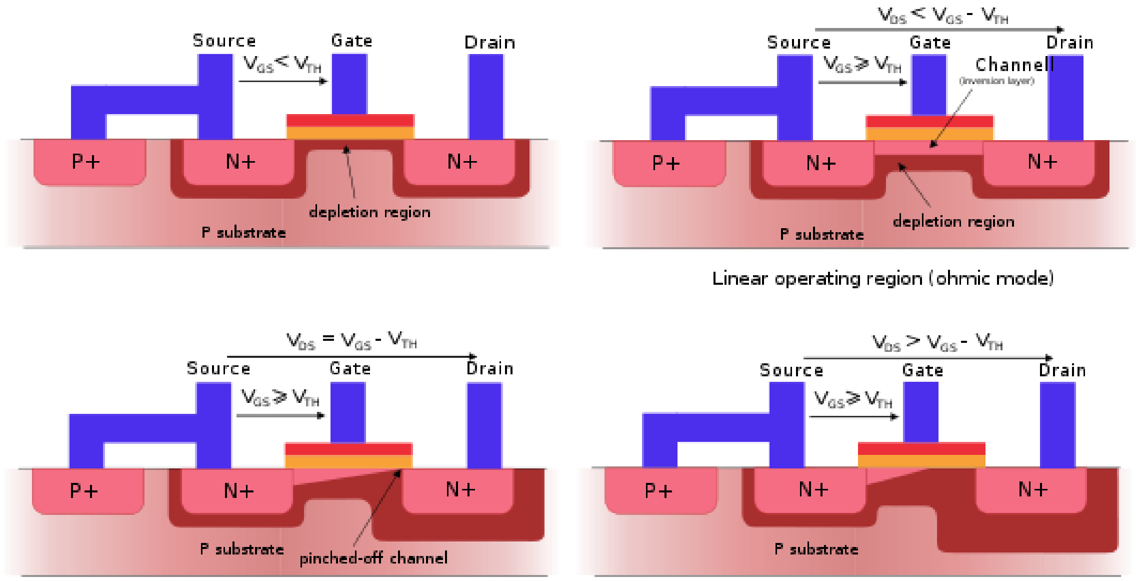
$$r_{out} = \frac{1}{\lambda I_D}.$$

r_{out} is the inverse of g_{DS} where $g_{DS} = \frac{\partial I_{DS}}{\partial V_{DS}}$. V_{DS} is the expression in saturation region.

If λ is taken as zero, an infinite output resistance of the device results that leads to unrealistic circuit predictions, particularly in analog circuits.

As the channel length becomes very short, these equations become quite inaccurate. New physical effects arise. For example, carrier transport in the active mode may become limited by velocity saturation. When velocity saturation dominates, the saturation drain current is more nearly linear than quadratic in V_{GS}. At even shorter lengths, carriers transport with near zero scattering, known as quasi-ballistic transport. In addition, the output current is affected by drain-induced barrier lowering of the threshold voltage.

Body effect



Linear operating region (ohmic mode)

Saturation mode at point of pinch-off

Saturation mode

Ohmic contact to body to ensure no body bias; top left:subthreshold, top right:Ohmic mode, bottom left:Active mode at onset of pinch-off, bottom right: Active mode well into pinch-off – channel length modulation evident

The body effect describes the changes in the threshold voltage by the change in the source-bulk voltage, approximated by the following equation:

$$V_{TN} = V_{TO} + \gamma \left(\sqrt{V_{SB} + 2\phi} - \sqrt{2\phi} \right),$$

where V_{TN} is the threshold voltage with substrate bias present, and V_{TO} is the zero-V_{SB} value of threshold voltage, γ is the body effect parameter, and 2φ is the surface potential parameter.

The body can be operated as a second gate, and is sometimes referred to as the "back gate"; the body effect is sometimes called the "back-gate effect".

The primacy of MOSFETs

In 1959, Dawon Kahng and Martin M. (John) Atalla at Bell Labs invented the metal–oxide–semiconductor field-effect transistor (MOSFET). Operationally and structurally different from the bipolar junction transistor, the MOSFET was made by putting an insulating layer on the surface of the semiconductor and then placing a metallic gate electrode on that. It used crystalline silicon for the semiconductor and a thermally oxidized layer of silicon dioxide for the insulator. The silicon MOSFET did not generate localized electron traps at the interface between the silicon and its native oxide layer, and thus was inherently free from the trapping and scattering of carriers that had impeded the performance of earlier field-effect transistors. Following the (expensive) development of clean rooms to reduce contamination to levels never before thought necessary, and of photolithography and the planar process to allow circuits to be made in very few steps, the Si–SiO_2 system possessed such technical attractions as low cost of production (on a per circuit basis) and ease of integration. Largely because of these two factors, the MOSFET has become the most widely used type of transistor in integrated circuits.

CMOS circuits

The MOSFET is used in digital CMOS logic, which uses p- and n-channel MOSFETs as building blocks. Overheating is a major concern in integrated circuits since ever more transistors are packed into ever smaller chips. CMOS logic reduces power consumption because no current flows (ideally), and thus no power is consumed, except when the inputs to logic gates are being switched. CMOS accomplishes this current reduction by complementing every nMOSFET with a pMOSFET and connecting both gates and both drains together. A high voltage on the gates will cause the nMOSFET to conduct and the pMOSFET not to conduct and a low voltage on the gates causes the reverse. During the switching time as the voltage goes from one state to another, both MOSFETs will conduct briefly. This arrangement greatly reduces power consumption and heat generation. Digital and analog CMOS applications are described below.

Digital

The growth of digital technologies like the microprocessor has provided the motivation to advance MOSFET technology faster than any other type of silicon-based transistor. A big advantage of MOSFETs for digital switching is that the oxide layer between the gate and the channel prevents DC current from flowing through the gate, further reducing power consumption and giving a very large input impedance. The insulating oxide between the gate and channel effectively isolates a MOSFET in one logic stage from earlier and later stages, which allows a single MOSFET output to drive a considerable number of MOSFET inputs. Bipolar transistor-based logic (such as TTL) does not have such a high fanout capacity. This isolation also makes it easier for the designers to ignore to some extent loading effects between logic stages independently. That extent is defined by the operating frequency: as frequencies increase, the input impedance of the MOSFETs decreases.

Analog

The MOSFET's advantages in most digital circuits do not translate into supremacy in all analog circuits. The two types of circuit draw upon different features of transistor behavior. Digital circuits switch, spending most of their time outside the switching region, while analog circuits depend on MOSFET behavior held precisely in the switching region of operation. The bipolar junction transistor (BJT) has traditionally been the analog designer's transistor of choice, due largely to its higher transconductance and its higher output impedance (drain-voltage independence) in the switching region.

Nevertheless, MOSFETs are widely used in many types of analog circuits because of certain advantages. The characteristics and performance of many analog circuits can be designed by changing the sizes (length and width) of the MOSFETs used. By comparison, in most bipolar transistors the size of the device does not significantly affect the performance. MOSFETs' ideal characteristics regarding gate current (zero) and drain-source offset voltage (zero) also make them nearly ideal switch elements, and also make switched capacitor analog circuits practical. In their linear region, MOSFETs can be used as precision resistors, which can have a much higher controlled resistance than BJTs. In high power circuits, MOSFETs sometimes have the advantage of not suffering from thermal runaway as BJTs do. Also, they can be formed into capacitors and gyrator circuits which allow op-amps made from them to appear as inductors, thereby allowing all of the normal analog devices, except for diodes (which can be made smaller than a MOSFET anyway), to be built entirely out of MOSFETs. This allows for complete analog circuits to be made on a silicon chip in a much smaller space.

Some ICs combine analog and digital MOSFET circuitry on a single mixed-signal integrated circuit, making the needed board space even smaller. This creates a need to isolate the analog circuits from the digital circuits on a chip level, leading to the use of isolation rings and Silicon-On-Insulator (SOI). The main advantage of BJTs versus MOSFETs in the analog design process is the ability of BJTs to handle a larger current in a smaller space. Fabrication processes exist that incorporate BJTs and MOSFETs into a single device. Mixed-transistor devices are called Bi-FETs (Bipolar-FETs) if they contain just one BJT-FET and BiCMOS (bipolar-CMOS) if they contain complementary BJT-FETs. Such devices have the advantages of both insulated gates and higher current density.

Advantages of BJT over MOSFET

BJTs have some advantages over MOSFETs for at least two digital applications. Firstly, in high speed switching, they do not have the "larger" capacitance from the gate, which when multiplied by the resistance of the channel gives the intrinsic time constant of the process. The intrinsic time constant places a limit on the speed a MOSFET can operate at because higher frequency signals are filtered out. Widening the channel reduces the resistance of the channel, but increases the capacitance by exactly the same amount. Reducing the width of the channel increases the resistance, but reduces the capacitance by the same amount. R*C=Tc1, 0.5R*2C=Tc1, 2R*0.5C=Tc1. There is no way to minimize the intrinsic time constant for a certain process. Different processes using different channel lengths, channel heights, gate thicknesses and materials will have different intrinsic time constants. This problem is mostly avoided with a BJT because it does not have a gate.

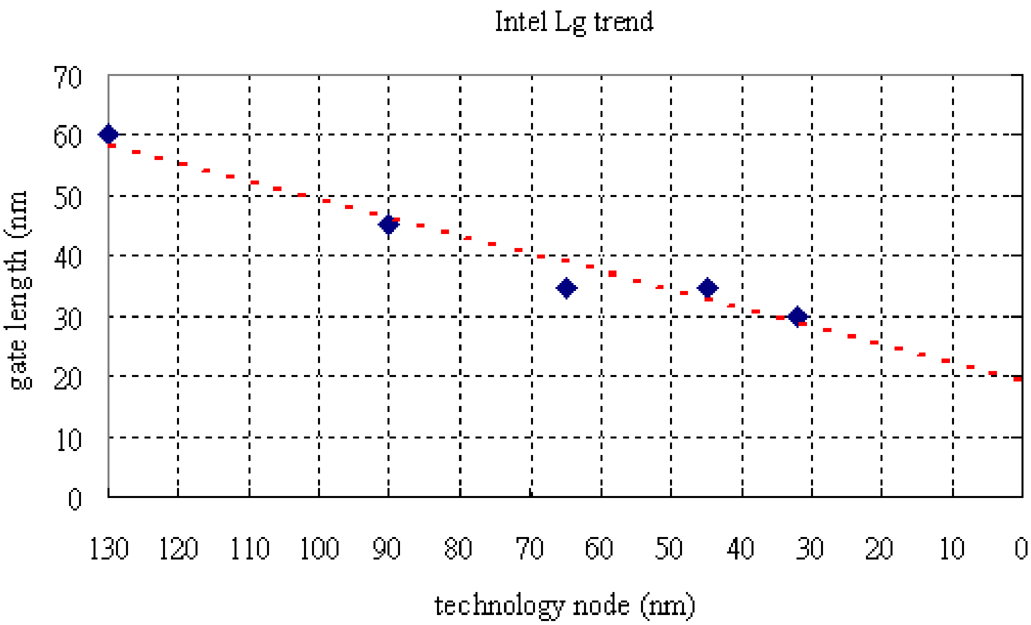
The second application where BJTs have an advantage over MOSFETs stems from the first. When driving many other gates, called fanout, the resistance of the MOSFET is in series with the gate capacitances of the other FETs, creating a secondary time constant. Delay circuits use this fact to create a fixed signal delay by using a small CMOS device to send a signal to many other, many times larger CMOS devices. The secondary time constant can be minimized by increasing the driving FET's channel width to decrease its resistance and decreasing the channel widths of the FETs being driven, decreasing their capacitance. The drawback is that it increases the capacitance of the driving FET and increases the resistance of the FETs being driven, but usually these drawbacks are a minimal problem when compared to the timing problem. BJTs are better able to drive the other gates because they can output more current than MOSFETs, allowing for the FETs being driven to charge faster. Many chips use MOSFET inputs and BiCMOS outputs (see above).

MOSFET scaling

Over the past decades, the MOSFET has continually been scaled down in size; typical MOSFET channel lengths were once several micrometres, but modern integrated circuits are incorporating MOSFETs with channel lengths of tens of nanometers. Intel began production of a process featuring a 32 nm feature size (with the channel being even shorter) in late 2009. The semiconductor industry maintains a "roadmap", the ITRS, which sets the pace for MOSFET development. Historically, the difficulties with decreasing the size of the MOSFET have been associated with the semiconductor device fabrication process, the need to use very low voltages, and with poorer electrical performance necessitating circuit redesign and innovation (small MOSFETs exhibit higher leakage currents, and lower output resistance, discussed below).

Reasons for MOSFET scaling

Smaller MOSFETs are desirable for several reasons. The main reason to make transistors smaller is to pack more and more devices in a given chip area. This results in a chip with the same functionality in a smaller area, or chips with more functionality in the same area. Since fabrication costs for a semiconductor wafer are relatively fixed, the cost per integrated circuits is mainly related to the number of chips that can be produced per wafer. Hence, smaller ICs allow more chips per wafer, reducing the price per chip. In fact, over the past 30 years the number of transistors per chip has been doubled every 2–3 years once a new technology node is introduced. For example the number of MOSFETs in a microprocessor fabricated in a 45 nm technology is twice as many as in a 65 nm chip. This doubling of the transistor count was first observed by Gordon Moore in 1965 and is commonly referred to as Moore's law.



Trend of Intel CPU transistor gate length

It is also expected that smaller transistors switch faster. For example, one approach to size reduction is a scaling of the MOSFET that requires all device dimensions to reduce proportionally. The main device dimensions are the transistor length, width, and the oxide thickness, each (used to) scale with a factor of 0.7 per node. This way, the transistor channel resistance does not change with scaling, while gate capacitance is cut by a factor of 0.7. Hence, the RC delay of the transistor scales with a factor of 0.7.

While this has been traditionally the case for the older technologies, for the state-of-the-art MOSFETs reduction of the transistor dimensions does not necessarily translate to higher chip speed because the delay due to interconnections is more significant.

Difficulties arising due to MOSFET size reduction

Producing MOSFETs with channel lengths much smaller than a micrometer is a challenge, and the difficulties of semiconductor device fabrication are always a limiting factor in advancing integrated circuit technology. In recent years, the small size of the MOSFET, below a few tens of nanometers, has created operational problems.

Higher subthreshold conduction

As MOSFET geometries shrink, the voltage that can be applied to the gate must be reduced to maintain reliability. To maintain performance, the threshold voltage of the MOSFET has to be reduced as well. As threshold voltage is reduced, the transistor cannot be switched from complete turn-off to complete turn-on with the limited voltage swing available; the circuit design is a compromise between strong current in the "on" case and low current in the "off" case, and the application determines whether to favor one over the other. Subthreshold leakage (including subthreshold conduction, gate-oxide leakage and reverse-biased junction leakage), which was ignored in the past, now can consume upwards of half of the total power consumption of modern high-performance VLSI chips.

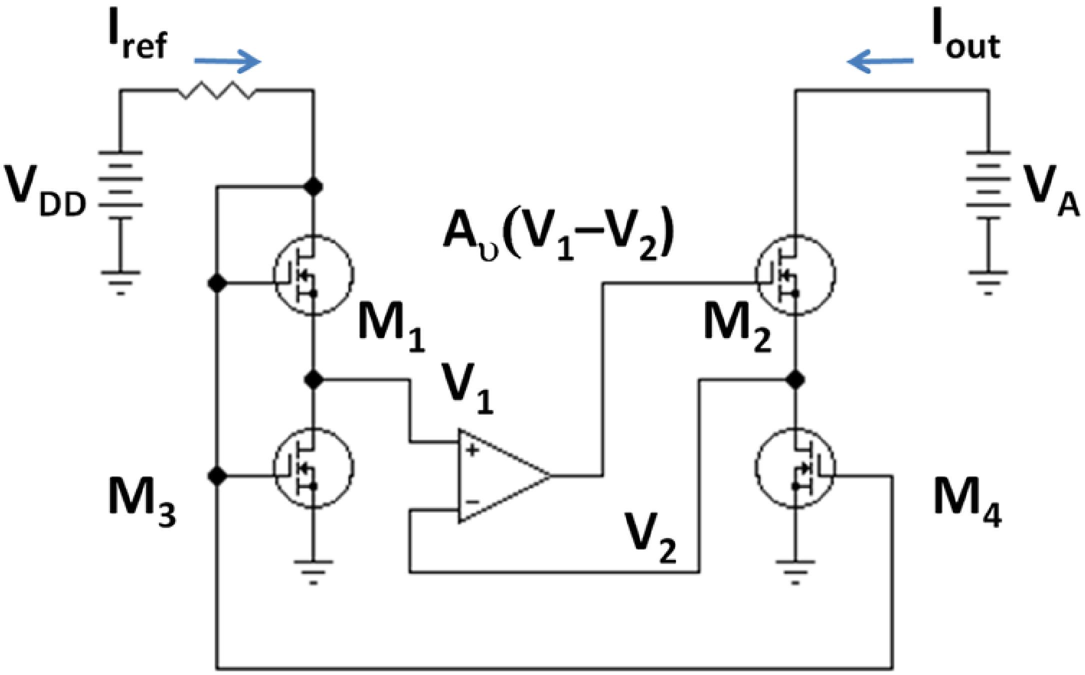
Increased gate-oxide leakage

The gate oxide, which serves as insulator between the gate and channel, should be made as thin as possible to increase the channel conductivity and performance when the transistor is on and to reduce subthreshold leakage when the transistor is off. However, with current gate oxides with a thickness of around 1.2 nm (which in silicon is ~5 atoms thick) the quantum mechanical phenomenon of electron tunneling occurs between the gate and channel, leading to increased power consumption.

Insulators that have a larger dielectric constant than silicon dioxide (referred to as high-k dielectrics), such as group IVb metal silicates e.g. hafnium and zirconium silicates and oxides are being used to reduce the gate leakage from the 45 nanometer technology node onwards. Increasing the dielectric constant of the gate dielectric allows a thicker layer while maintaining a high capacitance (capacitance is proportional to dielectric constant and inversely proportional to dielectric thickness). All else equal, a higher dielectric thickness reduces the quantum tunneling current through the dielectric between the gate and the channel. On the other hand, the barrier height of the new gate insulator is an important consideration; the difference in conduction band energy between the semiconductor and the dielectric (and the corresponding difference in valence band energy) also affects leakage current level. For the traditional gate oxide, silicon dioxide, the former barrier is approximately 8 eV. For many alternative dielectrics the value is significantly lower, tending to increase the tunneling current, somewhat negating the advantage of higher dielectric constant.

Increased junction leakage

To make devices smaller, junction design has become more complex, leading to higher doping levels, shallower junctions, "halo" doping and so forth, all to decrease drain-induced barrier lowering. To keep these complex junctions in place, the annealing steps formerly used to remove damage and electrically active defects must be curtailed increasing junction leakage. Heavier doping is also associated with thinner depletion layers and more recombination centers that result in increased leakage current, even without lattice damage.



MOSFET version of gain-boosted current mirror; M_1 and M_2 are in active mode, while M_3 and M_4 are in Ohmic mode, and act like resistors. The operational amplifier provides feedback that maintains a high output resistance

Lower output resistance

For analog operation, good gain requires a high MOSFET output impedance, which is to say, the MOSFET current should vary only slightly with the applied drain-to-source voltage. As devices are made smaller, the influence of the drain competes more successfully with that of the gate due to the growing proximity of these two electrodes, increasing the sensitivity of the MOSFET current to the drain voltage. To counteract the resulting decrease in output resistance, circuits are made more complex, either by requiring more devices, for example the cascode and cascade amplifiers, or by feedback circuitry using operational amplifiers, for example a circuit like that in the adjacent figure.

Lower transconductance

The transconductance of the MOSFET decides its gain and is proportional to hole or electron mobility (depending on device type), at least for low drain voltages. As MOSFET size is reduced, the fields in the channel increase and the dopant impurity levels increase. Both changes reduce the carrier mobility, and hence the transconductance. As channel lengths are reduced without proportional reduction in drain voltage, raising the electric field in the channel, the result is velocity saturation of the carriers, limiting the current and the transconductance.

Interconnect capacitance

Traditionally, switching time was roughly proportional to the gate capacitance of gates. However, with transistors becoming smaller and more transistors being placed on the chip, interconnect capacitance (the capacitance of the metal-layer connections between different parts of the chip) is becoming a large percentage of capacitance. Signals have to travel through the interconnect, which leads to increased delay and lower performance.

Heat production

Large heatsinks to cool power transistors in a TRM-800 audio amplifier

The ever-increasing density of MOSFETs on an integrated circuit creates problems of substantial localized heat generation that can impair circuit operation. Circuits operate slower at high temperatures, and have reduced reliability and shorter lifetimes. Heat sinks and other cooling methods are now required for many integrated circuits including microprocessors.

Power MOSFETs are at risk of thermal runaway. As their on-state resistance rises with temperature, if the load is approximately a constant-current load then the power loss rises correspondingly, generating further heat. When the heatsink is not able to keep the temperature low enough, the junction temperature may rise quickly and uncontrollably, resulting in destruction of the device.

Process variations

With MOSFETS becoming smaller, the number of atoms in the silicon that produce many of the transistor's properties is becoming fewer, with the result that control of dopant numbers and placement is more erratic. During chip manufacturing, random process variations affect all transistor dimensions: length, width, junction depths, oxide thickness *etc.*, and become a greater percentage of overall transistor size as the transistor shrinks. The transistor characteristics become less certain, more statistical. The random nature of manufacture means we do not know which particular example MOSFETs actually will end up in a particular instance of the circuit. This uncertainty forces a less optimal design because the design must work for a great variety of possible component MOSFETs.

Modeling challenges

Modern ICs are computer-simulated with the goal of obtaining working circuits from the very first manufactured lot. As devices are miniaturized, the complexity of the processing makes it difficult to predict exactly what the final devices look like, and modeling of physical processes becomes more challenging as well. In addition, microscopic variations in structure due simply to the probabilistic nature of atomic processes require statistical (not just deterministic) predictions. These factors combine to make adequate simulation and "right the first time" manufacture difficult.

MOSFET construction

Gate material

The primary criterion for the gate material is that it is a good conductor. Highly-doped polycrystalline silicon is an acceptable but certainly not ideal conductor, and also suffers from some more technical deficiencies in its role as the standard gate material. Nevertheless, there are several reasons favoring use of polysilicon:

1. The threshold voltage (and consequently the drain to source on-current) is modified by the work function difference between the gate material and channel material. Because polysilicon is a semiconductor, its work function can be modulated by adjusting the type and level of doping. Furthermore, because polysilicon has the same bandgap as the underlying silicon channel, it is quite straightforward to tune the work function to achieve low threshold voltages for both NMOS and PMOS devices. By contrast, the work functions of metals are not easily modulated, so tuning the work function to obtain low threshold voltages becomes a significant challenge. Additionally, obtaining low-threshold devices on both PMOS and NMOS devices would likely require the use of different metals for each device type, introducing additional complexity to the fabrication process.
2. The Silicon-SiO_2 interface has been well studied and is known to have relatively few defects. By contrast many metal–insulator interfaces contain significant levels of defects which can lead to Fermi-level pinning, charging, or other phenomena that ultimately degrade device performance.

3. In the MOSFET IC fabrication process, it is preferable to deposit the gate material prior to certain high-temperature steps in order to make better-performing transistors. Such high temperature steps would melt some metals, limiting the types of metal that can be used in a metal-gate-based process.

While polysilicon gates have been the de facto standard for the last twenty years, they do have some disadvantages which have led to their likely future replacement by metal gates. These disadvantages include:

1. Polysilicon is not a great conductor (approximately 1000 times more resistive than metals) which reduces the signal propagation speed through the material. The resistivity can be lowered by increasing the level of doping, but even highly doped polysilicon is not as conductive as most metals. In order to improve conductivity further, sometimes a high-temperature metal such as tungsten, titanium, cobalt, and more recently nickel is alloyed with the top layers of the polysilicon. Such a blended material is called silicide. The silicide-polysilicon combination has better electrical properties than polysilicon alone and still does not melt in subsequent processing. Also the threshold voltage is not significantly higher than with polysilicon alone, because the silicide material is not near the channel. The process in which silicide is formed on both the gate electrode and the source and drain regions is sometimes called salicide, self-aligned silicide.
2. When the transistors are extremely scaled down, it is necessary to make the gate dielectric layer very thin, around 1 nm in state-of-the-art technologies. A phenomenon observed here is the so-called poly depletion, where a depletion layer is formed in the gate polysilicon layer next to the gate dielectric when the transistor is in the inversion. To avoid this problem, a metal gate is desired. A variety of metal gates such as tantalum, tungsten, tantalum nitride, and titanium nitride are used, usually in conjunction with high-k dielectrics. An alternative is to use fully-silicided polysilicon gates, a process known as FUSI.

Insulator

As devices are made smaller, insulating layers are made thinner, and at some point tunneling of carriers through the insulator from the channel to the gate electrode takes place. To reduce the resulting leakage current, the insulator can be made thicker by choosing a material with a higher dielectric constant. To see how thickness and dielectric constant are related, note that Gauss' law connects field to charge as:

$$Q = \kappa\epsilon_0\ E,$$

with *Q* = charge density, κ = dielectric constant, ε_0 = permittivity of empty space and *E* = electric field. From this law it appears the same charge can be maintained in the channel at a lower field provided κ is increased. The voltage on the gate is given by:

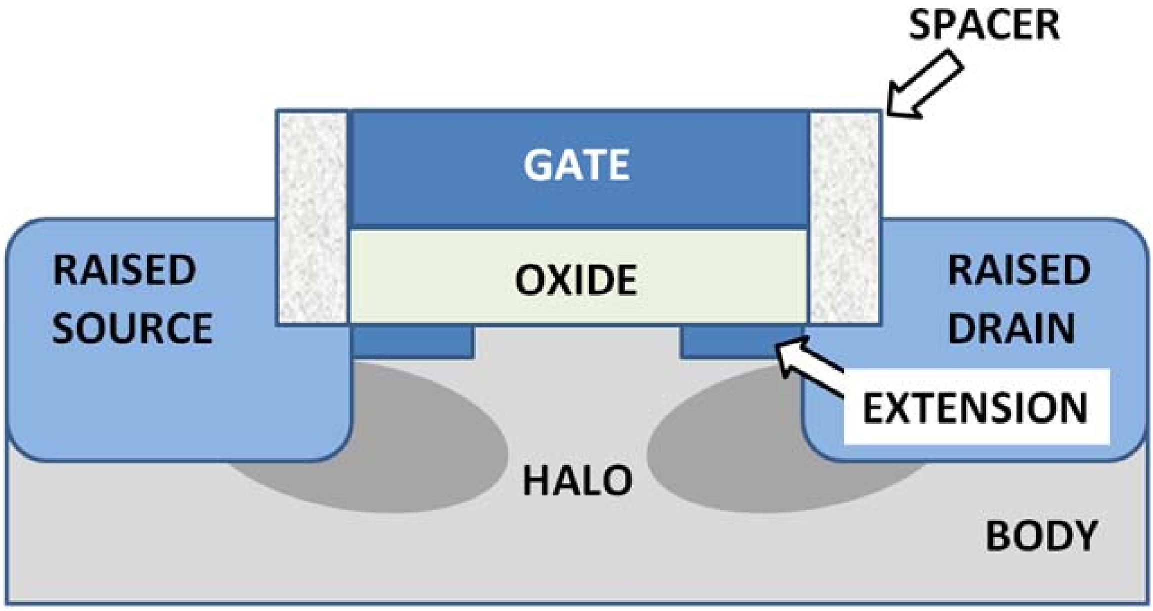
$$V_G = V_{ch} + E\ t_{ins} = V_{ch} + \frac{Qt_{ins}}{\kappa\epsilon_0},$$

with V_G = gate voltage, V_{ch} = voltage at channel side of insulator, and t_{ins} = insulator thickness. This equation shows the gate voltage will not increase when the insulator thickness increases, provided κ increases to keep $t_{ins}/\kappa = constant$.

The insulator in a MOSFET is a dielectric which can in any event be silicon oxide, but many other dielectric materials are employed. The generic term for the dielectric is gate dielectric since the dielectric lies directly below the gate electrode and above the channel of the MOSFET.

Junction design

The source-to-body and drain-to-body junctions are the object of much attention because of three major factors: their design affects the current-voltage (*I-V*) characteristics of the device, lowering output resistance, and also the speed of the device through the loading effect of the junction capacitances, and finally, the component of stand-by power dissipation due to junction leakage.



MOSFET showing shallow junction extensions, raised source and drain and halo implant. Raised source and drain separated from gate by oxide spacers.

The drain induced barrier lowering of the threshold voltage and channel length modulation effects upon *I-V* curves are reduced by using shallow junction extensions. In addition, *halo* doping can be used, that is, the addition of very thin heavily doped regions of the same doping type as the body tight against the junction walls to limit the extent of depletion regions.

The capacitive effects are limited by using raised source and drain geometries that make most of the contact area border thick dielectric instead of silicon.

These various features of junction design are shown (with artistic license) in the figure.

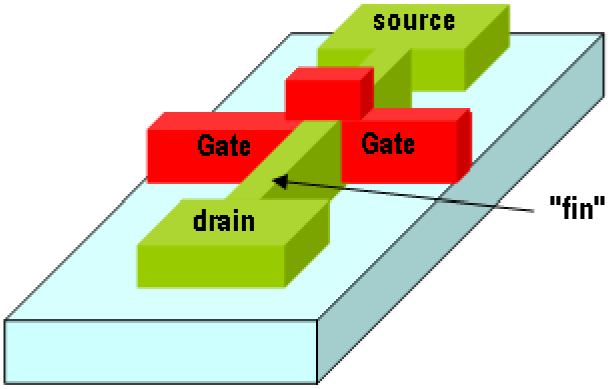
Junction leakage is discussed further in the section increased junction leakage.

Other MOSFET types

Dual gate MOSFET

The dual gate MOSFET has a tetrode configuration, where both gates control the current in the device. It is commonly used for small signal devices in radio frequency applications where the second gate is normally used for gain control or mixing and frequency conversion.

FinFET



A double-gate FinFET device.

The Finfet, see figure to right, is a double gate device, one of a number of geometries being introduced to mitigate the effects of short channels and reduce drain-induced barrier lowering.

Depletion-mode MOSFETs

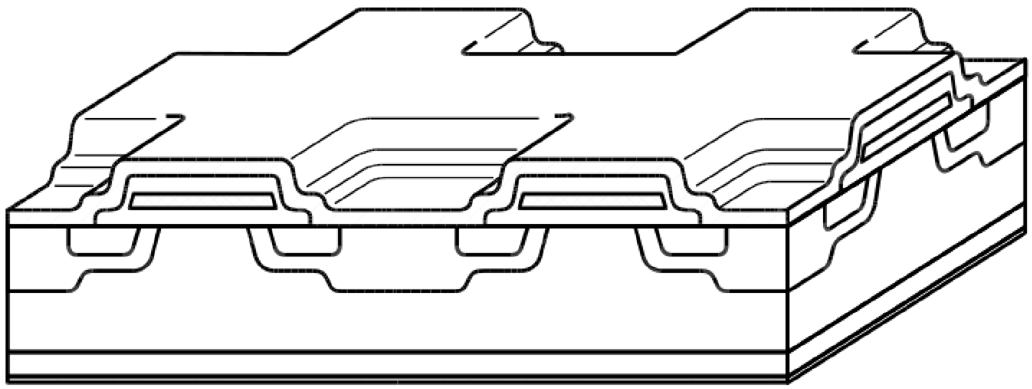
There are *depletion-mode* MOSFET devices, which are less commonly used than the standard *enhancement-mode* devices already described. These are MOSFET devices that are doped so that a channel exists even with zero voltage from gate to source. In order to control the channel, a negative voltage is applied to the gate (for an n-channel device), depleting the channel, which reduces the current flow through the device. In essence, the depletion-mode device is equivalent to a normally closed (on) switch, while the enhancement-mode device is equivalent to a normally open (off) switch.

Due to their low noise figure in the RF region, and better gain, these devices are often preferred to bipolars in RF front-ends such as in TV sets. Depletion-mode MOSFET families include BF 960 by Siemens and BF 980 by Philips (dated 1980s), whose derivatives are still used in AGC and RF mixer front-ends.

NMOS logic

n-channel MOSFETs are smaller than p-channel MOSFETs and producing only one type of MOSFET on a silicon substrate is cheaper and technically simpler. These were the driving principles in the design of NMOS logic which uses n-channel MOSFETs exclusively. However, unlike CMOS logic, NMOS logic consumes power even when no switching is taking place. With advances in technology, CMOS logic displaced NMOS logic in the mid 1980s to become the preferred process for digital chips.

Power MOSFET



Cross section of a Power MOSFET, with square cells. A typical transistor is constituted of several thousand cells

Power MOSFETs have a different structure than the one presented above. As with all power devices, the structure is vertical and not planar. Using a vertical structure, it is possible for the transistor to sustain both high blocking voltage and high current. The voltage rating of the transistor is a function of the doping and thickness of the N-epitaxial layer, while the current rating is a function of the channel width (the wider the channel, the higher the current). In a planar structure, the current and breakdown voltage ratings are both a function of the channel dimensions (respectively width and length of the channel), resulting in inefficient use of the "silicon estate". With the vertical structure, the component area is roughly proportional to the current it can sustain, and the component thickness (actually the N-epitaxial layer thickness) is proportional to the breakdown voltage.

Power MOSFETs with lateral structure are mainly used in high-end audio amplifiers and high-power PA systems. Their advantage is a better behaviour in the saturated region (corresponding to the linear region of a bipolar transistor) than the vertical MOSFETs. Vertical MOSFETs are designed for switching applications.

DMOS

DMOS stands for double-diffused metal–oxide–semiconductor. Most power MOSFETs are made using this technology.

RHBD MOSFETs

Semiconductor sub-micrometer and nanometer electronic circuits are the primary concern for operating within the normal tolerance in harsh radiation environments like space. One of the design approaches for making a radiation-hardened-by-design (RHBD) device is Enclosed-Layout-Transistor (ELT). Normally, the gate of the MOSFET surrounds the drain, which is placed in the center of the ELT. The source of the MOSFET surrounds the gate. Another RHBD MOSFET is called H-Gate. Both of these transistors have very low leakage current with respect to radiation. However, they are large in size and take more space on silicon than a standard MOSFET.

Newer technologies are emerging for smaller devices for cost saving, low power and increased operating speed. The standard MOSFET is also becoming extremely sensitive to radiation for the newer technologies. A lot more research works should be completed before space electronics can safely use RHBD MOSFET circuits of nanotechnology.

When radiation strikes near the silicon oxide region (STI) of the MOSFET, the channel inversion occurs at the corners of the standard MOSFET due to accumulation of radiation induced trapped charges. If the charges are large enough, the accumulated charges affect STI surface edges along the channel near the channel interface (gate) of the standard MOSFET. Thus the device channel inversion occurs along the channel edges and the device creates off-state leakage path, causing device to turn on. So the reliability of circuits degrades severely. The ELT offers many advantages. These advantages include improvement of reliability by reducing unwanted surface inversion at the gate edges that occurs in the standard MOSFET. Since the gate edges are enclosed in ELT, there is no gate oxide edge (STI at gate interface), and thus the transistor off-state leakage is reduced very much.

Low-power microelectronic circuits including computers, communication devices and monitoring systems in space shuttle and satellites are very different than what we use on earth. They are radiation (high-speed atomic particles like proton and neutron, solar flare magnetic energy dissipation in earth's space, energetic cosmic rays like X-ray, Gamma-ray etc.) tolerant circuits. These special electronics are designed by applying very different techniques using RHBD MOSFETs to ensure the safe space journey and also space-walk of astronauts.

MOSFET analog switch

MOSFET analog switches use the MOSFET channel as a low–on-resistance switch to pass analog signals when on, and as a high impedance when off. Signals flow in both directions across a MOSFET switch. In this application the drain and source of a MOSFET exchange places depending on the voltages of each electrode compared to that of the gate. For a simple MOSFET without an integrated diode, the source is the more negative side for an N-MOS or the more positive side for a P-MOS. All of these switches are limited on what signals they can pass or stop by their gate-source, gate-drain and source-drain voltages, and source-to-drain currents; exceeding the voltage limits will potentially damage the switch.

Single-type MOSFET switch

This analog switch uses a four-terminal simple MOSFET of either P or N type. In the case of an N-type switch, the body is connected to the most negative supply (usually GND) and the gate is used as the switch control. Whenever the gate voltage exceeds the source voltage by at least a threshold voltage, the MOSFET conducts. The higher the voltage, the more the MOSFET can conduct. An N-MOS switch passes all voltages less than (V_{gate}–V_{tn}). When the switch is conducting, it typically operates in the linear (or Ohmic) mode of operation, since the source and drain voltages will typically be nearly equal.

In the case of a P-MOS, the body is connected to the most positive voltage, and the gate is brought to a lower potential to turn the switch on. The P-MOS switch passes all voltages higher than ($V_{gate}+|V_{tp}|$). Threshold voltage (V_{tp}) is typically negative in the case of P-MOS.

A P-MOS switch will have about three times the resistance of an N-MOS device of equal dimensions because electrons have about three times the mobility of holes in silicon.

Dual-type (CMOS) MOSFET switch

This "complementary" or CMOS type of switch uses one P-MOS and one N-MOS FET to counteract the limitations of the single-type switch. The FETs have their drains and sources connected in parallel, the body of the P-MOS is connected to the high potential (V_{DD}) and the body of the N-MOS is connected to the low potential (Gnd). To turn the switch on the gate of the P-MOS is driven to the low potential and the gate of the N-MOS is driven to the high potential. For voltages between (V_{DD}–Vtn) and (Gnd+Vtp) both FETs conduct the signal, for voltages less than (Gnd+Vtp) the N-MOS conducts alone and for voltages greater than (V_{DD}–Vtn) the P-MOS conducts alone.

The only limits for this switch are the gate-source, gate-drain and source-drain voltage limits for both FETs. Also, the P-MOS is typically three times the width of the N-MOS so the switch will be balanced.

Tri-state circuitry sometimes incorporates a CMOS MOSFET switch on its output to provide for a low ohmic, full range output when on and a high ohmic, mid level signal when off.